

MY SCALE OR YOUR METER? EVALUATING METHODS OF MEASURING THE INTERNET

Giampiero Giacomello and Lucio Picci[#]

This version: November 2002

Abstract

Measuring the Internet - the size of its infrastructure, how many people use it, and their prevalent uses - is of obvious interest. However, the wealth of available quantitative information regarding the Internet so far has fallen short of satisfying the many needs that it would fulfill.

We set the problem of measuring the Internet into a framework that allows us to derive insights on the peculiar nature of the Internet as a piece of infrastructure. After reviewing the current measures available, while drawing a distinction between the object of measurement, and the types of institutions involved in it, we provide some indications on what data should be trusted more, and how better measures of the Internet could be obtained.

Keywords: Internet, measuring the Internet, infrastructure, public capital.

JEL codes: C81, C82, H54, O31

[#] Giampiero Giacomello: Department of History and Political Science, University of Bologna, (giampiero.giacomello@iue.it) and Lucio Picci: Department of Economics, University of Bologna, Strada Maggiore 45, 40125 Bologna, Italy. (l.picci@ei.unibo.it, <http://www.spbo.unibo.it/picci>). The authors thank for the comments received Larry Press and other participants at the Internet Society INET 2002 Conference (Washington, DC June 13, 2002), Alessandra Colecchia and Hans-Jurgen Engelbrecht. All URL's last checked on 15 November, 2002.

*“In the Internet era,..., the need for
a new measure is emerging”
(Franklin Daniel, 2001)*

1. Introduction

The appreciation of how “big” is the Internet has considerably risen in the second half of the 1990s, as the Net became an object of much interest at least in the industrialized world. Nevertheless, despite increasing awareness about the phenomenon, determining the size and other features of the Internet has remained a daunting task.

True, valuable progresses have been made, by scholars coming from very different fields. As an example of a healthy diversity of backgrounds, consider the involvement of mathematicians and of geographers. Mathematicians and physicists are busy trying to determine the abstract nature of the Internet as a network, for modeling and forecasting purposes. Geographers are shaping what is already called "cybergeography", an attempt to reinterpret geography and space in the age of the Internet¹. These efforts are certainly contributing to a better understanding of the quantitative aspects of the Net. However, to date, some very basic measurement problems remain unsolved. We argue that this is so not just due to the lack of enough numbers but, more important, to some deep rooted unsolved measurement issues.

The end result of this state of affairs is known: confusion. Consider the rate of growth of Internet traffic, estimated by much press - at least before the collapse of the "dot.com" stocks in 2000 - to double “every three months”. Despite such claims, “[...] there have been no hard data to substantiate it” (Odlyzko, 2000). Indeed, in the same paper Odlyzko remarks that the belief that Internet traffic could continue doubling every three months *ad eternum* shows “the lack of simple quantitative reasoning” and, ultimately, a case of “innumeracy”.

The implications of having sometimes wildly diverging "data" on the Internet could be serious: Companies, and governments often base their decisions regarding new technologies precisely on those data whose reliability is shaky. The lack of good data is also a serious drawback for any scientific research aiming at measuring the impact of the Internet itself. Such researches often use econometric tools to estimate the elasticity of output to changes in given inputs, such as a piece of infrastructure. Without reliable data, the mere possibility of measuring the impacts of the new phenomenon, and, as a consequence, of carrying out sound

policies, is left to mere deductive reasoning, or to impressionistic assessments.

There are at least three reasons to explain why, in an age when information of a quantitative nature is ubiquitous, with the Internet we do get numbers, but often not very meaningful ones. The first one is conceptual, and has to do with the distinction between infrastructure and the use that economic agents make of them. The point is clear, say, in the case of roads: kilometers of asphalt on the one hand, decisions to buy gas and drive on the other. Such a distinction is blurred with the Internet. For instance, a Web server represents a use of the Internet, but it can also be a "ramp" to access the infrastructure, as in a Web-based e-mail facility. Also, the same computer that allows uses of the infrastructure could also be dedicated to purposes that are indispensable for the working of the infrastructure itself, such as the routing of traffic².

The second reason is that, even after discounting for this conceptual problem, there is a sheer difficulty in finding the necessary data, in large part due to the decentralized – and, for the most, private – nature of the Internet. Much has been written about the intrinsic decentralization of the Internet. However, such a decentralized structure is not incompatible, in principle, with the presence of a single national provider of at least the backbone connections³. A single national provider would be in the best position to collect and to make public data about the network and its use for one obvious reason – it would control the whole network, and it could define a unique standard of measurement – and for a less obvious one: once there are more providers in a competitive market, they may have reasons to keep their data private, or possibly even to misrepresent them.

Last, there is what we could define a cultural element representing an obstacle to obtaining good measures of the Internet, having to do with the professional cultures of the people involved. The professionals who would be in the best position to make many measurements are the engineers who run the infrastructure, who, unfortunately, generally do not have a firm grasp of economics and statistics. On the other hand, the people who do have that kind of expertise, for example the statisticians at the national statistical offices, are far away from where many useful measures could be taken, and often do not have a clear

¹ For the former, see, among others, Barabási et al. (2000) and Yook et al., (2001). For the latter, among the many contributions available, see the paper by Gorman and Malecky (2002). See also the Web site <http://www.cybergeography.org>

² Gorman and Malecky (2002) rank "fixity" - closer to infrastructure - and "fluidity" - the use of the infrastructure - using OSI seven layer netplex, from the physical layer (the cables), to network processes and applications.

³ This was the situation in the United States until 1995, when the Internet was privatized, in

appreciation of the phenomenon, nor an understanding of the technical issues that are relevant for coming up with meaningful definitions of what should be measured, and how.

The whole issue of developing good measures of the Internet, however, is best seen from a more detached point of view: A new technological revolution is happening, and an impressive technology adoption process is at work. Economic historians underline that understanding and adopting a new technology is always a long and painful ordeal. They argue that, for example, the second Industrial Revolution of electricity and chemistry actually occurred decades after the relevant inventions took place (David, 1990). In the fields of numbers, state statistics was invented in its modern form in Europe already in the XVIII century, in the age of enlightened absolutism, but we had to wait until the 1930s to see the first systematic data on national accounts (Stigler, 1999).

