An Analysis of International Internet Diffusion

by

Gonzalo Figuera Díez-Picazo

Ingeniero Industrial
Universidad Politécnica de Madrid, 1993

Ingénieur des Arts et Manufactures
Ecole Centrale Paris, 1993

Submitted to Technology and Policy Program
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Technology and Policy

at the

Massachusetts Institute of Technology

June 1999

© 1999 Gonzalo Figuera Díez-Picazo
All rights reserved

The author hereby grants permission to MIT to reproduce and publicly distribute
paper and electronic copies of this thesis document in whole or in part

Signature of Author ____

Technology and Policy Program
May 14, 1999

Certified by ____

Sharon E. Gillett
Research Associate, MIT Center for Technology, Policy and Industrial Development
Thesis Supervisor

Certified by ____

Dr. William H. Lehr
Research Affiliate, MIT Center for Technology, Policy and Industrial Development
Thesis Reader

Certified by ____

Professor Lee W. McKnight
Associate Professor of International Communications, Tufts University
Thesis Reader

Accepted by ____

Professor Richard de Neufville
Chairman, MIT Technology and Policy Program

AN ANALYSIS OF INTERNATIONAL INTERNET DIFFUSION

by

GONZALO FIGUERA DÍEZ-PICAZO

Submitted to the Technology and Policy Program
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Technology and Policy
June 1999

ABSTRACT

This thesis is concerned with the diffusion of the Internet in an international perspective. Although the booming growth of the Internet is common knowledge, the differences in growth across countries are difficult to measure and explain. The thesis addresses several approaches to measure and analyze these geographical differences and draws some conclusions on the causal factors of the diffusion of the Internet. Policy implications of this uneven international diffusion are also presented.

Before trying to answer the question on the factors that determine Internet diffusion, we consider different ways in which Internet diffusion has been measured and some of the issues involved in those measurements. We find that most of the analysis done to date is based on a very limited set of data that only partially suits the intent of the research. There is a clear need for better, more insightful metrics, and some of the possibilities are listed in the part devoted to metrics.

The differences in diffusion among countries are analyzed by means of statistical and econometric analysis. The number of Internet hosts per capita in each country is regressed on several economic and infrastructure determinants. We find that although there is a close relationship between a country's level of income (GDP per capita), availability of telecommunications and computer infrastructure and level of Internet penetration, there remain important unexplained variations in diffusion, in particular between countries with similar socioeconomic and cultural conditions. The effect of Internet access and transport costs in diffusion is also studied. Finally, an econometric model is proposed and its main results are presented. In particular, we find that little catch-up has happened in the last decade between early adopters and laggards.

The thesis concludes by calling for an increase in the effort devoted by governments and organizations to collect data more suited for better tracking and understanding Internet diffusion, as well as for policies aimed at increasing the global availability of this powerful and promising new technology.

Thesis Supervisor: Sharon E. Gillett
Title: Research Associate, MIT Center for Technology, Policy and Industrial Development

Thesis Reader: Dr. William H. Lehr
Title: Research Affiliate, MIT Center for Technology, Policy and Industrial Development

Thesis Reader: Professor Lee W. McKnight
Title: Associate Professor of International Communications, Tufts University

a Soledad y Javier, mis padres

Acknowledgments

This thesis was made possible with the support of many advisors, colleagues, friends and family. I would first like to thank my advisors, Sharon E. Gillett and Prof. Lee McKnight, for their valuable guidance and support along the way. Lee was responsible for my initial involvement with the Research Program on Communications Policy since my arrival at MIT, while Sharon took over the baton last fall and led me adroitly through the sometimes intricate process of drafting and completing a thesis. Thanks also to Bill Lehr who provided me with valuable directions and advice for the analytical part of this thesis.

