Turning Dirt Roads into Information Highways

The Conceptual Misformation of Internet Diffusion

Meelis Kitsing and Philip N. Howard

Contents

Abstract ... 3

About This Research ... 4
The World Information Access Project ... 4
Working Papers ... 4
Author Biographies ... 5
Acknowledgments ... 7
Review Panel ... 7

Introduction ... 9

Conceptual Misformation in the Social Sciences ... 12

Stretching the Concept of Internet Diffusion ... 14

Operationalization and Measurement of Internet Diffusion ... 22
Hosts as Indicator of Diffusion ... 23
Noncorrelation of Hosts and Users in the CEE ... 25
Noncorrelation of Hosts and Users in the World ... 26
Table 2. Ranking for Top Ten Countries for Internet Diffusion with Common Metrics ... 28
Users 2007 ... 28
Hosts 2007 ... 28
Weaknesses of Hosts as a Measure of Internet Diffusion ... 29
Internet Users as a Measure of Internet Diffusion ... 30
Internet Users Compared with Subscribers ... 31
Weaknesses of Survey Data on Internet Use ... 33
Contextual Effects on the Data on Internet Use ... 34
Summary ... 36

Measuring Effective Internet Diffusion ... 38
Table 3. Three Dimensions for Indexing Effective Internet Diffusion ... 40
Network Nature ... 40
Sophistication of Use ... 42
Social Distribution ... 42
The Index of Effective Internet Diffusion ... 45

Conclusion ... 47

References ... 50

Abstract

As we analyze the use of the concept of internet diffusion found in economics, political science, and public policy research in this paper, we find that the research on internet diffusion is often characterized by extensive errors in operationalization and measurement. These errors are not caused only by imperfections and inherent biases in commonly used datasets. Rather, they stem primarily from a conceptual stretching whereby the indicators used are rarely connected to the properties of the concept. The following factors account for the outcome:

- First, the nature of the internet and its diffusion is better understood as a network good, but researchers rely on data about internet connectivity as an independent good.
- Second, whereas the large international datasets report on internet hosts, internet users, and subscribers, these categories mean very different things from country to country and result in peculiar patterns in national ranking that are not generalizable.
- Third, there is no correlation between the different variables that have been used by scholars in measuring internet diffusion.

These factors result in inherent biases in measuring the diffusion outcomes; in addition, the use of different indicators considerably affects the primacy of a particular causal explanation over the alternatives. Researchers should develop explicit methodologies for analyzing internet diffusion. This avoids substituting a particular research technique for methodology, as well as using unrepresentative nominal variables.

Citation

Kitsing, Meelis, and Philip N. Howard. *Turning Dirt Roads into Information Highways: The Conceptual Misformation of Internet Diffusion*. World Information Access Project Working Paper #2009.2. Seattle: University of Washington.

About This Research

The World Information Access Project

The World Information Access Project (www.wiareport.org) is an international team of researchers dedicated to investigating global trends in technology distribution, information security, and personal privacy. The researchers look for practical ways to improve equity in information access and for ways to use communications technologies to improve the quality of our economic, political, and cultural lives. Please direct correspondence to:

Philip N. Howard
Director, World Information Access Project
Department of Communication
141 Communications Building, Box 353740
University of Washington
Seattle, Washington, 98195-3740
Telephone: (206) 543-2660, Facsimile: (206) 616-3762
www.wiareport.org

Working Papers

This working paper series supports junior scholars by providing stipends, modest travel allowances, and mentorship to translate innovative research into accessible manuscripts for circulation to journalists, policymakers, and scholars. Papers receive substantive feedback from at least three internationally prominent senior scholars or practitioners with expertise in the relevant field.

Author Biographies

Meelis Kitsing is a doctoral fellow at the National Center for Digital Government and a Ph.D. candidate in political science at the University of Massachusetts Amherst. His primary research interests lie in comparative and international political economy with a focus on technology issues. His doctoral dissertation focuses on both conceptual issues related to internet diffusion and empirical aspects of the diffusion as experienced by the countries of Central and Eastern Europe. He has presented his research at the annual conferences of the American Political Science Association (APSA), the International Studies Association (ISA), and the Association for the Study of Nationalities (ASN). Kitsing graduated with a Master of Arts in Law and Diplomacy from the Fletcher School at Tufts University and a Master of Science in Politics of the World Economy from the London School of Economics and Political Science. He has taught at Clark University, the College of the Holy Cross, Emmanuel College, the Estonian Business School, and the University of Massachusetts. He has been a researcher at Harvard University, an international policy fellow at the Center for Policy Studies of the Open Society Institute and Central European University in Budapest, and a research associate at the Stockholm School of Economics in Riga. He serves on the executive board of the Information Technology and Politics Section of the American Political Science Association. He has received Fulbright, Peacock, Hammer, Institute for Humane Studies, Templeton, and numerous other scholarships and fellowships for his studies and research.

Philip N. Howard (B.A. Toronto, M.Sc. London School of Economics, Ph.D. Northwestern) is an associate professor in the Department of Communication at the University of Washington. His current research and teaching interests include the role of new information technologies in the political communication systems of advanced democracies and the role of new information technologies in the social development of poor countries. He is working on a new book, *The Internet and Islam: Digital Origins of Dictatorship and Democracy* for Oxford University Press. He is the author of *New Media Campaigns and the Managed Citizen* (New York: Cambridge University Press, 2006), which considers how information technology is used by political elites to structure public opinion and political culture in the United States. The book was awarded the 2007 CITASA Best Book prize from the American Sociological Association and the 2008 Best Book prize from the International Communication Association. He has coedited, with Steve Jones, *Society Online: The Internet in Context* (Thousand Oaks, Calif.: Sage, 2004) and, with Andrew Chadwick, the *Handbook of Internet Politics* (London: Routledge, 2008). He is the author of numerous journal articles examining the role of new information and communication technologies in politics and social development, including pieces in the *American Behavioral Scientist*, the *Annals of the American Academy of Political and Social Science*, and *New Media & Society*. He has worked on several National Science Foundation projects, serving on the advisory board of the Survey2000 and Survey2001 Projects, comanaging a project about information and communication technologies in Central Asia and directing the World Information Access Project. Howard has been a consultant to the Canadian International Development Agency and has held fellowships at the Pew Internet & American Life Project in Washington D.C., the Stanhope Centre for Communications Policy Research in London, and the Center for Advanced Study in the Behavioral Sciences in Palo Alto.

Acknowledgments

This research was supported by the World Information Access Project (www.wiareport.org) and funded by the National Science Foundation under award IIS-0713074. All opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Authors would like to thank the Graduate School and the National Center for Digital Government at University of Massachusetts Amherst for providing support to present an early version of this study at the 2008 International Studies Association meetings. The authors are particularly grateful to the members of the review panel for their detailed feedback.

Review Panel

- Jeffrey Hart is Professor of Political Science at Indiana University, Bloomington, USA.
- Christopher May is Professor of Political Economy in Politics and International Relations at Lancaster University, UK.

- Helen V. Milner is B. C. Forbes Professor of Politics and International Affairs at Princeton University, USA.

Introduction

Despite its seemingly global nature, the internet diffuses with different extensity and intensity in different locations. Recent studies in economics, political science, and public policy have either implicitly or explicitly used the concept of internet diffusion to explain other changes in our economic, political, and cultural lives. This implies that the importance of the concept is not limited to studies exploring information technology issues such as the digital divide. Indeed, some recent scholarly articles and books have used internet diffusion as a broad measure of the knowledge economy. Furthermore, indicators of internet diffusion are often used in political rhetoric for the purpose of indicating the level of development of information society. It follows naturally that conceptualization and measurement have real-life consequences beyond academic debates. Development agencies use data gathered by international organizations to direct aid to developing countries to help bridge the digital divide. Governments in the developed world rely on the indicators to make public policy and investments for improving internet connectivity.