This paper intends to contribute to this new effort of data construction in three steps. First, we investigate the implications of the distinction between infrastructure and its use - or, in different words, between the capital stock proper, and the flow of services that it allows. Second, we provide a snapshot of the current status of research in this field, first by outlining the different objects of measure, and then by investigating the “who” collects the observations, i.e., national statistical offices, international organizations, etc. Last, we propose some policy recommendations to help address the problem of measuring the Internet. We anticipate that our recipe involves a joint effort by many actors, both to overcome the methodological issues, and eventually to collect and organize better measurements of the Internet.

2. Infrastructure data and their use

Economists and economic statisticians distinguish between infrastructure (often also called “public capital”, since it is typically publicly provided, unlike its “private” variety), and the use of infrastructure by economic agents. There are good reasons to keep these concepts separate. One is that, for example, roads are public goods and are usually built with public funds, whereas cars and trucks are not. Moreover, infrastructure very often are networks and natural monopolies. Modern information technologies, by making possible such things as congestion pricing tolls on freeways, may in time put a private touch in many traditionally public goods. However, as of today, traditional infrastructure are still fairly close to the usual

concomitance with the National Science Foundation stopping its funding.

definition of a public good, with all the implications about their provision and pricing.

When it comes to actually measuring infrastructure, two main alternatives are available. Sometimes a physical measure of its consistence is adopted, for example the kilometers of roads, of train tracks, or the number of school rooms. To aggregate these measures, some weighting is often necessary: a six-lane freeway counts more than a one-lane country road. Aggregation is also an obvious problem, whenever a combined measure of different types of infrastructure – say, roads and railroads together – is desired⁴. Besides these physical measures (public) capital is very often measured using a "permanent inventory" technique, that involves adding up past investment flows expressed at constant prices, while deducing the value of assets as they reach the end of their service lives. In order to do it, a sufficiently long time series of the investment flow is needed, as well as information on how the prices of investment goods change in time - or, in different words, the investment data must be expressed at constant prices, in order for their sums to be meaningful (Organization for Economic Cooperation and Development, 2001).

Good data on infrastructure allow for interesting analyses. In particular, it is possible to use a production function to estimate the role of the Internet in determining output, both at the economy wide level, and possibly at the firm level also. The availability of permanent inventory data on public capital has allowed many researchers, particularly after a seminal work of Aschauer (1989), to carry out similar exercises for traditional types of infrastructure. The big questions that inspired at least the first wave of these studies was provided by the economic growth slow-down that was observed, in the United States and elsewhere, from the 1970's, much before the more recent debate on "new economy" high growth rates. The role played by infrastructure in determining output is still to some extent controversial, but there is a greater consensus pointing to its significance and economic relevance⁵. These studies have focused on the role of traditional infrastructure, for which data are available for a sufficiently long period of time, and so far have not considered the potential role of the Internet, for which permanent inventory data are not available.

Just as finding a reason for the growth slowdown ranked high in economists preoccupations in the 1980's and during the early 1990's, later on the question become how to explain the observed revival in growth, and the concomitant Internet revolution provided a

⁴ For such an assessment of infrastructure in Europe, see Biehl (1986).

⁵ For an early survey of the literature, see Gramlich (1994). More recent works (on the US economy) include Pereira (2001), Pereira and Flores-de-Frutos (2000) and Fernald (1999), all pointing to a significant role of infrastructure.

natural suspect. Economists reverted again to econometric tools, and a large literature developed, that we can conveniently divide into two main strands. In both of them, economists had to resort to investment data, referring to private capital goods such as computers and software, or to the stock that they form, that is, to varieties of private capital. One strand of the literature involves "growth accounting" exercises, where, under certain conditions, it is possible to split up the observed economic growth into several factors, to determine, for example, the role played by the labor input, the capital services, etc. What the relation does not explain is usually called the "Solow residual", typically seen as a proxy of technological progress. One important issue in these studies, when they are carried out at the sectoral level, is whether parts of the economy other than the Information and Communication Technologies (ICT) sector are growing faster thanks to new technologies. If this is the case, then we would conclude that any aggregate effect is not just due to the observed progress in the ICT industry, but (also) to the spillovers to other sectors of those technologies, and of the related new organizational practices that they enable.

Using national accounts data, many researchers found that the new technologies were indeed responsible for part of the observed economic growth in the US and in other countries. These results, however, did not go unchallenged; the Fall 2000 issue of the *Journal of Economic Perspectives* provides a review of both camps. These studies employed data on the use of ICT collected from several sources, but that typically did not distinguish Internet related investments from more general ICT expenditure.

A second strand of the literature used data originated from surveys of firms' behavior⁶. As we will see, there are several surveys of this type available internationally, typically run by the national statistical agencies and also by other organizations, where firms are asked questions related to their performances and choice of inputs. The data allows to assess whether firms investing more in ICT perform better, and by how much. These data do not always draw a clear distinction between ICT investment in general, and investment that pertains to the Internet proper, even though questions related to Internet usage have started to appear within already existing surveys.

Overall, the vast body of studies that have tried to measure the "new-economy", while not unanimously, in most cases found evidence of what Robert Solow was missing in a well-known aphorism: "You can find the computer age everywhere but in the productivity statistics" (Solow, 1987).

⁶ See, among others, Brynjolfsson and Hitt (2000), and Brynjolfsson et al. (2002).

3. The object of observation

After having considered how economists often use data on investments and new technologies, we now move on to an analysis of several measures of the Internet, from the ones that are more focused on the infrastructure, to the ones related to its use.

3.1 The cables

According to a well-known metaphor, the Internet is an "electronic highway": cables would be the Internet analogue of roads, possibly with a parallel hierarchy - just as there are many types of roads, so there are many kinds of cables, from telephone lines to fiber optics. Focusing on the cables that make up the Internet would apparently be the closest thing to looking at the Internet as a piece of infrastructure. However, and for several reasons, the electronic highway metaphor is not very helpful in understanding the infrastructure nature of the Internet. Traditional roads are characterized by a more or less fixed (in time) relationship between the type of road and the traffic it can sustain. This is not the case for electronic highways. A good example is provided by the telephone lines used for dial-up, and now also xDSL, connections from home. Dial-up connections speed increased by an order of magnitude in a decade, as modem technology improved, and xDSL technologies then pushed the limit much further.