I thank all the members and staff of the MIT Internet and Telecommunications Convergence Consortium (formerly the Internet Telephony Consortium), for creating a privileged and exciting environment in which to work and learn about the complex and fast changing world of the Telecommunications Industry. In addition to my advisors, they include Dr. David Clark, Thomas Kiessling, Terry McGarty, Shawn O'Donnel and John Wroclawski. I also thank Merrick L. Berman and Elizabeth Yoon for their efficient and friendly assistance. My fellow students helped in keeping the balance between work, constructive discussion and fun: thanks to Mugo Kibati, David Pearah, David Sirkin, Massimo Sabbatucci, Brett Leida, Annabelle Gawer, Yuki Naruse, Donya Krairit and Isao Sugino. I am also grateful to Antonio Castillo, Antonio Golderos, Antonio Nebreda, Miguel Angel Mariezcurrena, Miguel Angel Blanco and my other former colleagues and current members of the Consortium at Telefónica de España, to Prof. Carlos Vera and Prof. Claudio Feijóo at the Polytechnic University of Madrid, and to Ellinor Sigler at Ecole Centrale Paris, for helping me link my professional and academic interests.

I am most grateful to the "La Caixa" Foundation of Spain for supporting my studies at MIT, and to their staff in Barcelona and in Indiana University for their diligence and kindness.

I wish to make a special mention to the team formed by Prof. Richard de Neufville, Dr. Richard Tabors, Gail Hickey, Jennifer Knupp-Stamp and Linda Manion, to mention only a few, for creating and supporting the Technology and Policy Program at MIT, one of the most interesting, diverse and enriching academic and human groups I have ever encountered after more than ten years in contact with some of most renowned academic institutions in various countries. I do really feel grateful for having been given the chance to be part of this group, and hope to be able to pay back for this opportunity in the future.

My friends at MIT and in Boston (Carlos Gutiérrez, Caroline Proust, David Prosper and Cristina Cuesta, Elena Gil, Estelle Andrin, Jesús Río and Isabel Fernández, Luis García Delgado, Miquel Ortega and Inés Aschl, Pablo Fernández de la Torre and Belén Garrigues, Patricia Mejía, Ricardo Balzola and many others) made working and living here a much enjoyable experience. My MIT host family, Don and Louesa Gillespie have made me very fond of New England through their exquisite hospitality. A special mention goes to Pablo Fernández de la Torre, whose counseling has been invaluable while writing this thesis and through the overall MIT experience. Finally, thank you to Arayna Lindsay Yearwood for her unfailing support and friendship.

Most of all, I thank my parents, Javier and Soledad, and my sister Celia for allowing me the freedom and loving support to pursue my interests and dreams.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

The "Internet" has been the major development of the last decade in the telecommunications world. The rate of adoption of the "Net" by consumers around the planet has been the fastest ever measured for any technological innovation, even compared to the Telephone service and Radio broadcast early in the century, Television in the 50's or Personal Computers in the 80's and 90's. The reasons for such demand are at least twofold: first, the Internet is a way of providing existing "traditional" communication services in new, more flexible and often more efficient ways, the layered architecture of Internet protocols making Internet Services to a great extent architecture independent. But additionally, the Internet is allowing the development of a broad new set of communication services such as email, world wide web and most recent multicasting and video conferencing systems.

The above facts make the Internet not only a huge business opportunity for both existing and new companies, but also an important economic and social variable. In the U.S. alone, for instance, the Department of Commerce estimated that Information Technologies account for about 30% of the real economic growth since 1995, with a growing share directly related to the Internet[1]. Because the Internet is also perceived as becoming an essential education and development tool, governments and non government organizations (NGOs) around the world are pushing to speed up the availability of Internet access to a broad range of consumers. However the diffusion of Internet varies broadly from country to country and even in different market segments within countries, and the reasons for such differences are not always easy to explain. The ability of both predicting the speed of

[1][DOC98, P. 6].

diffusion of the Internet as well as of identifying the key levers that may alter this speed in a particular region would be a valuable resource.