We argue in this paper that conceptual stretching has occurred in many studies about internet diffusion. Any meaningful discussion of the diffusion—particularly when it is used as a broad measure of a big concept such as the knowledge economy—must avoid concept malformation. It must rely on explicit operationalization and measurement that will not mishandle the background concept, because such stretching can lead to mistakenly categorizing dirt roads and information highways as the same concept.

This so because measuring diffusion of equipment and counting internet hosts or thin surveys that ask whether people have used the internet reveal very little about the nature and actual diffusion patterns of the internet (Slater and Miller 2000). This is especially so if only one set of data is used. The nature of internet use differs from one social context to another. Many large-N studies rely on limited datasets that assume away the context. Further, in addition to ignoring contextual aspects, these datasets suffer from several methodological flaws. Some studies use the number of internet hosts in a country as an aggregate measure of internet diffusion. Others use the number of internet users as a percentage of population as an aggregate measure. Those two measures do not correlate well.

Instead of relying on single datasets for measuring nominal and formal internet diffusion, we propose an index of "effective internet diffusion." By capturing the characteristics of both the quantity and the quality of internet diffusion, this index allows the contextual aspects to overcome the problems of single metrics of internet diffusion. Most importantly, the internet will be analyzed as a network good, not an independent good, which reveals network externalities that are fundamental for understanding the essence of the knowledge economy. The nature of these network externalities goes beyond the fact that the value of the internet depends on the network to which these technologies are connected. In addition to technological and physical networks, there are social networks in which the externalities come into play even more profoundly. Essentially, this stems from the understanding that the nature of the internet and its diffusion is epistemological. Both technological and social externalities are reinforced by the fact that the internet is by nature a decentralized network; that is, applications are hosted at the edge of the network by absolutely anyone. The internet is much less controllable than a smart network such as a telephone network, where applications are hosted in the network's core.

We critique existing approaches and propose some solutions that can contribute to the literature on technology diffusion in general and on the internet diffusion in particular, as well as to our understanding of the internet. The paper proceeds with a discussion of the concept misformation; this discussion constitutes the backbone of the operationalization and measurement issues considered in the study. We next offer an overview of the literature on the internet diffusion. This is followed by an in-depth analysis of the concept operationalization and measurement that references the scholarly literature and examines several empirical examples to discuss the shortcomings in measurement. The paper concludes by highlighting the different dimensions of the effective internet diffusion index and its potential indicators.

Conceptual Misformation in the Social Sciences

The concept of road as used in the title of this paper has very different meanings in different contexts. Some may understand the road to mean a four-lane highway. For others it may simply be a dirt road. Considering the highway and the dirt road as essentially the same and measuring how many roads countries have on the basis of this conceptualization is very unhelpful for someone who wants to travel and/or transport goods from point A to point B.

Similarly, the concept of internet diffusion has sometimes been misformed in the research on the digital divide. Yet the problem of conceptually stretching observations from qualitative and single-case studies to larger quantitative modeling research is common in the social sciences. Giovanni Sartori observed already in 1970 that comparative politics scholarship has engaged in “conceptual stretching” in an attempt to make concepts universally applicable (Sartori 1970). Sartori’s pioneering work on how extending concepts into unsuitable domains undermines the validity of research has become an important issue for methodologists (Munck 2004). This so because conceptual stretching is a theoretical problem with consequences for operationalizing research and data collection. This, in turn, implies that the measurements used in analysis are irrelevant or misrepresentative of the concept. “As a result, the very purpose of comparing—control—is defeated, and we are left to swim in a sea of empirical and theoretical messiness,” writes Sartori (1970).

Seawright and Collier (2004) echo Sartori's point by offering the following definition of conceptual stretching: "A form of measurement error that arises when scholars inappropriately apply established concepts and theories to new situations. Prior assumptions about the meaning of some components of the concept, and about the interrelations among these components, are not met in these new contexts."

Many concepts in social science research are difficult to reproduce as variables (George and Bennett 2005). Many large-N studies, which employ weak conceptulizations and limited data in order increase their sample size, ultimately run a high risk of conceptual stretching. Similarly, a qualitative research design that employs the quantitative methods as articulated in *Designing Social Inquiry* by King, Keohane, and Verba (1994) runs the same risks because such research design aims at increasing the number of observations in a qualitative setting (George and Bennett 2005; Munck 2004). Such quantitative and qualitative approaches may lead "to a meaningless togetherness based on pseudo-equivalences" (Sartori 1970).

The solution to this problem is to engage in "contextualized comparisons" that address the challenge of establishing equivalence by attempting to identify "analytically equivalent phenomena" (George and Bennett 2005). This implies that the same phenomena may be represented in different contexts in fundamentally distinct ways (George and Bennett 2005). The following sections of the paper will try to operationalize this insight in discussing more analytically valid ways of measuring the internet diffusion.

Stretching the Concept of Internet Diffusion

There is no uniform definition of internet diffusion in the literature. Often scholars use the terms internet diffusion, penetration, adoption, connectivity, access, use, and digital divide are used to describe the same phenomena. For example, the Internet Encyclopedia offers the following definition: “diffusion of an innovation is a macro process concerned with the spread of the innovation from its source to the public” (Dholakia et al. 2004). The adoption is defined as “a micro process that focuses on the stages through which an individual passes when deciding to accept or reject the innovation” (Dholakia et al. 2004). However, individual decisions to adopt or not to adopt the internet have clear consequences on the macro level. And diffusion on the macro level impacts individual decisions to adopt the internet. In practice, it is not possible to draw clear lines between adoption and diffusion, as these processes are mutually interdependent. Indeed, the same encyclopedia talks about “adoption and diffusion patterns of internet” without clearly distinguishing between the two (Dholakia et al. 2004). Hence, we see internet diffusion as a process indicating increased adoption and use of network technologies by growing numbers of people in their daily lives. As the adoption, use, penetration, and diffusion of the internet describe the same phenomena, then this discussion of the literature draws from all these different branches of the literature. The following review discusses both diffusion outcomes and causes of these outcomes. Discussion of both sets of variables can contribute to improved understanding of technology and its diffusion.

Many studies have focused on internet diffusion as a dependent variable and explained its variance in a large-N setting. Some scholars have focused on the internet diffusion outcomes as such (Beilock and Dimitrova 2003). Others have used it as a proxy for the digital divide (Milner 2006) or the knowledge economy (Inglehart and Welzel 2005). Several studies have outlined a strong correlation between the rate of per capita internet penetration and per capita gross domestic product (GDP) (Kiiski and Pohjola 2002, 2001). Norris (2001) sees the level of economic development as a key variable in explaining the different outcomes in internet diffusion. Kiiski and Pohjola (2001) point out that, in addition to income, the cost of internet access also helps explain the observed growth in computer hosts in per capita terms. Beilock and Dimitrova (2003) found a strong correlation between the level of infrastructure development (defined as main telephone lines per 100 inhabitants) and internet diffusion in addition to per capita income.

However, other scholars have emphasized the role of institutions and policies in explaining the outcomes of technology diffusion (Dasgupta, Lall, and Wheeler 2001; Milner 2006; Caselli and Coleman 2001; Coralles and Westhoff 2006). Such institutionalist approaches reveal shortcomings of both resource-based and technological determinist explanations. The latter would see the spread of the internet as the sufficient objective technology for spreading openness or increasing productivity (Jensen 2007). Contrary to this view, institutionalists stress the role of domestic rules of the game in facilitating the diffusion. The importance of political openness is supported, for example, by the work of Beilock and Dimitrova (2003). They found in their statistical study that openness of a society (defined as a noneconomic factor) was strongly correlated with outcomes in the internet diffusion (in addition to the level of infrastructure development and national income). Corrales and Westhoff (2006) have confirmed this finding by providing evidence that authoritarian regimes discourage internet use. They also point out that per capita income has a stronger explanatory power for internet diffusion in authoritarian countries (Coralles and Westhoff 2006). This implies that open political institutions can act to a certain degree as substitutes for the level of national income; that is, a wealthier authoritarian country and a somewhat poorer democracy may have similar levels of internet diffusion simply because the latter has a more liberal institutional framework in place. Similarly, Milner (2006) finds in her quantitative study of 184 countries that on average democracies have a higher rate of internet adoption than autocracies. This again implies that political institutions are key for understanding the diffusion. Milner argues that

> political institutions in particular matter for the adoption of new technologies because they affect the manner and degree to which winners and losers from the technology can translate their preferences into influence. Groups that believe they will lose from the internet try to use political institutions to enact policies that block the spread of the internet. These “losers” hope to slow down or stop its diffusion, and some institutions make this easier to do than others (Milner 2006).