This limitation notwithstanding, the knowledge of the quantity and of the types of cables used for Internet traffic still provides very useful information to get some idea of the size of the Internet. The role of technological progress limits the possibilities of comparisons in the time dimension, but, given that the same technologies are today available at least in principle all over the world, it does not preclude useful cross sectional comparisons. Also, an assessment of the quantities of cables somehow provides an upper bound for the overall transmission capacity of the network, to be reached in case the best technological practices were available everywhere and to everyone.

Data on the quantity and quality of Internet cables are not always readily available, mainly because different organizations - mostly private firms - are responsible for the investments. The main backbone connections, and other major elements of the Internet, such as Network Access Points, or "NAPs", are fairly well documented, even if with limitations (see Gorman and Malecki, 2002).

However, as we move away from the backbones, and away from the United States,

information deteriorates. In some instances, there may exist some information on the kind of Internet connections that are available within a given community - for example, we may know that somewhere there is a "Metropolitan Area Network" with a given data transmission capacity. Where there exist networks dedicated to research and educational purposes, good data about their nature (and sometimes about their use) are often available. However, these dedicated networks only represent a shrinking subset of the whole Internet.

Part of the cables that make the Internet, moreover, also serve the telephone network, the most significant example being the lines that allow dial-up and xDSL access to an Internet Service Provider. In this case, an available measure of the infrastructure is provided by the number of people who have a telephone service subscription. The rationale for this is very simple: if a person has a telephone subscription, then the cable is also there for a potential Internet dial-up or xDSL access⁷. This examples introduces a theme which recurs frequently in measuring the Internet: often we would like to have a measure of the supply of the infrastructure, but we can only proxy it with some measure of its use. This is also the case with Domain Names counts, that we now consider.

3.2 Internet Domain Names

IP (Internet Protocol) numbers are the unique 4-part numbers assigned to each and every computer linked to the Internet, such as "137.204.152.151", and a "domain" is a mnemonic string for a set of IP numbers. Domain names are certainly informative with respect to the size of the Internet infrastructure: roughly speaking, the more cables there are, and the more complex is the way they connect with each other, the more domain names have to be allocated. However, domain names are also a measure of *demand*. Given the size of the infrastructure, there is a relationship between the number of domain names and the use of the Internet: the more Web sites there are, the more email services, etc, the more domain names are allocated.

Worse than the usual blurring between measures of the infrastructure, and its use, a host of technical reasons dictate that the direct relationship between domain names and size of the Internet is only approximate. First, it is not easy to judge the geographic correspondence of a domain name. One reason is that the popular "top level domain" (TLD's), such as .com,

⁷ This is true only as a first approximation. The quality of the telephone connection influences the maximum speed of transmission in a dial-up connection. Moreover, xDSL may not be available at a given place due to technical reasons or simply because a market for its provision has not (yet) formed.

.net, .org are not geographic, and the only way to know where the computers using them are, is to ask their administrators. Moreover, even the "Country Code TLD" (ccTLD), such as .fr for France, or .it for Italy, are not really "geographic", in the sense that a computer in a ccTLD could physically be anywhere.

Second, not all allocated domain names are actually used. Over the last few years, and particularly so during the ".com" frenzy of 1999 and 2000, domains with possible commercial value were taken up by people who, in different ways, were hoping to make a profit. There is no available reliable estimate of the dimension of "cybersquatting", as this phenomenon has been named, even if we know that it was primarily confined to a subset of rich countries, where it may make sense to pay the fees required to hold a domain name idle. Cybersquatting is hard to characterize and to control for, also because it depends on several factors related to developments in technology, such as search engines (Gillmore, 2002), that make it less important to have one's content associated with a domain name, and maybe in the future the Semantic Web (Berners-Lee et al., 2001).

Third, while to an assigned domain name there used to correspond a single computer, or "host", now to a single host could correspond many domain names, due to "virtual hosting", whereby a single server can effectively act as if it were several hosts (for example, by providing Web sites to several organizations, with their distinct domain names).

Notwithstanding these limitations, and pretty much due to the lack of better alternatives, counting domain names is a widely used way to measure the size of the Internet infrastructure, and it is also informative with respect to uses of the Net.

3.3 Servers and their use

Computers connected to the Internet and offering services, or servers, are at the same time part of an infrastructure definition of the Internet, since they contribute to its working, and a manifestation of the uses of the infrastructure. Analyses of servers or, more to the point, of the software that they have installed, can be carried out both through surveys to system administrators, and with "bots". A bot is a generic name for a program often based on Artificial Intelligence algorithms that may be employed to study design and implementation of self-organizing and self-assembling artifacts⁸ or computer games. In the case at hand, so called "webbots" or "web spiders" may be "launched" to automatically analyze computer

⁸ On this, see for instance the "Future and Emerging Technologies" program of the European Commission at <http://www.swarm-bots.org/>.

connected to the Internet, in order to determine some relevant characteristics. Such automated programs can provide information, for example, on market penetration rates for Web servers - for an example see Table 3 - or of the number of "secure hosts", a family of software that allows for secure economic transaction on the Internet, and that is often taken to signal the presence of economic commerce activities⁹.

Such analysis can also provide useful information, more generally, of Web related activities, such as the number of Web services active, and the amount of information offered by Web servers.

3.4 Counting the users

Users of the Internet can be individuals or organizations. In both cases, counting them is difficult. First, given the decentralized nature of the Internet, there is no unique registry, akin to the telephone white pages, of the people who subscribe to the service. Moreover, at least for individual users, in order to access the Internet there is no need to register to a service, given that an access at least to the Web may be available in the workplace, at a public library, at a friends' house, or through a cellular phone. Another problem has to do with the definition of "using the Internet". The number of people who meet the requirement of "accessing the Web at least once a month", for example, may be significantly greater than the numbers who satisfy the requirement of "checking e-mail at least five times a week". Moreover, even once we have taken this into account, different communities (or countries) could differ in the time spent using given applications, as stressed in Kirkman et al. (2002), thus complicating comparisons.