This thesis aims at investigating such cross-country differences and their possible causes by looking at the evolution of penetration rates over time in a set of representative countries, in comparison with a series of possible causal factors such as: existing telecommunication infrastructure, market structure, degree of competition and regulation, cost and pricing patterns and cultural specificities (in particular, language) together with economic and demographic control variables. The methodological approach used is based on statistical and econometric analysis. Once the key levers are identified, the results are used to suggest several ways in which they could be used by different stakeholders to further encourage Internet diffusion.

Previous to that analysis, this thesis addresses the various possible definitions of Internet Access and which metrics should be used to take into account those different cases. Then I make a review and critical assessment of the data publicly available on Internet penetration and discuss the possible bias that they might contain, to justify my choice of the variables for the analysis.

Related Work

Despite the large amount of media attention that the booming development of the Internet has had, there is a relative scarcity of analytical work on international Internet diffusion to date. The most comprehensive reports on the topic are being produced by international organizations such as the International Telecommunications Union (ITU) and the Organization for Economic Co-operation and Development (OECD). The two ITU reports *Challenges to the Network: Telecoms and the Internet*, 1997 [ITU97], and *Challenges to the Network: Internet and Development*, 1999 [ITU99] are amongst the most reliable and comprehensive references on the topic, and in particular concerning the gap between developing and developed countries. Various OECD reports, in particular *Information Infrastructure Convergence and Pricing: The Internet* [OECD96] and

Internet Infrastructure Indicators [OECD98], also provide useful information and analysis on this issue, although restricted to its 29 member countries.

Some government agencies are also producing some insightful work on the environment of Internet use. However those studies are usually restricted to one particular country or to comparisons between a reduced group of countries. Two good examples of such work are the U.S. Department of Commerce report on The Emerging Digital Economy, 1998 [DOC98] and the set of three reports entitled *Benchmarking the UK in the Information Age* commissioned by the U.K. Department of Trade and Industry between 1996 and 1998 [DTI98]. Also worth mentioning is the report *Internet Counts: Measuring the Impacts of the Internet*, published by the U.S. National Academy Press in 1998 [NAS98]. This report sketches an analytical framework and provides a most useful list of indicators related to the environment for Internet use, that we have included in Appendix F of this thesis.

A few papers have been written on the specific issues of Internet metrics and diffusion patterns, with an approach that is often more qualitative than quantitative. [Petrazzini98] gives a very insightful overview on the main variables affecting differences in Internet development among countries. [Elie98] provides an interesting categorization of Internet statistics that we adopted in this thesis, together with some analysis and a call for the creation of a distributed "Observatory of the Uses of the Internet", which is similar to one of the recommendations of this thesis. [Hargittai98] also explores international differences in the spread of the Internet by means of cross-tabulations.

There is more comprehensive literature on econometric analysis of basic telecommunication infrastructure demand. The econometric model presented in Chapter 4 builds in particular on [Henisz98], that analyses cross-national variations in the growth rates of basic telecommunications infrastructure.

Finally, [Press97] provides a good starting point to understand Internet data sources, and [Press98] describes a particular diffusion framework to which we refer in Chapter 3.

Overview of Thesis

Chapter 2 provides the background of Internet diffusion: a brief historical review of the origins and phases of Internet diffusion (§2.1), a discussion on the importance of Internet diffusion for economic growth and social development (§2.2), and a description of the basic determinants that will be analyzed in the following chapters (§2.3).

Chapter 3 focuses on the metrics and data sources available for measuring Internet diffusion. Metrics are categorized as endogenous, exogenous or compound. For each type, available data sources are listed and limitations are discussed (§3.1). Finally, the relationship between some of the most important metrics is analyzed (§3.2).

Chapter 4 presents the analysis and results of the variations of Internet development among countries. The first part of the chapter describes how Internet diffusion can be modeled as a growth process (§4.1). Then the differences in the growth of the Internet between countries are described (§4.2) and analyzed in light of different causal factors (§4.3). Section §4.4 presents the results from a cross sectional statistical model combining time series of various explanatory factors.

Chapter 5 summarizes the findings of the thesis (§5.1), gives policy recommendations both in the area of data collection and in the one of Internet development (§5.2) and suggests areas for future research (§5.3).