According to Milner, authoritarian governments have most to lose from internet diffusion, and quite obviously they are more able than democracies to minimize its diffusion.

Scholars studying more specific institutions and policies have confirmed that institutions are more important than resources in explaining the digital divide—more so than a simple democracy-authoritarianism dichotomy would suggest. Dasgupta and others (2001) used econometric analysis to conclude that income differentials do not entirely explain the digital divide between countries. They reason that the digital divide is not a new phenomenon but rather reflects the persistent disparity in main telephone lines. Dasgupta et al. demonstrate that a state's competition policies matter a great deal, given that low-income countries with high World Bank ratings for competition policy can have surprisingly high rates of internet subscriptions per main telephone line (Dasgupta, Lall, and Wheeler 2001). The emphasis on competition is supported by studies on telecom regulation by Heimler (2000) and Wu (2008). In addition, the econometric study of eighty-six developing countries by Fink and others (2003) demonstrates that complete telecom liberalization pays off by increasing teledensity (which refers to the main telephone lines in per capita terms) by 8 percent. Caselli and Coleman (2001) find evidence that the larger the size of a government, the smaller the computer adoption rate across a country; they also have collected considerable evidence that the rate of computer diffusion across countries is associated with sound property-rights protection. Even though their study looks at personal computers, it is also relevant for understanding internet diffusion, because as John Gage, a cofounder of Sun Microsystems, once said, “the network is the computer” (Gapper 2007).

The common denominator of the studies discussed above is their reliance on indicators that are easily accessed through data bases of international organizations such as the World Bank and the International Telecommunications Union (ITU). A particular shortcoming seems to be that these studies do not incorporate the insights of earlier studies on technology diffusion that emphasize social shaping, social capital, the epistemological nature of technology, and informal institutions for understanding technology diffusion such as the work by Mackay and Gillspie (1992). Certainly, the factors are primarily independent variables, but bringing them into the analysis can help to construct a richer dependent variable as well. This literature encourages us to think of the internet as a network good, not an independent good, where social relations, existing technology infrastructure, and physical proximity can profoundly influence diffusion. Despite this more meaningful way of defining the internet, much of the data used by existing literature relies on singular properties for explaining diffusion. For instance, asking whether people have used the internet in the last six months or whether they subscribe to an internet provider is not be sufficient for drawing meaningful conclusions. It is important to know where they use the internet (home, work, and/or public places), what the purposes of the use are, and so on (concrete measures are developed in the last section of this paper).

Hence, our understanding of internet diffusion would benefit from linking it to and borrowing from a broader literature on technology and knowledge creation. Technology historian Joel Mokyr has outlined the role of the broader institutional environment in his historical study of technology adoption. He highlights the importance of openness to new information and "exposure effects" to new knowledge as crucial elements for technology diffusion (Mokyr 1990). Furthermore, his contribution to our understanding of the role of informal institutions and the role of knowledge in the facilitation of technology adoption can be linked to studies of social capital. Ghosal and Nahapiet (1998) point out that social capital facilitates knowledge creation and that more dense social capital creates “organizational advantage.” Ghosal and Tsai (1998) demonstrate how fundamental social capital is for understanding the value creation within the intrafirm networks of a multinational electronics company. Honig and others (2006) explored the importance of social capital in creating linkages between small high-tech firms and the military—essentially, some of the most dynamic, most hierarchical organizations in the world. Sociological studies show interactions of social networks with markets and organizations (White 2002; Wasserman and Faust 1994; Eccles and Nohria 1992) and their contributions to businesses in Silicon Valley (Castilla 2003) and to such hierarchical forms of business units as call centers (Castilla 2005). Even though these studies do not focus on the diffusion of internet or technological equipment, they matter for understanding technology diffusion. The diffusion is not simply about the spread of technical equipment (such as personal computers or connections at particular speed) but is also dependent on accompanying ideas about how to use the technology. Certainly we cannot measure people's ideas but we can have a richer understanding of the purposes to which they put technology and where they do so. Since the users of technology interact socially, these social interactions create network effects, as the recent triumph of Web 2.0 type technologies has clearly demonstrated. However, the evolution and social shaping of these networks is ultimately dependent on context. This is well demonstrated by the frequency, intensity, and purpose of technology use in different countries.

The exclusion of the network good aspect of internet diffusion in the research is not compensated by sophisticated use of statistical methods if the dependent variable is operationalized on the basis of a single dataset. Employing different datasets for operationalizing the dependent variable would certainly do a better job of capturing the network aspects of the diffusion. Essentially, having a large-N output of these studies does not necessarily mean that the result is improved understanding of how the internet actually diffuses on the ground. Even the data that are collected on the use of internet carry very different meanings in the diverse contexts, despite the fact that these different meanings may still be represented by one and the same data point in the quantitative analysis. Qualitatively, the difference between internet use may vary so significantly that use in one country may carry completely different meaning from use in other country. For example, the data on use are often collected by asking whether people have used the internet in the last six months. Thus the same data point would represent a person who used the internet once in a library during this period and a person who uses the internet every day for 10 hours at home and office. Obviously, there is a fundamental difference between these two users that is not captured in the analysis. The next section of the paper builds on these insights to explore how the concept of internet diffusion is captured in the various operationalizations and measurements.

Operationalization and Measurement of Internet Diffusion

Usually, the concept behind a metric of internet diffusion is to determine how widely the internet is used. The concept is typically not measured in absolute terms, as the total number of internet users in a country, but in relative terms, as an indicator weighted by population size: internet use per capita. This approach establishes equivalence by taking into account a single specific context—population size in a country (Adcock and Collier 2001). In addition, given how much research suggests that economic wealth is an important predictor of technology diffusion, it might make more sense to weight internet access per dollars of national income. Most important, standardizing by population is useful because it avoids effects that are the result of population size (Adcock and Collier 2001; Jacob 1984). There are two standard ways of measuring internet diffusion. First, scholars measure the number of internet hosts as a percentage of population or of a specific numerical group in the population, such as per 10,000 inhabitants (Kiiski and Pohjola 2001) or per 1,000 inhabitants (Inglehart and Welzel 2005). This approach implies that internet hosts can be applied as a proxy for the internet diffusion. Other scholars prefer measuring the number of internet users per 10,000 inhabitants (Beilock and Dimitrova 2003) or per 1,000 people (Coralles and Westhoff 2006) . This approach implies that compared with a country with 5 hosts per 10,000 inhabitants, a country in which there are 50 hosts per 10,000 people has an internet that is 10 times faster, larger, or better in some way.

Good examples of conceptual stretching are found in the datasets and methodological notes of the ITU. Some scholars have used this dataset, whereas others s have preferred different datasets such as the World Development Indicators. But the ITU dataset is representative of some fundamental issues that scholars have to face when they try to operationalize internet diffusion and develop measurement variables. The measurement issues are discussed on the basis of a sample of Central and Eastern European (CEE) countries by comparing different indicators in 2001 and 2007. This limited comparison of fairly similar countries in the broader context of the world permits some basic analytical points to be made. In addition, the findings about the CEE are compared according to the measure of the top 10 countries in the world in 2007.