One way to estimate the number of Internet users is, simply, by surveying random samples of people and by asking them. These surveys may ask several questions so that interesting relations between characters - say, use of the Internet and age, gender, or income - can be ascertained. Another solution consists in using the domain name count that we have considered, multiplying it by a proportionality factor that typically depends on the country, to take into consideration that in poorer countries a single e-mail account tends to be used by more people. However, given the very imprecise relation existing between the number of

⁹ Bots also can operate in two logically distinct way. They can "crawl" the net, that is, go from one computer to others linked to it (for example, via links in the pages of a Web server), or they can analyze a sample of servers selected under some criterion.

domain names and the various characteristics of the Internet, such estimates, while popular, amount to no more than educated guesses.

Last, information on the number of users can implicitly be obtained from the analysis of log files¹⁰. For instance, a geographic area where Internet users make up many hits in Web server logs, is a place where there should be many users. We will describe the limitations of this technique later on.

Estimating how many organizations are on the Internet (and doing what) is largely left to surveys. These can either aim at the whole population of interest or, more often, at randomly selected samples. The existing applications of surveys to this purpose are many, involving both the private sector – we have already considered how these data are used to assess the importance of the “new economy” – and the public administrations, where the focus may be, for example, on the assessment of “e-government” practices. Beyond doubts, at least in the industrialized world, the shift is rapidly changing from counting who is on the Internet, to trying to understand what they are doing there, more or less assuming that all organized entities are already connected to the Net at least to some extent, or will be so shortly.

3.5 What the users do

Activities on the Internet may be characterized in many ways. On the one hand, we may distinguish between uses of different applications, such as the Web. Also, it may be of interest to qualify further these uses, for example to assess whether, within an organization, e-mail is used only for job-related purposes, or also to organize parties after work.

A first solution to the problem, again, is provided by surveys. In the present context, these can take several forms. They can be carried out on random samples of the population at large. In fact, the relevant population for a survey of Internet users is the whole set of people who use the Internet, so that surveys could be carried out by interviewing a sample of users only. However, selecting appropriate samples is difficult, given the lack of a general directory of users. Also, the problem in defining what a user is, blurs the boundaries of the users' population. These limitations notwithstanding, surveys of these type have been conducted in a variety of ways, sometimes using email to solicit an answer, and sometimes via Web forms.

¹⁰ Log files are the files where accesses (for example, to a Web site) leave their mark. They provide three basic types of information: what page or service was accessed, when, and by what Internet address (IP number).

In most cases, the statistical properties of the sample are dubious to say the least, as is the statistical inference drawn from them. Very often, these surveys are not unlike what goes under the jargon name of "straw polls", where self-selection invalidates statistical inference. In some instances, samples of people are selected to willingly install on their computer a dedicated software that effectively tracks relevant behaviors, such as which software is used, at what time of the day, and for how long. The software may also record what sites have been visited, which plug-ins or other applications have been used, etc. Such technique, in principle, allows for obtaining a data panel, that is, a cross section of individual observations repeated in time.

A general problem with surveys is that the definitions of the phenomenon that they employ are *ad hoc*, since no common set of definitions has emerged so far. Another problem is that, as we will see, they are often carried out by private institutions who are wary about disclosing the methodological information that would allow an outside observer to judge their scientific rigor. The mere fact that most times only point estimates are published, with no reference about intervals of confidence, indicates that such rigor is generally lacking.

Another way to inquire into people's use of the Internet is by looking at Web log files. The focus on the Web arguably is not a severe limitation, given that the Web has increasingly become the general interface to the Internet. While informative, log files analyses have shortcomings. First, they are collected at each site, so that, in order to obtain them on a wider basis, agreements have to be reached with the (usually private) entities who own them. Such agreements have to imply trust, because log files are simple text files that can be easily tampered with. Moreover, a log file analysis cannot yield any particulars of the specific users, nor of their preferences or identities. It cannot even determine if, during a log session, one or more users accessed the computer. Last, Web pages are often seen without causing a "hit" on the servers' log file because they had been previously "cached" in "proxy servers". One last element of trouble in analyzing Web logs is the presence of "bots" themselves. Such programs, now ubiquitous on the Internet, leave a mark on the log files, and tend to complicate their analysis since they are not always clearly recognizable as such.

Separate attention has to be dedicated to what organizations, instead of individual users, do on the Internet. The range of possible activities is very wide. On one end of the spectrum we could place a firm having an e-mail box that a willing employee checks from time to time. On the opposite end, a big organization that has successfully reorganized itself through a massive use of Internet related technologies. In all cases, no standard definitions for

the different uses of the Internet have emerged yet, and this, even regardless of other measurement problems, is an obvious obstacle for the comparisons and the analysis of data. The issue is relevant also because it has an impact on the measurement of the effects of ICT investments itself. In particular, there is both anecdotal and quantitative evidence that ICT usefulness increases when it is accompanied by a parallel investment in "organizational capital", in the form restructuring, training, etc. (see Brynjolfsson et al., 2002, and the discussion therein). Better understanding this problem requires a clearer perspective on the prevalent uses of the technology, to understand, for example, to what extent within organizations they are used to delegate more, to control better, or maybe both.

3.6 Composite indexes

At the end of the list of the different objects of measurements, we recognize that a possibility is to wish to look at many things at the same time. Such a desire is at the base of composite indexes aggregating several primitive measures in order to provide a summary view of how people, places, organizations use the Internet, and of what is their potential for use. As always with composite indexes, their advantage is their capability to condensate in a single number a complex reality, and their shortcoming is their sensitivity to the underlying assumptions, to the aggregation procedure, and to the quality of the primitive indexes themselves.

Relevant examples of broad-based composite indexes are the Networked Readiness Index (Kirkman et al., 2002) created by the Harvard Center for International Development¹¹ or the Information Warfare Index (Giacomello, 2002). The former combines several national indexes from advanced technologies (especially in telecommunications and computers) with other "social" and "psychological" measures (such as the level of entrepreneurship of individualism), to rank 75 selected countries according to their propensity toward the "networked society". The latter, applying similar methodology, ranks some 60 countries according to their capabilities in waging offensive and defensive information warfare.