Chapter Two

The context of Internet Diffusion

This chapter provides the background and context for understanding Internet diffusion: a brief historical review of the origins and phases of Internet diffusion (§2.1), a discussion on the importance of Internet diffusion for economic growth and social development (§2.2), and a description of the basic determinants that will be analyzed in the following chapters (§2.3).

1. Origins and Phases of Internet Diffusion

What today is called the Internet technology was initially developed during the 1960's by researchers from a department at the U.S. department of Defense named the Advanced Research Projects Agency (ARPA). The purpose of the agency, created in 1957 in the context of the Cold War, was to strengthen long term high risk/high payoff U.S. research and development, and to do so placed great emphasis in the then emerging computer technologies. In particular, the agency realized that "machines needed greater capability to interact with each other to gather relevant information, solve problems, anticipate data requirements, communicate effectively across distances, present information visually, and do all this automatically".[2]

Several projects were undertaken by ARPA in the areas of computer time-sharing and networking between 1957 and 1965. In 1966, a new networking project was initiated for connecting all computers in the research community via dial-up telephone lines. The basis of the technology was packet switching, consisting in breaking up the data to be sent from

[2] Norberg, Arthur L. "Changing Computing: The Computing Community and DARPA." IEEE Annals of the History of Computing, Vol. 18, 2, 1996, p. 42. Quoted from [Laursen97].

one computer to another in small packets containing information about their origin, destination and information allowing them to be reassembled at the destination. The packets would travel over a network of computers acting as peers with each having equal status and data transmitting capabilities. There would be redundant paths from origin to destination and all network elements would communicate with a common communication protocol. The aim of the technology was to define communications protocols to connect heterogeneous computer systems in a flexible, distributed way, in order for the system to keep working even when some nodes would be "down".

In 1969, the first four remote computers were connected to what was then called the ARPANET. All those computer nodes were located in the United States.

During the 1970s the ARPANET kept evolving in size and stability, as a growing number of institutions joined the network and new protocols were tested to perform a number of tasks, such as remote log-in (telnet), file transfers (ftp) or electronic mail (email) on top of the basic communications protocols in charge of establishing and maintaining communication. The basic protocols themselves were improved and scaled to become the Transmission Control Protocol / Internet Protocol (TCP/IP) in 1982. At the same time, experiments were performed to connect various types of packet switching networks such as satellite, radio and cable networks. The first transatlantic connections were established with the U.K. and Norway in 1973. By the late 1970's, about 100 host computers were connected at institutions in the U.S. and a few NATO countries.

The "civilian" use of the ARPANET was highly influenced by the creation of the Usenet in 1979. The Usenet was a hierarchy of discussion groups ("newsgroups") which were distributed between computers running UNIX at academic institutions via modems and phone lines. Soon a link was established to the ARPANET in order to share the mailing lists. Usenet traffic became a major component of the ARPANET, and contributed to the establishment of additional international connections to Europe and Australia. This raised the discussion about the creation of dedicated research data networks, leading to the creation in the early eighties of two new networks between different groups of U.S.

universities: the CSNET, sponsored by the U.S. National Science Foundation, and BITNET, sponsored by IBM corporation. Both networks and the original ARPANET were all connected together via the TCP/IP protocols, giving birth to the "Internet". By 1985, there were about 1000 hosts (end-user computers) connected to the Internet, and almost all users were still academics and researchers at U.S. and Canadian institutions. The monthly volume of traffic in the backbone amounted to about 17 GB (gigabytes) in January 1988.[3]

The creation of the CSNET and BITNET proved the need for networking for researchers and universities and prompted the NSF to design the NSFNET in 1986 as a new trans-continental network based on TCP/IP protocols. The core of the network consisted of 5 super-computing centers interconnected by means of high-speed connections (the "backbone") and allowing smaller institutions to access the network. Between 1986 and 1995, the NFSNET program spent $200 million making the transition from a research network to one based on commercial equipment; in eight years, the network grew from five nodes with 56 kbps links to 21 with multiple 45 Mbps links, and from about 1000 to 50,000 networks connected. By 1995 TCP/IP had supplanted or marginalized most other wide-area protocols, and IP was on its way to becoming "the" bearer service for the Global Information Infrastructure.[4]