Hosts as Indicator of Diffusion

The term "internet hosts" refers to organizations or firms that have computers directly linked to the worldwide internet network. For instance, an Internet Service Provider (ISP) serves as the host through which individuals can connect to the internet via a modem connected to a telephone line, a cable, a wireless link, or a satellite link. ITU identifies hosts by a two-digit country code, for example, France is .fr, the United Kingdom is .uk., and so on, or by a three-digit code referring to a specific classification of organization, for example .org, .com, .edu, and so on (ITU 2006). Data come from the Internet Software Consortium and Reseaux IP Europeens. This method is a reliable means of measurement because errors in collecting the data are minimal and, from a technical standpoint, data are easily accessible (for a discussion of reliable data collection, see Jacob 1984).

Problems do arise with content validity, however. This is broadly related to the issue how top-level domain names are distributed internationally: .com can be used by any company in the world, just as .eu can be used by any country in the European Union. The bottom line is that this method of measurement of internet hosts does not necessarily tell whether a counted host is physically located in a certain country. Further, the method of measurement changes over time, different international organizations and national agencies use different methodologies, and hosts do not measure accessibility (Minges 1999). Identifying the geolocation of hosts is possible but is a difficult process (Zook 2006). In light of these fundamental challenges in using hosts as an indicator of internet diffusion, it is quite surprising that many scholarly studies rely on these data as the sole dependent variable. As Michael Minges of the ITU pointed out almost ten years ago, hosts are the "most widely "misused" indicator of internet access" (Minges 1999). More recently, the ITU has been clear in its discussion of data-collection methodology that the indicators are simply an "approximation" (ITU 2006). More recent data from the ITU (2008) exclude internet hosts as an indicator.

The shortcoming of using hosts as a measure of internet diffusion is particularly true of hosts offering services under internet names ending with .com or .org. Discriminant validity is missing, as the measurement does not differentiate between different types of hosts, for example, internet hosts based in the domestic economy and those based outside. Nor is convergent validity present because internet users and internet hosts do not correlate well. The latter is a fundamental flaw in using hosts as an indicator of internet diffusion because it would allow researchers to manipulate the dependent variable by picking a dataset that presents quite different outcomes in the dependent variable than would be the case with alternative datasets.

Noncorrelation of Hosts and Users in the CEE

In order to demonstrate this lack of correlation, we look at a sample of countries from the same region and at top countries with highest user and host penetration globally. Table 1 demonstrates the lack of correlation simply on the basis of a sample of the CEE countries. If we look at the two years of comparative data about hosts and users, it is obvious that the correlation between the two datasets is highly imperfect. For example, comparing the ranking of hosts per 100 inhabitants in 2001 with the ranking of users per 100 inhabitants in 2001 shows clearly that all countries change their respective positions in the list, with the exception of Estonia and Slovakia, which remain numbers 1 and 5, respectively, on both lists. The nonexistent correlation is even more remarkable in 2007. Only Estonia ranks number 2 on both hosts' and users' list; all other countries change their respective position on the two lists. Belarus and Romania, which make it to the top of the list on users, do not even make it into the top 10 by measuring by hosts. Lithuania is number 1 by hosts but ranks number 9 by users. The Czech Republic ranks 10 by users but number 4 by hosts. Slovenia is number 5 by users but number 10 by hosts. These are huge variations. Hence, if only one of the indicators were used for the operationalization of the dependent variable, then quite different outcomes and explanation could emerge as to why the internet diffuses differently in the CEE.

Table 1. Ranking Internet Diffusion in Central and Eastern European Countries with Common Metrics

Rank	Users 2001	Subscribers 2001	Hosts 2001	Users 2007	Subcribers 2007	Hosts 2007
1	Estonia (31.67)	Slovenia (14.24)	Estonia (3.57)	Belarus (61.93)	Croatia (29.07)	Lithuania (36.49)
2	Slovenia (30.51)	Czech R (12.25)	Czech R (2.11)	Estonia (58.41)	Russia (21.40)	Estonia (29.62)
3	Czech R (14.63)	Croatia (7.37)	Hungary (1.68)	Romania (52.24)	Estonia (21.28)	Croatia (24.74)
4	Hungary (14.51)	Estonia (7.06)	Slovenia (1.48)	Latvia (51.68)	Slovenia (20.86)	Czech R (23.81)
5	Slovakia (12.48)	Hungary (3.50)	Slovakia (1.35)	Slovenia (49.57)	Czech R (15.88)	Hungary (23.29)
6	Croatia (11.52)	Poland (2.41)	Poland (1.27)	Croatia (43.57)	Lithuania (15.14)	Slovakia (15.07)
7	Poland (9.84)	Latvia/ Slovakia (1.85)	Latvia (1.06)	Slovakia (42.89)	Hungary (14.87)	Poland (14.76)
8	Bulgaria (7.62)	Latvia/ Slovakia (1.85)	Lithuania (0.96)	Poland (42.01)	Ukraine (13.85)	Latvia (10.42)
9	Latvia (7.21)	Lithuania (1.69)	Croatia (0.50)	Lithuania (39.33)	Macedonia (13.42)	Bulgaria (7.07)
10	Lithuania (7.18)	Macedonia (1.49)	Bulgaria (0.33)	Czech R (39.27)	Romania (10.05)	Slovenia (6.69)

Source: Based on authors' calculations from ITU (2003, 2008) and CIA (2008).
Note: The rankings of selected CEE countries in internet users, subscribers, and hosts are from 2001 and 2007. Countries with the highest number of users, subscribers, and hosts per 100 inhabitants have higher positions in the list.

Noncorrelation of Hosts and Users in the World

The previous discussion of data reveals numerous problems with the reliability and validity of data from the CEE. However, these issues, such as the imperfect correlation between hosts and users, are not unique to CEE countries, but hold globally as well. A simple comparison of per capita hosts and users in the top ten countries by each indicator supports this claim. Table 2 demonstrates the lack of correlation. Hosts and users also show imperfect correlation in Figure 1a and Figure 1b of Milner's (2006) research on 184 countries. In Figure 1a, Milner correlates users with regime type (democracies and autocracies) from 1991 to 2000. In Figure 1b she correlates hosts with regime type from 1994 to 2000. Democracy and internet diffusion are correlated more strongly if diffusion is measured by hosts rather than by users.

In Table 2 all countries change their positions on the respective lists and there is no correlation between users and hosts. Even more striking, that six countries in the list of top ten users (Greenland, Andorra, Canada, Bermuda, Greenland, Andorra, Canada, Bermuda Luxembourg, South Korea Luxembourg, South Korea) do not even make it to the list of top ten hosts tells us that something more fundamental is going on. Similarly, six countries (Tuvalu, Iceland, Denmark, Finland, Australia, and Lithuania) on the list of top ten hosts do not appear on the list of top ten users.

The number 1 rank of Tuvalu, which has 248 hosts per 100 inhabitants in 2007, may be the result of the country becoming a "host farm" because of the popularity of domain names ending with .tv. However, this does not mean that the entire Tuvalu population benefits from the high host penetration. This may be an extreme case, but it is a striking example of the inaccuracy that can result from using hosts as a measure of internet diffusion or as a proxy for something more abstract, such as the digital divide or the knowledge economy.

Table 2. Ranking for Top Ten Countries for Internet Diffusion with Common Metrics

Rank	Users 2007	Subscribers 2007	Hosts 2007
1	Netherlands (91.36)	Bermuda (58.82)	Tuvalu (248)
2	Greenland (90.75)	Sweden (44.46)	United States (92.82)
3	Norway (80.88)	Singapore (43.69)	Iceland (89.02)
4	Andorra (79.08)	Hong Kong (39.77)	Netherlands (67.11)
5	Canada (76.77)	Denmark (38.61)	Denmark (56.78)
6	Sweden (76.76)	Switzerland (37.59)	Australia (53)
7	Bermuda (74.42)	Netherlands (36.55)	Norway (44.87)
8	Luxembourg (73.95)	Andorra (36.46)	Finland (44.29)
9	South Korea (73.80)	Iceland (35.20)	Sweden (36.68)
10	United States (71.94)	Norway (34.19)	Lithuania (36.49)

Source: Based on authors' calculations from ITU (2008) and CIA (2008).
Note: The rankings of top ten countries in the world either by internet subscribers or users are from 2007 (except Australia, where host data are from 2008). The countries with highest number of users, subscribers, and hosts per 100 inhabitants have higher positions on the list.