4. The data available.

We have seen what types of measures and of techniques are available, together with their qualities and shortcomings. We now describe the broad categories of institutions who are

¹¹ Details are at http://www.cid.harvard.edu/cr/gitrr_030202.html

involved measuring various characteristics of the Internet, also to consider how their nature intersects with the scope and the quality of the data that they collect. We begin with an assessment of the role of international organizations.

4.1 International Organizations

Many international organizations, such as the United Nations, the OECD (the Organization for Economic Co-operation and Development), the International Telecommunication Union (ITU), the World Bank and the European Commission statistical office have devoted resources and attention to measuring the Internet. One of the activities of international organizations is the provision of data that are internationally comparable; however, “[...] the global statistical system is founded, in statistical work, at the national level”,¹² and while they can choose to collect data themselves, in general they rely on the information provided by the national offices or by other organizations. Thus, if national figures are not dependable; if data are not collected nationally on certain features or activities, or if the data are collected, but according to contrasting definitions of the object of analysis, then comparisons at the international level becomes problematic.

First and foremost, among international organizations, the United Nations Organization (UN) has a considerable expertise in gathering data on different countries to make comparisons possible, and the UN statistical division offers a comprehensive picture of world statistics on a very broad spectrum of topics. To monitor the fulfillment of the UN Millennium Declaration (adopted by the UN General Assembly in September 2000, and describing the role of the UN in the century ahead) (United Nations, 2000), the UN statistical division has assembled a global database on 48 social and economic indicators,¹³ including the numbers of Internet users and of personal computers. These data, however, are collected by the International Telecommunication Union (ITU). With 189 member states, ITU was founded in 1865, is one of the oldest international organizations, and is now the UN agency dedicated to telecommunications. ITU collects data on Internet users and number of computer users per country¹⁴.

Among the several international economic organizations that are involved in measuring the Internet, the Organisation for Economic Cooperation and Development

¹² United Nations Statistical Division, “About Global Statistics”, <http://www.un.org/Depts/unsd/global.htm>

¹³ <http://millenniumindicators.un.org/>.

¹⁴ <http://www.itu.int/ITU-D/ict/statistics/>.

(OECD) has set up an ambitious plan to measure various aspects of the “new economy”. These include measures of ICT industries (Organization for Economic Cooperation and Development, 2000) of the types used for the studies on the emergence of the “new economy” that we have considered, assessment of E-commerce and measures of the infrastructure. By "measure of the infrastructure", however, the OECD means those measures, such as the number of domain names and analysis of log files of various type, that we have argued to provide in fact an estimate of a mixture of supply and demand elements (Organization for Economic Cooperation and Development, 1998). Many of the data collected by OECD are published in the so called "Scoreboard"¹⁵. Part of these studies are carried out by using data developed by other institutions, and part through surveys conducted by OECD. To this purpose, OECD is developing a "model questionnaire" to measure "ICT use and electronic commerce in enterprises". A summary of the various activities of OECD in this field, with indications of future plans, is in Gault and McDaniel (2002).

4.2 National Statistical Offices and other governmental entities

So important is national statistical offices (NSOs) official validation of social phenomena and economic performance that some authors (e.g. Minges, 2000) demand greater involvement in helping to measure the Internet. The NSOs of some countries, typically the ones where the Internet is more widespread, have begun to think about measures of the Internet and, in some instances, have started collecting the data. The US Census Bureau is a noteworthy example, both because of the quantitative and qualitative relevance of the United States, and because it was among the first NSO's to consider the issue. The US Census Bureau made available on-line the data on Internet use in 1996, albeit only on the preferences and opinions of users accessing the Bureau. For more detailed data on users, however, even the Census relied on figures collected by private companies.¹⁶

More recently, the U.S. Census has devoted a project to “measuring the electronic economy” (E-stats)¹⁷. In a series of papers (Atrostic et al., 2000, Mesenbourg, 2000, Mesenbourg, 2001), the US Census laid out its strategy with respect to measuring the Internet or, more to the point, since the focus is explicitly on its economically relevant uses, the "electronic economy". The US Bureau of Census first developed and adopted some preliminary concepts of "three key components" of the electronic economy: "Electronic

¹⁵ <http://www.oecd.org/EN/document/0,,EN-document-41-1-no-1-17270-0,FF.html>

¹⁶ See for instance the data available on <http://www.census.gov/statab/freq/00s0913.txt>

business", defined as "any process that a business organization conducts over computer-mediated network channels"; "Electronic commerce" is "any transaction completed over computer-mediated network channels", and "E-business infrastructure", the "economic infrastructure used to support electronic business process and conduct electronic commerce transaction. It includes the capital (hardware, application software, human capital, and telecommunication networks) used in electronic business and commerce". Note, in particular, the broad infrastructure concept, that includes very diverse types of capital, with human capital among them. Also, note that the concept of "E-business" subsumes phenomena that are often given different names, such as "technology enabled reengineering of organizational processes", "E-government" (and E-governance), etc.

Given these broad definitions, the US Census Bureau aims at answering two distinct sets of questions: measuring "the dimension of the electronic economy", and describing the "impact of the electronic economy on businesses, workers, sectors, regions, and the entire economy" (Altrostic et al, 2000). This activity is carried out in part by setting up new surveys, and in part by adding appropriate questions to existing Census survey. For example, the data on E-commerce¹⁸ are obtained as part of the Census Monthly Retail Trade Survey, that exists, in its monthly format, since 1951¹⁹.

Besides NSO's, data are sometimes collected, or at least organized, also within ministries or other governmental organizations. For example, the U.S. Department of Commerce (DOC) has been involved in projects to assess the digital economy (Department of Commerce, 2000 and 2000). However, the data reported, for instance, in Department of Commerce (2002), come from external sources, and emphasis is placed on data analysis, not in data collection.

4.3 Academic research institutions

Academic research institutions have conducted research programs whose aim is to collect data on the Internet, either as an end in itself, or instrumentally in order to estimate the impact of the Internet, often involving surveys to users. Given the numbers involved, it is impossible to briefly review this field of research, and we limit ourselves to few representative examples summarizing the main interests addressed.