In parallel, personal computers (PCs) started to become broadly available in developed countries during the 1980s, growing at incredible rates: in the U.S., there was about 1 PC per 100 people in 1980; In 1985, 1 PC per 10 people (a 58% 1980-85 compound annual growth rate –CAGR), and in 1990 about 1 PC per 5 people (a 38% 1985-90 CAGR). The growth rate has since then slowed down, growing to "only" about 1 PC every 2 people in 1998.[5] However the massive use of the PC as an Internet terminal did not start until the end of the decade: the first e-mail links between the NSFNET and commercial mail carriers were established in 1989 (with MCI mail and Compuserve), and the first commercial dial-up Internet access provider was created in 1990 (The World).

[3] [McKnight97], p. 29. Figures based on NSFNET data.
[4] [Leiner97].

The next major step driving massive Internet adoption was the development of the World Wide Web (WWW), an hypertext based user interface that made navigation in the Internet much more user friendly. The first WWW prototype was developed at the CERN in 1990, and in two years browser software (Mosaic) became available for the most common operating systems (Microsoft Windows and Apple Macintosh), leading to a new exponential growth process with respect to WWW adoption.[6]

By the end of 1998, there were about 147 million Internet Users Worldwide, for about 43 million hosts. More than half of the users are still located in the U.S. and Canada (76.5 + 6.5 million). The users in North America (83 M), eastern Europe (36 M), Japan (9.5 M), Australia (4.4 M) and Taiwan (1.65) alone account for 92% of worldwide users.[7] The volume of traffic has become almost impossible to measure due to the multiplicity of networks that are part of today's backbone, but an order of magnitude would be around 1 million GB/monthly.[8]

Today, most Internet traffic in developed countries is WWW traffic, including a wide range of applications from traditional web browsing to advanced electronic commerce applications. In the near future, new applications such as webcasting, multicasting and Internet telephony are likely to increase their share of the pie. But in many developing countries, email is still the main or the only driver of Internet use. These are important differences that are often difficult to take into account when considering just number of machines or users instead of the usage they make of the net. In the future, new applications and new Internet appliances will appear that will challenge again our metrics and our understanding of Internet drivers.

---

[5] Source: [#CIA].
[6] For figures on WWW growth, see for instance Hobbes' Internet Timeline v.4.1, URL: http://info.isoc.org/guest/zakon/Internet/History/HIT.html.
[7] Souce: [#CIA], survey on *Top 15 Countries in Internet Use at Year-End 1998.*
[8] Packetcom, Inc. estimate, quoted from [Gromov99].

2. The Importance of Internet Diffusion.

2.1. Economic Growth

In the U.S. alone, the IT sector (computing and telecommunications) contributed to 40% of 1998 real growth.[9] A growing share of that growth is directly related to the Internet, that allows connections outside organizations to become much more easy to implement. The Department of Commerce report on the digital economy distinguishes the following four types of economic activity that are boosted by the spread of the Internet:

1) The building of the Internet itself: the expansion of the Internet drives important increases in computer, software, services and telecommunications investments.

2) Electronic commerce among businesses: using the Internet for creating, buying, distributing and selling service products and services allows significant productivity improvements for two reasons: first, any product made available over the Internet does instantaneously get more potential customers (any one having an Internet connection) than over any other distribution channel. Second, because the Internet is a standard, worldwide protocol, the different players in the market place do not need to agree “a priori” on a custom procedure to exchange information or transactions.

3) Digital delivery of goods and services: a growing number of goods, including Software programs, newspapers, music CDs, banking and securities transactions, airline tickets or theater reservations and can be delivered electronically over the Internet, saving in packaging, distribution and sales costs.

4) Retail sale of tangible goods: the Internet is also an efficient tool for ordering and providing information on tangible products: computers, books or flowers, for instance.