South Korea offers a more mainstream case than Tuvalu. Its internet users' penetration rate was 73.8 percent in 2007, while its hosts' penetration rate was only 0.7 percent (CIA 2008). During the same year, the US internet users' penetration was 71.94 percent, while its hosts' penetration was 92.82 percent. In other words, South Korea achieved the same outcome in terms of per capita users with the number of per capita internet hosts being 133 times smaller. The United States ranks number 10 by the measure of users but number 2 by the measure of hosts. As the United States is literally the focal point of the information highway, the difference between its ranking in hosts and users in 2007 reveals that outside users may be an important factor in the demand for its host services. Indeed, there is a methodological flaw, as the US internet total host count includes the following top level domain host addresses: .us, .com, .edu, .gov, .mil, .net, and .org (CIA 2008). Obviously, these domains are used worldwide and hosts do not have to be physically located in the US to use these domain names. Hence, the radical difference in the US position on the two lists can be explained at least in part by the asymmetric role that the US plays in the supply side of the internet. Nevertheless, scholars could again easily manipulate the outcomes in the variable of per capita internet diffusion by using either users or hosts for any discussion of the US standing in the world. This simple fact should set off alarm bells among social scientist and convince them of the need to avoid any simple measurements of internet diffusion by looking at only one variable.

Weaknesses of Hosts as a Measure of Internet Diffusion

The bottom line of these incongruities is that the number of hosts in per capita terms as an indicator takes on different meaning in the context of different countries. Any researcher with area knowledge of a particular country can read these large datasets and find incongruities. This implies that the supply-equals-demand approach does not work because the way hosts are measured is inadequate for identifying the internet users and the extensity and intensity of internet diffusion in any given country. Empirically, the weakness of using hosts as a measure of internet diffusion is supported by looking at the number of internet users per 100 inhabitants and comparing these figures with the number of internet hosts per 100 inhabitants. Countries with a high number of internet users may actually have a lower number of internet hosts than countries with a lower number of internet users. These differences reflect the fact that it is possible for a user to be a supplier of internet services for other users and use the internet without becoming a host itself but by using an already registered host. The differences in the number of hosts and users per 100 inhabitants as demonstrated in Tables 1 and 2 are broad indicators of structural impediments on the supply side. Certain factors such as telecom regulations and/or high costs keep the barriers for market entry high and discourage host activity on the internet. The competitive environment in the market in general and in the telecom sector in particular could be analyzed to using hosts as a meaningful measure.

Overall, this implies that using the number of internet hosts per 100 inhabitants is a weak operationalization of the concept of internet diffusion. The indicator is imperfect because the fit between per capita internet hosts and the concept of internet diffusion defined as a percentage of internet users in society is not close (for a conceptual discussion of validity, see Jacob 1984, 16; and Adcock and Collier 2001, 538–39).

Internet Users as a Measure of Internet Diffusion

If the internet hosts are an indicator of the supply side, then, logically, this would imply that internet users are an indicator of the demand side of the internet. The number of users per 100 inhabitants is usually recorded by calling people and asking whether they used the internet during a specific period (for example, last year, last six months, and so on). As the operationalization of internet diffusion, such measurement fares better in validity than measuring hosts. Indeed, Minges points out that users are a “better indicator of diffusion than hosts” (Minges 1999). This is so because the measure discriminates between internet users and nonusers in a given country. At the same time, the approach based on internet hosts measured diffusion indirectly by making a number of assumptions, for example, that supply equals demand.

Yet the measurement of users scores much worse in terms of reliability. This is the case despite of the fact that data have improved over time as market researchers, statistical agencies, and other organizations that compile user statistics have become more sophisticated in their data-collection techniques. And even when specific issues in measurement are taken into account, internet users as an indicator do not score very well in validity either. Convergent validity (correlation with hosts and subscribers) is missing (Tables 1 and 2).

Internet Users Compared with Subscribers

As the extensive discussion of noncorrelation of hosts and users was already given above, the imperfect nature of the data on users is further demonstrated by comparing it with the data on subscribers. The ITU defines subscribers by the following way: “The number of total Internet subscribers with fixed access, which includes dial-up, total fixed broadband subscribers, cable modem, DSL internet subscribers, other broadband and leased-line internet subscribers. Only active subscribers who have used the system within a reasonable period of time should be included. This period (e.g., three months) should be indicated in a note”

(ITU 2007). The data on subscribers are not important only for their own sake for demonstrating noncorrelation with users; ITU also uses these data on subscribers to derive the number of users for countries where the survey data are not available (ITU 2007). Evidently, such categorization and collection of data does not indicate how many users may exploit the same subscription to the internet, and it excludes wireless connectivity as an important factor. As a result, the data are heavily biased toward countries where fixed-line access dominates wireless access.

As both Tables 1 and 2 demonstrate, there is no correlation between subscribers and users. All of the data from 2007 and 2001 in the CEE, as well as from 2007 for the rest of the world, show that all countries have different rankings depending on whether one looks at users or subscribers. Furthermore, the newest data from 2007 reveal several unexpected results that lead one to question the accuracy of the data collection. Most interestingly, Belarus has more Internet users per capita than any other country in the CEE. If the ITU data are correct, then Belarus has gone through a rapid, unprecedented internet revolution in which its penetration grew from 35 percent in 2005 to almost 62 percent in 2007. Certainly scholars should explore the causal factors behind such unnoticed growth in Europe's last dictatorship—especially as a look at the country's websites seems to indicate that Belarus still lives in the last century. In addition, the ITU reports that Belarus had only 0.39 subscribers per 100 inhabitants in 2005 and 4.2 subscribers in 2007. So it must be a wireless revolution and/or highly effective use of few subscriptions by the majority of people that has led to the tremendous growth in the internet use. Another anomaly is Russia, which has the second highest number of subscribers in per capita terms in the CEE (Table 1). Even more striking, Russia has more subscribers (21.40) than users (21.05) per 100 inhabitants (ITU 2008). This clearly distinguishes Russia from the rest of CEE, where the opposite correlation is true. Romania is third by users but tenth in subscribers.

Clearly, there is no justification for deriving the data on users from subscribers, as the ITU does. In addition, the high ranking of internet users in Belarus and Romania seems suspicious in light of the general state of the internet there and in light of some other studies of internet use and services in those countries, for example, the recent United Nations E-Government Survey (UN 2007).

Weaknesses of Survey Data on Internet Use

The ITU points out that the surveys differ across countries by the age groups they include and the frequency of use they cover (ITU 2007). Using loose definitions of users makes any comparison of these indicators across countries problematic at best. Different organizations produce a very wide variation in their estimates. Conceptually speaking, counting users by such thin and imperfect measures tells very little if anything about accessibility. This is particularly worrisome for researchers who are concerned about the digital divide and who might operationalize this concept by looking at the number of internet users as a percentage of population in a given country. Using surveys to ask whether inhabitants of a country have used the internet in the past six months, for example, fails to discriminate between fundamentally different types of users. A user who uses the internet once day for an hour or so is fundamentally different from a user who uses it for 10 hours every day. And of course grouping together someone who uses the internet once a week or once a month with heavy internet users—failing to distinguish between such different types of users—is profoundly problematic. These conditions create systematic error and bias for any cross-country analysis on the basis of ITU data and thereby undermine measurement validity.