Several researches have aimed at overall assessments of the changes that the new

¹⁷ <http://www.census.gov/eos/www/ebusiness614.htm>

¹⁸ For the latest available, see <http://www.census.gov/mrts/www/current.html>

technologies are causing. For example, the UCLA Center for Communication has recently published the "year two" report of its long term "Survey of the Digital Future"²⁰, a multinational inquiry into how the Internet is changing everyday life. In June 1999, the University of Texas at Austin (2001) launched a major study on "Internet Economy Indicators",²¹ aimed at measuring the complexity of the digital economy and producing viable indicators. The study stopped in 2001.

Other research institutions specifically focus on estimating the technical performance of the Internet, including (and especially) traffic analysis. In this respect, the Cooperative Association for Internet Data Analysis (CAIDA)²², based at the University of California San Diego, provides an interesting example of an attempt to "create a collaborative research and analytic environment in which various forms of traffic data can be acquired, analyzed, and (as appropriate) shared"²³. One of the output of CAIDA is maps of the Internet, a production that is becoming more important and more voluminous, as the need of visualization of the Net, and also the contribution of geographers, increases.

One aspect that many researches of this type share is their discontinuity. Motivated people come and go, as do research grants. Several projects that started with a long-term view eventually ended up being shot-term affairs.

4.4 Internet bodies

One of the noteworthy characteristics of the Internet has been the ability of loose communities of technologists to organize themselves to "govern" various aspects of the network. Among such organizations, the Internet Software Consortium (ISC) is responsible for the main domain names count available. The basis for such a count were laid out in an Internet Engineering Task Force (IETF) "Request For Comment" dated January 1992 (Internet Engineering Task Force, 1992)²⁴. Table 1 reports yearly observations from 1993 of the ISC total host count.

Other organizations collecting data on domain names are the area Regional Internet

¹⁹ See <http://www.census.gov/mrts/www/noverview.html>.

²⁰ <http://ccp.ucla.edu/pages/NewsTopics.asp?Id=27>

²¹ <http://www.internetindicators.com/internetindic.html>

²² <http://www.caida.org>

²³ <http://www.caida.org/home/about/>

²⁴ A "Request For Comment" is a typical example of Internet governance. While not being an "Internet standard" (see <http://www.rfc-editor.org/rfcfaq.html>), it may become a *de-facto* one if it obtains sufficient consensus.

Registries (RIRs), whose aim is to provide IP services worldwide²⁵. For example, RIPE provides data on the number of hosts²⁶, supplementing those already made available by the ISC.

Table 1. Internet Society Consortium Host count (in thousands)

Jul 2002	Jan2002	Jan2001	Jan2000	Jan1999	Jan1998	Jan1997	Jan1996	Jan1995	Jan1994	Jan1993
162,128	147,345	109,574	72,398	43,230	29,670	16,146	9,472	4,852	2,217	1,313

Source: <http://www.isc.org/ds/WWW-200201/index.html>

4.5 Pollsters

Many professional pollsters conduct surveys on Internet usage. The details of these polls are almost inevitably not made public; it is then not obvious whether their quality would stand accurate scrutiny. While serious pollsters should be expected to do their work professionally, not all pollster are serious, and pollster's reputation is the main way to judge the numbers that they produce. Traditional polls are mostly carried out by means of computer aided telephone interviews (or "CATI"). Recently, however, several pollster have begun conducting on line polls on many topics, Internet usage being one of them. The advantage of such practices is lower costs compared to traditional CATI surveys. The main limitation of such practices is the difficulty to come up with samples that correctly represent the population of interest²⁷.

4.6 Other private organizations

Several other private organizations are currently involved in the measuring effort. The characteristics of their analyses depends on their purposes. Firms involved in marketing analyses typically use polls on samples of the population or other surveying techniques, such as focus groups. Some firms employ dedicated software that keeps track of the behavior of a sample of users. Marketing and advertising researchers mostly rely on server log file analysis, on bots, and on surveys (Measurecast, 2001) of various types.

²⁵ There are three of them: ARIN, catering the Americas and Sub-Saharan Africa; APIC, responsible for Asia and the Pacific, and RIPE, in charge of Europe, part of Africa, and the Middle East.

²⁶ <http://www.ripe.net/ripenc/p-services/stats/hostcount/index.html>

²⁷ An interesting evaluation of the pros and cons of on-line polls is in Schonlau et al. (2002). The National Council on Public Polls (NCPP, an association of polling organizations established in 1969), acknowledging the problem, has published a set of 10 guidelines intended to help journalists to avoid major mistakes and blunder when using figures from on-line polls (<http://www.ncpp.org/internet.htm>).

These analyses are often prone to a serious lack of methodological transparency. As an example, consider the methodological note published by Nua²⁸ for its well known "How many on line" survey, which cannot be considered a "methodological note" by any scientific standards²⁹ (See Table 2 for a summary of the last data available). Engelbrecht (2001) has detected and exposed similar problems with the "Information Society Index" developed by World Times/IDC. As an example of a survey conducted using bots, Netcraft³⁰ publishes data on several aspects of Internet usage, such as the use of Web servers (see Table 3 for their latest data). Again, it is hard to judge the reliability of such estimate, whose details are not open to scientific scrutiny.

Table 2. Nua's "How many on line" count: February 2002.

World Total	Africa	Asia/Pacific	Europe	Middle East	Canada / USA	Latin America
605.60	6.31	187.24	190.91	5.12	182.67	33.35

Source: http://www.nua.com/surveys/how_many_online

Table 3. Market servers for top Web Servers - Active sites.

Developer	September 2002	Percent	October 2002	Percent	Change
Apache	10449418	64.85	10470848	65.39	0.54
Microsoft	4071863	25.27	4013397	25.06	-0.21
Iplanet	237802	1.48	227424	1.42	-0.06
Zeus	220729	1.37	215957	1.35	-0.02

Source: <http://www.netcraft.com/survey/>

Other organizations take a broader view at the quantitative aspects of the Internet, one good example being provided by TeleGeography³¹. TeleGeography provides detailed

²⁸ Nua, founded in 1996, is owned by the Scope Communications Group (Ireland). See: <http://www.nua.com/surveys/about/index.html>

²⁹ See http://www.nua.com/surveys/how_many_online/methodology.html. It has do be added that the survey is presented by NUA as nothing more than an "educated guess".