2.2. Social and Economic Development

The impact of the Internet in social and economic development is expected to be paramount as it allows easier access to communication, education and social resources.

Besides, the Internet has the potential to lower the barriers between producers and consumers, especially in developing countries: producers can now publicize their products in the Internet in a very cost effective way and get direct orders from final consumers. This can be particularly relevant for less developed countries that traditionally did not have the scale to output their products to consumers in the richer countries. Organizations committed to fair trade such as Peoplelink provide this kind of electronic commerce services.[10]

However, it is often quoted that half of the world population is still waiting for its first phone call as an illustration of the big remaining information inequalities between the have and have not. In such a world, does it make sense to talk about the Internet as a development tool, or should we proceed sequentially and care first about basic communication services (i.e. telephone) in developing countries before mentioning the Internet? We believe it does make sense because of the capability of the Internet to cost-effectively allow the integration of many telecom services. Because of the flexibility with which the Internet allows to provide different services over different networks, it could facilitate a certain telecommunications "catch up" in developing countries if deployed wisely. The first telephone call for which that many people are still waiting could prove to be, in the end, an Internet telephony call.

3. Determinants of Internet Services Diffusion.

This section provides the background for understanding what are the determinants of the Internet in an international perspective. Identifying those different determinants is important because their presence, absence or relative strength in each particular country play an important role in the availability and the pace of diffusion of Internet services.

The determinants can be categorized as follows:

---

[9] [DOC98].

[10] PeopleLink is a non profit organization that provides an e-commerce web site for arts and crafts producers from developing countries around the world. URL: http://www.peoplelink.org.

- **Infrastructure determinants:** They account for the availability of network and appliances that support the Internet, such as: terminal equipment (mainly PCs), access networks, Internet Service Providers and International Connectivity.
- **Regulatory and market structure:** Regulation and market structure have an impact in the costs and supply side of Internet services.
- **Cost determinants:** They reflect the relative cost of the elements listed above. Of particular interest are connectivity costs, both of local access and of leased capacity. Terminal equipment costs are roughly similar around the world.
- **Socioeconomic determinants:** the overall income level, income distribution and economic structure have an impact in demand of Internet services.
- **Cultural determinants:** the level of literacy, computer literacy, languages spoken or even religious beliefs can have an impact in adoption of the Internet.

Of all of the above, this thesis concentrates on infrastructure, costs and socioeconomic background. The reasons for the choice are linked to the availability of reliable quantitative data rather than to the belief that those are the more important factors. Indeed, both regulatory and market environment are thought to play an important role in Internet adoption.

3.1. Telecommunication Infrastructure

No matter its name, the “Internet” is not a network in itself, but rather a set of communication protocols that can run over a variety of networks, both public and private. The ways by which a particular individual at a particular location connects to the global Internet are far from being simple. At least the following elements are involved:

- User terminal/appliance
- User access network (network that connects user to ISP)
- ISP
- ISP connection network (network that connects ISP to backbone)
- Regional backbone provider
- International backbone provider

Internet connectivity relies on the availability of both backbone and access networks in order to allow users to connect. Therefore one can expect that Internet diffusion will be correlated with the development of telecommunication infrastructure in each country. Figure 2.1 provides an overview of the different elements involved in Internet connectivity.