Furthermore, the telephone surveys ask questions of people who have landlines, are eager to answer, and spend more time at home; that, too, creates systemic selection bias. People without phones and without internet cannot answer these surveys. This would show up in survey results by increasing the number of internet users.

Contextual Effects on the Data on Internet Use

Answers to surveys are based on people's perception and memory, which is selective. If society values internet use, people may actually lie and report that they have used the internet even if they have not. People may also respond differently to essentially the same question asked in a different way; this in turn creates context-specific problems that arise from comparing data across different countries where different organizations have been collecting data. Moreover, the respondents are often asked if they have "ever" used the internet. Many people can respond positively to this question: people who used the internet once several years ago, as well as people who use the internet daily. Moreover, this question increasingly suffers from the social desirability effect. In many countries, internet use is a symbolically modern activity and an important part of contemporary cultural consumption, so respondents are likely to overreport their internet use in an effort to appear modern. Hence, survey design choices have a big impact on the reliability of respondent answers.

Even within one country different subgroups of the population may have different response styles, and, thus, contextual specificity may challenge the validity of the results (Adcock and Collier 2001). The reliability of operating procedures of organizations and work habits of its employees may skew the results further (Jacob 1984, 35). The data on internet users may be also manipulated if the organization collecting the data has a contractual relationship with a domestic agency seeking to present the particular country in a favorable light in international arena (Jacob 1984, 40). For example, this may be a serious problem in dictatorships such as Belarus and be the reason for anomalies explored above. Once the data are collected from different national organizations by a large international organization such as ITU, then the results may be even more skewed because of the number of steps involved and the likelihood of human error. Often ITU corrects the data of previous years. This should be considered in cross-country analysis (Jacob 1984, 38–39).

Summary

In sum, the research on internet diffusion is misformed in four respects. First, several studies are misinformed by using hosts as a sole measure for the operationalization of the concept of internet diffusion. As this analysis demonstrates, hosts do not necessarily indicate the diffusion of the internet in society. Second, many studies that use internet users as a sole measure of internet diffusion ignore the highly imperfect nature of the data. Low validity and reliability of data-collection techniques makes cross-country comparisons imperfect at best and highly suspect at worst. Third, most research on internet diffusion relies on either hosts or users as a sole measure of internet outcomes. But these two variables do not correlate, making any findings highly suspect. One causal explanation may gain supremacy over an alternative explanation simply because a different operationalization of the dependent variable is used. Fourth, studies employing both per capita hosts and per capita users as a measure of internet diffusion, as in the research of Milner (2006), signal a way ahead. Nevertheless, such an approach also does not seem to be sufficient in light of the previous discussion, as both standard approaches for measuring internet diffusion have shortcomings in validity and reliability. Internet diffusion should be operationalized by using richer datasets. Such a dataset would still contain indicators of hosts, users, and subscribers, but additional variables would be included. Clearly, however, the data would still be imperfect. Indeed, Sir Josiah Stamp, a founder of economic statistics once reminded us that "governments are very keen on amassing statistics—they collect them, add them, raise to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn well pleases (Kay 2008)."

However, data seems more trustworthy when several sources of data make for consistent observations. For instance, by analyzing Table 1 it is clear that in 2007 Lithuania ranks highest by hosts but is low by users, whereas Belarus ranks highest by users but does not even make it to the top ten by hosts. Both countries rank low by subscribers: Lithuania is sixth and Belarus is not even among the top ten. At the same time Estonia ranks number 2 by both users and hosts and number 3 by subscribers. If the index were based on these three different datasets, then Estonia would have the highest per capita Internet diffusion among these countries. This is clearly not the case if only one of the datasets is picked for analysis. Last, the approaches relying on single datasets treat the internet as an independent good and do not do a good job reflecting the nature of the internet as a network good. The following section tackles some of these challenges and offers an imperfect solution.

Measuring Effective Internet Diffusion

The proposed solution is to create an index for measuring outcomes in per capita internet diffusion rates, as such an index would incorporate additional contextual factors that should increase analytical equivalence. It is called the index of effective internet diffusion because the aim is to reveal the intensity and extensity of internet diffusion on the ground. It is different from a formal internet diffusion index where only one measure is used (for example, hosts, users, or something else). In this sense, it is similar to debates concerning the operationalization of the concept of democracy by counting elections. At the same time, others hold that incorporating additional variables is an essential part of any measurement of democracy.

This approach borrows from Putnam, who used the index for measuring institutional performance in Italy (Putnam 1993) and from the analytical framework of the Global Diffusion of the Internet Project (Press et al. 1998). Press et al. (1998) proposed to use countries as the unit of analysis and investigate internet diffusion along six dimensions: pervasiveness, geographic dispersion, sectoral absorption, connectivity infrastructure, organizational infrastructure, and sophistication of use. These dimensions are modified and some new ways are developed to measure them. Press and others use the concept of pervasiveness, which refers to a degree to which people who are not technology experts use the internet, but this concept is not relevant for the index, in particular, since their analytical framework was published ten years ago. In the early days of internet use, this may have been an important criterion, but in the interim the widespread use of internet by diverse populations has rendered this dimension obsolete. In addition, this dimension can be captured by sectoral absorption and sophistication of use (see below). However, five other dimensions are important in capturing the extensity and intensity of internet diffusion. The index incorporates these five different factors, expands on these dimensions, and adds additional variables to the three dimensions of internet diffusion: network nature, sophistication of use, and social distribution. The index also incorporates insights from Barzilai-Nahon (2006), who distinguishes between focused monotopical indices and comprehensive indices of digital divide. The Network Readiness Index of the World Economic Forum (WEF 2008) is perhaps the most well known attempt to provide a comprehensive index for measuring the ICT. However, its somewhat narrow nature and various methodological flaws where several properties of the index are also their function still places it under the monotopical index.

In her theoretical contribution Barzilai-Nahon proposes a six-dimensional index consisting of infrastructure access, affordability, use, social and governmental constraints, socio-demographic factors, and accessibility (Barzilai-Nahon 2006). The index of effective internet diffusion includes these factors within each of our three dimensions. By relying heavily on the previous attempts to compose monotopical indices, Barzilai-Nahon seems to be treating the internet as an independent good in the integrated index. In comparison with her contribution as well as with that of Press et al. (1998) and the World Economic Forum (WEF 2008), the index of effective internet diffusion relies on conceptualizing the internet as a network good and highlighting the network nature of its diffusion. Table 3 identifies three important dimensions of the measurement of effective internet diffusion, along with potential indicators for geographic dispersion, sectoral absorption, and levels of infrastructure connectivity.

Table 3. Three Dimensions for Indexing Effective Internet Diffusion

Dimensions	Techniques for measurement	Some potential indicators
Network nature	Quantitative, qualitative, network analysis	Asymmetrical or symmetrical distribution between cities and rural areas, quality and quantity of connections to hubs, intensity (what for) and extensity of use (percentage) in different sectors: types of connectivity by dial-up, broadband, wireless and mobile users, number of internet hosts, number of internet providers, degree of independence of telecom regulator, and market openness.
Sophistication of use	Quantitative and qualitative	The purposes of internet use, availability of services and content, nature of discourses about the internet, search skills, language skills.
Social distribution	Quantitative, qualitative, network analysis	Percentage of males, females, young, old, different ethnic groups in using the internet, differences in use of these groups, connections between the different groups, gini coefficients for technology distribution.

Network Nature

The network nature of internet diffusion captures geographical dispersion, sectoral absorption and connectivity, and organizational infrastructure. Geographical dispersion refers to the geographical distribution of internet use. Is it concentrated only in the main cities or in certain clusters or is it widely dispersed across all geographical areas? This geographical analysis borrows from network theory by identifying the network hub and how different network nodes are connected to the hub in a given country.

It analyzes the extensity and intensity of these connections and is interested in whether this geographical network is characterized by asymmetrical or symmetrical distribution.