³⁰ See <http://www.netcraft.com>

³¹ It Web site is at <http://www.telegeography.com>

summary data on backbone connections, location of NAPs and other characteristics of the Internet (and telecommunication) infrastructure, including pricing and traffic flows³².

Unfortunately, the price at which private organizations sell their data make them not affordable to many. For example, the valuable TeleGeography yearly almanacs sell for a hefty US \$ 2195. A viable alternative to such high costs, more coherent with an idea of "open science", could come from the involvement of non profit organizations. As an example, the Pew Internet & American Life project, funded by the Pew Charitable Trust, collects and makes available to all survey data, reports and analysis on the Internet. Organizations of this type are well suited for providing useful and independent information about the Internet and the Information Society.

Several non profit organizations are also involved in trying to measure other aspects of the Internet. An example is provided by CERT Coordination Center (CERT/CC), which publishes data on the number of security "incidents" reported to them each year. Other CERT offices are "nationally based" (i.e. they monitor incidents on a "national" computer network), usually with governments, universities or private corporations. There is an ongoing attempt at international coordination in this field through associations such as the Forum of Incident and Response Security Teams (FIRST).³³ However, international comparisons of incidents data are still in their infancy, and international organizations statistical offices do not seem to be concerned with this type of information.³⁴ This circumstance clearly hampers improvements of measurement techniques in this specific subject.

5. Conclusions: Where should we go from here?

We have investigated the implications for measuring the Internet of the distinction between infrastructure and its use, and we have provided a snapshot of the current status of knowledge. The question that we now pose is: Where do we go from here? How can the present situation be improved, and who is in the best position to make a contribution?

The first problem that we have highlighted has to do with the conceptualization of the

³² Without an ability to distinguish among the different applications that generate the traffic (personal communication to the authors, Internet Society INET 2002 Conference, Washington, DC June 13, 2002). On TeleGeography data, see also Engelbrecht (2001) and Gorman and Malecki (2002).

³³ The CERT Web sites at the Carnegie Mellon University is at <http://www.cert.org/>. The FIRST Web site is at <http://www.first.org>.

³⁴ Occasional analyses, mostly based on CERT data, on the occurrence of incidents on the Internet are published (see, for instance, Howard, 1997).

Internet as a piece of infrastructure. Among econometricians there is the saying that there can't be "measurement without theory"³⁵: A conceptual effort in trying to better understand the nature of the Internet as a piece of infrastructure, that is, a better theory, is a prerequisite for making progresses in obtaining good measurements of the Internet. The burden of such an effort would obviously fall on the shoulders of the academic community, and an interdisciplinary effort is needed. In this respect, the fact that researchers with very different background are already involved is highly positive.

The benefits of clarifying our thoughts about the infrastructure nature of the Internet may well spillover to the analysis of more traditional types of infrastructure. Even roads are not what they used to be: New technologies are taking over, making possible congestion pricing, advanced forms of planning of infrastructure use, and futuristic approaches to the management of logistics. Computers and networks are being integrated into buildings, to the point that a word, "domotics", has been coined to analyze and describe such a process. The use of the new information technologies defines new types of infrastructure, whose components, just as it happens with the Internet, could be seen as services, as an expression of demand, and at the same type as an integrant part of the infrastructure proper. In other words, the analytical ambiguity that today we witness for the Internet, is spreading to more traditional realms. This makes all the more urgent an updating to our conceptual framework.

The need for a common definition of the objects of measurement also limits comparisons of data. In this respect, two types of organizations are called to action. First, the effort of international organizations, such as OECD and ITU, firmly rooted into the broader issue of international harmonization of statistical information (e.g. Lynn, 2001), should be made more intense and visible. However, no real progress can be made without a greater involvement of the engineers who run the infrastructure. The Internet governance institutions so far have operated on networking standards (the Internet Engineering Task Force, IETF, or the Internet Architecture Board, IAB), Web standards (the World Wide Web Consortium, W3C), on the managing of domain names (Internet Corporation for Assigned Names and Numbers, ICANN), or on Internet public policies (the Internet Society, ISOC). Several of these organizations have published guidelines on various technical matters; however, guidelines on "how to measure" the Internet have been remarkably lacking. The model for such an action could be a Request For Comments scheme akin to what has accompanied the technical developments of the Internet - including, for instance, the introduction of the

³⁵ With reference to Koopmans (1947).

TCP/IP protocol - or any suitable formalization of a similar "Internet governance" approach.

A coordinated effort from this camp would have two advantages. First, it would be the best way to build consensus on the need for measurement within the community of technologists that runs the Internet, and that are in the best position for many types of measurements (traffic, log files, etc.). Second, if carried out with the involvement of economists and economic statisticians, it could help put together the different expertise that are needed to obtain good measures, and bridge that cultural gap that is one of the reasons for the lack of good data today. The lesson of Internet governance should also be learned by the international organizations such as OECD and ITU. Too often their working is not very transparent to outside observers, and it fails to build the necessary consensus of the interested parties, including, ultimately, the users.

Such criticisms and suggestions for improvement, however, are best seen in perspective. Nobody foresaw that the Internet would become so important and few, if any, among the first Internet users and developers thought about assessing its long-term impact. We should then not be surprised that it took time before the need for good measures affirmed itself. One of the reasons for the impressive success of the Internet rests in its technological soundness, which is guaranteed by an interesting governance that puts a premium to professional expertise. Such expertise, quite understandably, so far has been devoted to guarantee the working of the infrastructure, not its measurement *per se*.

Experts from several fields, the engineers and among them, should now work together to come up with better solutions to the problem of measurement. Units of measure are standards that individuals in society agree to use. A coordinated and structured collaboration of many actors is necessary in order for standards to emerge, to be agreed upon, and to be adopted by all interested parties.