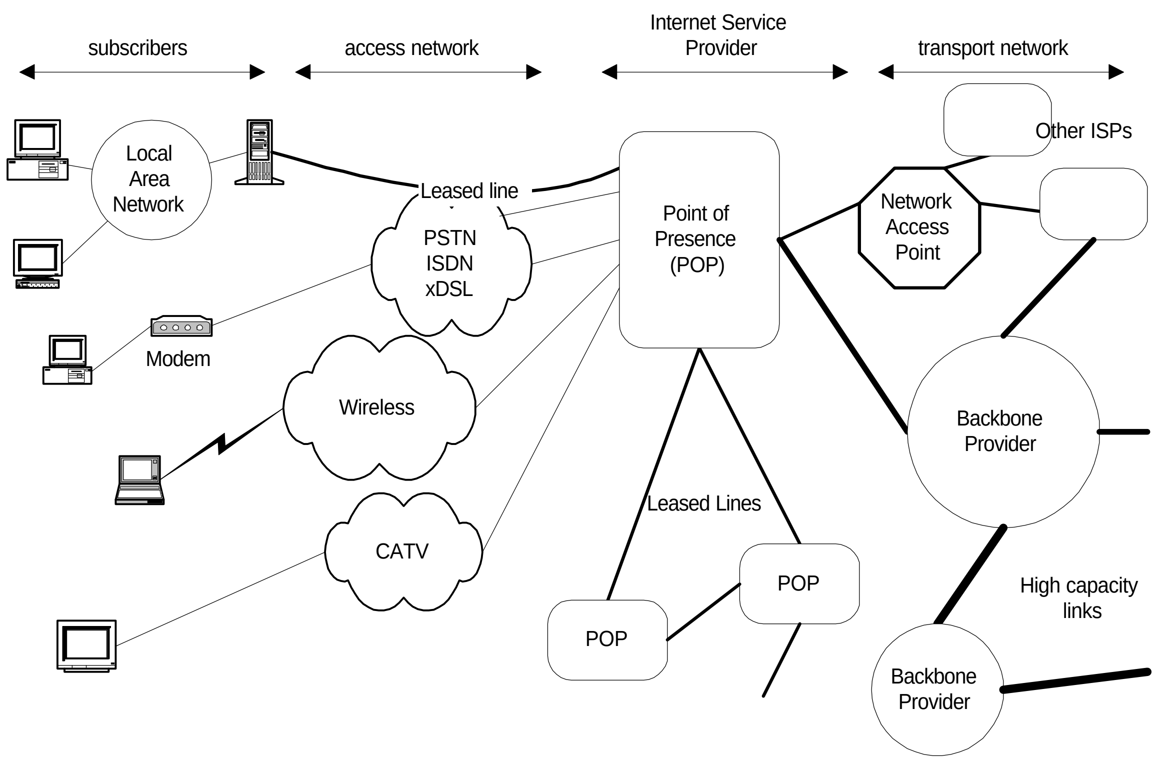


Figure 2.1: Overview of Internet Connectivity

Teledensity

A first approach to test this hypothesis is to look at the relationship between teledensity and Internet diffusion. Teledensity, or number of main telephone lines per 1000 inhabitants, is one of the most used infrastructure indicators.[11] This approach is taken in Section §4.3.2.

However, teledensity is a very rough measure of communication infrastructure in a country, and does not take into account the very different technologies used today in the telecommunication infrastructure. This is very relevant because some infrastructures are more Internet-friendly than others. In particular, networks with a high degree of digitization facilitate the introduction of Internet services. Similarly, broadband-capable networks will allow the introduction of advanced services faster than others. In general, telecommunication infrastructure investments are planned with very long amortization periods, what means that once in place, operators expect to exploit networks for a long period of time before replacing or upgrading them. For this reason, in some cases, the lack of extensive infrastructure can be an advantage for some developing nations in the adoption of Internet services, as they can potentially install more advanced networks faster (assuming that developments are planned with long term objectives in mind, which often contrast with shorter term political objectives). If that was the case, we should be able to observe a certain catch-up, i.e. countries with an initial low level of legacy infrastructure being able to increase penetration rates of new Internet technologies at a faster pace than more advanced countries. This hypothesis is further investigated in §4.1 and §4.4.

Terminal Equipment

In order to access the Internet, a user needs an interface that connects her to the network. In the early years of the Internet, such interface might be a terminal console of a large corporate or university mainframe that could potentially serve many users. Today, most Internet users use a Personal Computer (PC) to connect to the Internet, either at home or at work. Therefore it is of interest to study the relationship between PC penetration and Internet penetration. This is done in Section §4.3.2. In the near future, it is more likely that there would be a large number of appliances other than PCs that will allow Internet access, including Internet phones, handheld Personal Digital Assistants (PDAs), Email-capable pagers, Set top boxes allowing access through the TV, and Network Computers. All those developments are likely to further facilitate Internet access and will make it less dependent on the availability of PCs.