Sectoral absorption refers to the degree of internet use in different sectors. Is it heavily used in some sectors such as IT and banking or is the use widespread? What are the purposes of the use? Connectivity infrastructure refers to international and intranational backbone bandwidth and access methods of internet. How widespread is broadband connectivity? How widespread is wireless? What is the percentage of dial-up users? It would use variables such as per capita telephone lines (lines are needed to connect to the internet) and businesses with internet connections.

Organizational infrastructure provides some idea of the prevailing supply-side conditions in the market of the internet access. This is separate from infrastructure connectivity because connectivity could be provided by a monopolist provider. Or it can be provided by a market with structure that is monopolistic, oligopolistic, or a perfect competition. Internet hosts, number of ISPs, and other indicators of the degree of openness and competition in the provision of internet and telecommunication services would be used. Here again, the network analysis can be used for mapping out the network of ISPs. This would enrich the understanding of the diffusion of the internet as a network good, not an independent good.

Sophistication of Use

Sophistication of use is the measure that was operationalized by Press et al. (1998) in functionalist terms by simply looking at whether the internet is used as a substitute for traditional communication (such as writing letters and making phone calls) or in a new, innovative ways. This means looking at businesses with webpages, individuals with personal webpages, government use of the internet, the availability of interactive services, and the intensity and extensity of using content provided by the internet. However, in incorporating the indicators of this dimension, the index would go beyond their functionalist approach. The epistemological nature of technology and the importance of context must be carefully considered here. The provision of content and its use may differ across different countries, not because of technical factors but because of cultural and social factors. The sophistication of use could aim to capture the nature of the discourse about the internet occurring both online and offline. If the perception of the internet is positive in a particular country, for example, then it may influence how people answer survey questions about internet use.

Social Distribution

In addition to these two dimensions, an important third variable should cover the distribution of internet access by important social categories. Technology resources are not evenly distributed among countries and peoples. One common way of measuring how evenly a resource is distributed is through gini coefficients (Milanovic 2005; Atkinson 1970; Berrebi and Silver 1985). In a perfectly equal society, 23 percent of the population controls 23 percent of the resources, and 90 percent of the population controls 90 percent of the resources. A more equal society will have a low gini coefficient, and a society in which resources are highly concentrated will have a high gini coefficient. Gini coefficients for internet hosts, mobile phones, personal computers, secure servers, and internet users could reveal how equally they are distributed, and some summation could represent the overall inequality that is the digital divide. Gini coefficients could range from equal IT distribution across a sample of countries (valued at 0.00) to a condition of complete inequality in which all IT resources are held by one country (valued at 1.00).

For example, the mathematical expression for a gini coefficient of the distribution of mobile phones among the countries of world in a given year is Expression A, where μ_i represents the mean unweighted average number of mobile phones in each country and 1/n represents the weight of each country (n being the number of countries). All countries are ranked by the number of mobile phones within the country, so that $y_j > y_i$, and the relative difference between the number of mobile phones in two countries $y_j - y_i / y_i$, is weighted by the product of the technologically poorer country's (i^{th}) share in the world's mobile phones and the technologically advanced country's (j^{th}) share in the world's mobile phones. This formula renders a measure of the world's mobile phones distributed among a sample of countries.

Equation 1: Gini Coefficient for ICT Access by Country

$$Gini_T = \frac{1}{\mu_i} \bullet \frac{1}{n^2} \sum_{i}^{n} \sum_{j>i}^{n} (y_j - y_i)$$

Since countries vary quite significantly in terms of population size, it would be more accurate to weight the gini coefficient for the distribution of mobile phones by the relative portion of the world's population residing in each country. This expression allows us to weight the relative contribution of countries with different populations and many mobile phones, such as the United States and Belgium (a large and a small country with many mobile phones), with Indonesia and Laos (a large and a small country with few mobile phones). In other words, this allows for a more reasonable fit between large and small countries on both sides of the global digital divide.

Equation 2: Gini Coefficient for ICT Access by Country, Weighted by Population

$$Gini_T = \frac{1}{n^2} \sum_{i}^{n} \sum_{j>i}^{n} (y_j - y_i) p_i p_j$$

Using these expressions to compute the concentration of mobile phones, personal computers, or internet hosts (adjusted for geographical location) would reveal much about how well diffused a technology is by categories of age, education, or income. Indeed, economists often use gini coefficients to measure the distribution of wealth in a society and find that it can be highly concentrated. Since technology resources can also be concentrated within countries, any reasonable metric for internet diffusion must include its distribution across known categories of social inequality.

The Index of Effective Internet Diffusion

This approach to measuring the dependent variable would facilitate minimizing the problems concerning validity and reliability. Ultimately, this approach gives a more meaningful dependent variable than per capita internet diffusion measurements, which really measure what we call "formal internet diffusion." The index of internet diffusion is in essence a new variable representing "effective internet diffusion" that captures intangible elements of technology (Table 3).

The effective internet diffusion captures not only the quantity of internet use but also the quality of its use. This operationalization of the background concept of effective internet diffusion is a more approximate means of measurement than simply looking at the number of hosts or users. As there are substitutes and complements to the internet, the index demonstrates the actual use and diffusion of the internet in society more closely than would be achieved by looking at hosts or users. The number of hosts can be significantly skewed, and there is no symmetry between supply and demand, as we demonstrated above. To a great extent the index would still rely on the singular variables such as data on users, subscribers, and hosts, all of which have serious shortcomings, as discussed above. However, using richer datasets would reduce anomalies and cancel out biases. Obviously, for example, a country with very high host count but low subscription and user rates would not rank high in the index. The addition of further indicators will reduce potential anomalies of data on users and subscribers as well.

Furthermore, measurements of per capita internet penetration rates achieved by looking at either hosts or users do not capture the epistemological nature of technology. Even if the internet is based on the same or fairly similar tangible technology everywhere, the use of this technology may differ in different societies. For instance, it may be the case that in a wealthy society, the internet may be available to many, but a majority of population may use it for low-quality activities. At the same time, in a poorer society where the relative costs of using the internet are higher than in a wealthy society, the use of the internet may be characterized by significantly higher quality.

Conclusion

This review of some major studies on internet diffusion and the analysis of some datasets show that weak operationalization is the main reason for biased outcomes in the dependent variable. These biases obviously alter any causal explanations for internet diffusion by increasing the relative importance of some variables over others depending on the dependent variable employed by researchers. This finding cannot be dismissed by blaming it on the imperfections of the data or particular datasets or studies employed in this analysis. Certainly, future research could be improved by incorporating more studies and different datasets. However, data from many of the sources reviewed here will always remain imperfect.

The nonexistence of correlation between different measures of internet diffusion certainly reduces the validity and reliability of such indicators but the issue at stake is much more fundamental. Weak or nonexistent concept formation often leads to the conceptual stretching that undermines the entire research methodology. This implies that before starting to measure something, the operationalization of the concept must be made explicit, because “there is no methodology without logos, without thinking about thinking” (Sartori 1970). More explicit concept formation, operationalization, and development of measures can connect the properties of a digital divide metric with the theoretical concepts behind the digital divide.

The implications of improved concept formation and measurement are not limited to academic research. Indicators of internet diffusion are often used in the political domain. For instance, the number of internet users is exploited as a signal of the advancement of information society. There is also a tendency to use single datasets to demonstrate the level of internet access in comparison with other countries. However, the limited indicators may tell little about the intensity and extensity of internet penetration in a given country. Certainly, it is naïve to assume that improved research will necessarily create disincentives for politicians, for example, to stop abusing the data. Nevertheless, a richer understanding of internet diffusion will increase the probability that differences between political rhetoric and reality will be pointed out in the public debate. In addition, those who make important decisions concerning allocation of resources—whether in the public or private sector—for improving access to the scarce resource called the internet will benefit from a more sophisticated and complex understanding of internet diffusion.