BIBLIOGRAPHY

- Aschauer, D., 1989, Is Public Expenditure Productive?, *Journal of Monetary Economics*, vol. 23, pp. 177-200.
- Atrostic, B.K, J. Gates and R. Jarmin, 2000, Measuring the Electronic Economy, Current Status and Next Steps, US Census Bureau, mimeo, <<http://www.census.gov/eos/www/papers/3.pdf>>.
- Barabási, A., A. Réka, and J. Hawoong, 2000, Scale-free characteristics of random networks: The topology of the World Wide Web, *Physica, A* 281, 69-77.
- Berners-Lee, T., J. Hendler and O. Lassila, 2001, The Semantic Web, *Scientific American*, 17 May, 284(5), pp:34-43.
- Biehl, D., 1986, The contribution of infrastructure to regional development. The infrastructure study group, final report (Commission of the European Community, Brussels).
- Brynjolfsson, E., L. M. Hitt, 2000, Beyond Computation: Information Technology, Organizational Transformation and Business Performance, *Journal of Economic Perspectives*; 14(4), pages 23-48.
- Brynjolfsson, E, Hitt, L. M. and S. Yang, 2002, Intangible Assets: Computers and Organizational Capital, with discussion, *Brookings Papers on Economic Activity*: 1, pp. 137-198.
- Daniel, F., 2001, The Next Measure of National Machismo, *The World in 2001*, *The Economist*, pp.116-119
- David, P., 1990, The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox, *American Economic Review*, May, pp. 355-361.
- Department of Commerce, 2000, *Falling Through the Net*, Washington, DC, February

<http://www.ntia.doc.gov/ntiahome/ftn00/Falling.htm>

Department of Commerce, 2002, Digital Economy 2002, Washington, DC, February
<http://www.esa.doc.gov/508/esa/DIGITALECONOMY2002.htm>

Engelbrecht, H., 2001, Statistics for the Information Age, Information Economics and Policy,
Vol. 13, Issue 3, pp. 339-349

Fernald, J., 1999, Roads to prosperity? Assessing the link between public capital and
Productivity, American Economic Review, 89(3), pp. 619-638.

Gault, F., and McDaniel, S. A., 2002, Continuities and Transformations: Challenges to
capturing Information about the "Information Society", First Monday, February,
http://www.firstmonday.org/issues/issue7_2/gault/index.html

Giacomello, G., 2002, Measuring Infowars: Learning from the Experience of Peace Research
and Arms Control", mimeo, New York; Social Science Research Council, August

Gillmore, D., 2002, Google effect reduces need for many domains, Silicon Valley.com, Jan.
12, <http://www.siliconvalley.com/docs/opinion/dgillmor/dg011301.htm>

Gorman, S. P., and E. J., Malecki, 2002, Fixed and fluid: stability and change in the
geography of the Internet, Telecommunications Policy 26, 389-413.

Gramlich E., 1994, Infrastructure Investment: A Review Essay, Journal of Economic
Literature, Vol. 32, pp. 1176-1196.

Howard, J. D., 1997, An Analysis of Security Incidents on the Internet, 1989-1995, Ph.D.
dissertation, Pittsburgh, University of Pennsylvania,
<http://www.cert.org/research/JHThesis/Start.html>

Internet Engineering Task Force, 1992, Request For Comments n. 1296,
<http://www.ietf.org/rfc/rfc1296.txt?number=1296>

Kirkman, G., P. Cornelius, J. Sachs and K. Schwab (eds.), 2002, Global Information Technology Report 2001-2002: Readiness for the Networked World, World Economic Forum Report, Oxford and New York: Oxford University Press

Koopmans, T. C., 1947, Measurement without theory, *The Review of Economics and Statistics*, Vol XXIX, pp. 161-172.

Lemos, R., 2002, Security: What's Going On?, ZDNet News, 21 January, <<http://zdnet.com.com/2100-1105-819713.html>>.

Lynn, P., 2001, Developing Quality Standards for Cross-National Survey Research: Five Approaches, Working Papers of the Institute for Social and Economic Research, 2001-21, University of Essex, <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2001-21.pdf>

Measurecast, 2001, An Analysis of Streaming Audience and Measurement Methods, mimeo, 9 August, http://www.measurecast.com/docs/Audience_Measurement_Methods1.pdf

Mesenbourg, T.L., 2000, Measuring the Digital Economy, U.S. Bureau of Census, mimeo, <<http://www.census.gov/eos/www/papers/umdigital.pdf>>.

Mesenbourg, T.L., 2001, Measuring Electronic Business, U.S. Bureau of Census, mimeo, <<http://www.census.gov/eos/www/papers/ebusasa.pdf>>

Minges, M., 2000, Counting the Net: Internet Access Indicators, mimeo, <http://www.isoc.org/inet2000/cdproceedings/8e/8e_1.htm>

Odlyzko, A., 2000, Internet Growth: Myth and Reality, Use and Abuse, *iMP Magazine*, <http://www.cisp.org/imp/november_2000/odlyzko/11_00odlyzko.htm>

Organization for Economic Cooperation and Development, 1998, Internet Infrastructure Indicators, <<http://www.oecd.org/pdf/M000014000/M00014287.pdf>>

Organization for Economic Cooperation and Development, 2000, Measuring the ICT sector, <<http://www.oecd.org/pdf/M00002000/M00002651.pdf>>

Organization for Economic Cooperation and Development, 2001, Measuring Capital: A Manual on the Measurement of Capital Stocks, Consumption of Fixed Capital and Capital Services (OECD, Paris).

Pereira, A., 2001, Is all Public Capital created Equal?, Review of Economics and Statistics, 82(3), pp. 513-518.

Pereira, A., and R. Flores-de-Frutos, 2000, Public Capital and Private Sector Performance, Journal of Urban Economics, 46(2), pp. 300-322.

Schonlau, M., R. Fricker and M. Elliot, 2002, Conducting Research Survey Via Email and the Web (RAND, Santa Monica).

Solow, R. M. (1987), We'd better watch out, New York Times Book Review, July 12, p.36.

Stigler, S. M., 1999, Statistics on the Table. The History of Statistical Concepts and Methods (Harvard University Press, Cambridge).

United Nations (2000), The Millennium Declaration, <http://www.un.org/millennium/sg/report/>

United Nations Development Fund, 2001, Human Development Report: Making New Technologies Work for Human Development (Oxford University Press, Oxford)

Yook, S., H. Jeong, and A. Barabási, 2002, Modeling the Internet's Large-Scale Topology, Proceedings of the National Academy of Sciences, 99, pp. 13382-13386.