Most important, our current dependence on existing indicators is conflating dirt roads with information highways. Moreover, the goal of this paper has been to argue that we invest new energy into developing better indices. Thus, we explored the idea of creating an effective internet diffusion index as a solution. Such an index would not only minimize the risks of conceptual stretching but would also give us a much richer picture of internet use. The index would allow combining singular datasets, the use of which was criticized in this paper, and add additional datasets to the most commonly used ones. This integration would permit a reduction in the anomalies stemming from the use of one or a combination of two datasets. A more robust ranking of countries would emerge because anomalies in a singular data set would be canceled out by alternative datasets. Obviously, the development of such an index would be cumbersome and slow, but it is a prerequisite for improving our analysis. Without it there may not be much progress in our understanding of internet diffusion.

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3):520–46.

Atkinson, A. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2 (3):244–63.

Barzilai-Nahon, Karine. 2006 "Gaps and Bits: Conceptualizing Measurements for Digital Divides." *The Information Society* 22:269–78.

Beilock, Richard, and Daniela V. Dimitrova. 2003. "An Exploratory Model of Inter-Country Internet Diffusion." *Telecommunications Policy* 27 (3–4):237–52.

Berrebi, M., and J. Silver. 1985. "Income Inequality Indices and Deprivation: A Generalization." *Quarterly Journal of Economics* 100 (3):807–10.

Caselli, Francesco, and Wilbur J. Coleman II. 2001. "Cross-country Technology Diffusion: The Case of Computers." *The American Economic Review* 91 (2):328–35.

Castilla, Emilio J. 2003. "Networks of Venture Capital Firms in Silicon Valley." *International Journal of Technology Management* 25 (1/2):113–35.

———. 2005. "Social Networks and Employee Performance in a Call Center." *American Journal of Sociology* 110 (5):1243–83.

CIA. 2008. *The World Factbook*. Central Intelligence Agency 2008 [cited November 11 2008]. Available from <https://www.cia.gov/library/publications/the-world-factbook/>.

Coralles, Javier, and Frank Westhoff. 2006. "Information Technology Adoption and Political Regimes." *International Studies Quarterly* 50 (4):911–33.

Dasgupta, Susmita, Sonik Lall, and David Wheeler. 2001. *Policy Reform, Economic Growth and Digital Divide: An Econometric Analysis.* Washington DC: The World Bank.

Dholakia, Nikilesh, Ruby Roy Dholakia, and Nir Kshetri. 2004. Global Diffusion of Internet. In The Internet Encyclopedia, edited by H. Bidgoli. New York: Wiley.

Eccles, Robert G., and Nitin Nohria, eds. 1992. *Networks and Organizations: Structure, Form, and Action*. Cambridge, MA: Harvard University Press.

Fink, Carsten, Aaditya Matoo, and Randeep Rathindran. 2003. "Liberalising Basic Telecommunications: Evidence from Developing Countries." In *Quantifying the Benefits of Liberalising Trade in Service.* Paris: Organisation for Economic Cooperation and Development.

Gapper, John. 2007. "Stylish devices in search of a network." *Financial Times*, June 25, 11.

George, Alexander, and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences.* Cambridge, MA: MIT Press.

Ghoshal, Sumantra, and Janine Nahapiet. 1998. "Social Capital, Intellectual Capital and the Organizational Advantage " *The Academy of Management Review* 23 (2):242–66.

Ghoshal, Sumantra, and Wenpin Tsai. 1998. "Social Capital and Value Creation: The Role of Intra-firm Networks." *The Academy of Management Journal* 41 (4):464–76.

Heimler, Alberto. 2000. *Competition and Regulation of Public Utilities: Privatization, Competition and Regulation*. Paris Organization for Economic Cooperation and Development.

Honig, Benson, Miri Lerner, and Yoel Raban. 2006. "Social Capital and the Linkages of High-Tech Companies to the Military Defense System: Is There a Signaling Mechanism?" *Small Business Economics* 27 (4–5):419–37.

Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge: Cambridge University Press.

ITU. 2004. *Statistics of main telephone lines and internet users*. International Telecommunications Union 2003 [cited April 20, 2004]. Available from <http://www.itu.int/ITU-D/ict/statistics/>.

———. 2006. *Technical notes. Statistics of main telephone lines and internet users*. International Telecommunications Union 2006 [cited August 11, 2006]. Available from http://www.itu.int/ITU-D/ict/statistics/.

———. 2008. *Definitions of world telecommunication indicators*. International Telecommunications Union 2007 [cited November 11, 2008]. Available from <http://www.itu.int/ITU-D/ict/handbook.html>.

———. 2008. *World Telecommunication/ICT indicators*. International Telecommunications Union 2008 [cited November 11, 2008]. Available from http://www.itu.int/ITU-D/icteye/Indicators/Indicators.aspx.

Jacob, Herbert. 1984. *Using Published Data*. London: Sage Publications.

Jensen, Robert. 2007. "The Digital Provide: Information (Technology), Market Performance and Welfare in the South Indian Fisheries Sector." *The Quarterly Journal of Economics* 122 (3):879–924.

Kay, John. 2008. "Statistics, damned statistics and value-added." *Financial Times* August 12, 11.

Kiiski, Sampsa, and Matti Pohjola. 2003. *Cross-country Diffusion of the Internet.* United Nations University World Institute for the Development Economics Research 2001 [cited November 19, 2003]. Available from <http://www.wider.unu.edu/publications/dps/DP2001-11.pdf>.

———. 2002. "Cross-Country Diffusion of the Internet." *Information Economics and Policy* 14 (2):297–310.

King, Gary, Robert O. Keohane, and Sydney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press

Mackay, Hughie, and Richard Gillespie. 1992. "Extending the Social Shaping of Technology Approach: Ideology and Appropriation." *Social Studies of Science* 22 (4):685–716.

Milanovic, Branko. 2005. *Worlds Apart: Measuring International and Global Inequality*. Princeton, NJ: Princeton University Press.

Milner, Helen. 2006. "The Digital Divide: The Role of Political Institutions in Technology Diffusion." *Comparative Political Studies* 39 (2):176–99.

Minges, Michael. 2008. *Measuring the diffusion of the internet*. International Telecommunications Union 1999 [cited September 12, 2008]. Available from http://www.itu.int/ITU-D/ict/papers/1999/MM-Inet99-Jun99.ppt.

Mokyr, Joel. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. New York Oxford University Press.

Munck, Geraldo L. 2004. "Tools for Qualitative Research." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. H. E. Brady and D. Collier. Lanham, MD, and Oxford: Rowman & Littlefield Publishers.

Norris, Pippa. 2001. *Digital Divide. Civic Engagement, Information Poverty, and the Internet World-wide*. New York: Cambridge University Press.

Press, Larry, Grey Burkhart, Will Foster, Seymour Goodman, Peter Wolcott, and Jon Woodard. 1998. "An Internet Diffusion Framework." *Communications of the ACM* 41:21–26.

Putnam, Robert. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64 (4):1033–53.

Seawright, Jason, and David Collier. 2004. "Glossary." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. H. E. Brady and D. Collier. Lanham, MD, and Oxford: Rowman & Littlefield Publishers.

Slater, Don, and Daniel Miller, eds. 2000. *The Internet: An Ethnographic Approach* Oxford: Berg.

UN. 2007. *UN E-Government Survey 2008: From E-Government to Connected Governance*. New York: United Nations

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

WEF. 2008. *The Network Readiness Index*. Geneva, Switzerland: World Economic Forum.

White, Harrison. 2002. *Markets from Networks: Socioeconomic Models of Production*. Princeton: Princeton University Press.

Wu, Irene. 2008. "Who Regulates Phones, Television and the Internet? What Makes a Communications Regulator Independent and Why It Matters." *Perspectives on Politics* 6 (4):769–83.

Zook, Matthew. 2006. Top Level Domain Name Geographical Adjustments. Personal Communication, Seattle, January.

Turning Dirt Roads into Information Highways: The Conceptual Misformation of Internet Diffusion

Meelis Kitsing and Philip N. Howard