[11] The ITU defines main telephone lines as telephone lines connecting customer's equipment (e.g., telephone set, facsimile machine, modem) to the Public Switched Telephone Network (PSTN) and which have a dedicated port on a telephone exchange.

Leased Lines

Leased lines are an essential part of Internet connectivity, and probably one of its most obscure parts. A “leased line” is the informal name for a connection to the digital switching system used by telephone companies to carry data. Therefore a leased line is not just “dry copper”, i.e. plain copper wires from the Central Office to the user’s location: it is an interface to the carrier’s bit-synchronous data system. Leased lines are used by companies willing to have a high speed permanent connection to the Internet, as well as by Internet Service Providers in order to interconnect their different point-of-presence (POPs) between themselves and with Network Access Points (NAPs) or Backbone Providers in order to exchange traffic with the rest of the Internet.

In many countries, only local telephone companies offer leased lines. As competition develops, there might be alternative carriers or alternative networks that can be used. This situation of near monopoly has caused leased line prices to be highly expensive and arbitrarily priced all over the word (see Section §4.2.3).

3.2. Telecommunication Market Structure

The last decade has been characterized by very important changes in most countries’ telecommunications industry, that has been transformed from state-owned national monopolies to private companies facing global competition. This transformations have been achieved by means of a gradual series of privatization and deregulation processes taking place in the years between 1984 (with the breakup of AT&T in the U.S.) until today, with the public network operators from Brazil and Venezuela being privatized in the last year. Most countries in the E.U. have been through similar processes between 1992 and 1998. Many other countries have similar plans for the coming years. The main argument in favor of privatization is that private companies open to competition achieve better productivity than state-owned monopolies, allowing lower costs, better prices of services to consumers and higher incentives for innovation, and in particular for Internet diffusion.

Internet services have in general not been subject to the same regulatory restrictions as basic services, and therefore should be less affected by the changes in market structure. An exception to this are countries such as India that impose high license fees to ISPs, a significant barrier to entry in local markets. However, as the Internet relies on basic infrastructure for its development, the impact of market structure ends up being significant. In the end, competition and regulation affect the behavior of prices of telecommunications inputs which in turn have an impact in supply and demand of Internet services.

Modeling the impact of regulation or market structure faces the issue of which are the correct units of measure for "degree of regulation" or "degree of competition". In this thesis, we do not explicitly analyze the impact of regulation or market structure.[12] We do implicitly when we consider connectivity and access prices, that are highly influenced by them.

3.3. Costs

It is intuitive that there should be a relationship between costs and Internet penetration. However, measuring and studying costs in the Internet is a complex matter in which more theoretical and applied research is needed. The reason for that complexity comes from the polymorphic nature of Internet services: it is today almost impossible to determine what is a "typical" Internet service provider, in order to model its costs.[13] Similarly, from the end user point of view, there is a multiplicity of ways in which Internet connectivity can be achieved: via a dial-up subscription, a dedicated access line shared by a number of users, and since recently via alternative access networks (CATV, Wireless, Satellite). The cost of Internet service itself might not be easy to track when it is included as part of a bundle (e.g., cable TV + Internet, or long distance + Internet, or even "free Internet access" as part of a bank account benefits, for instance), or when the end-user is not the one paying for the service, as if she is using a dial-up connection to her company. In view of this

[12] For a very sound analysis considering explicitly this type of variables, for the case of the impact of political constraints in investment in basic telecommunications infrastructure, see [Henisz98].
[13] [Leida98] p. 32.

complexity, we have chosen to limit our cost analysis to that of two elements that we consider key to Internet costs: the costs of residential dial-up Internet access and leased capacity.

Internet Access costs

Given a certain level of infrastructure availability and demand for Internet services, the level of diffusion will be given by the cost of access to the network to the potential users. Here again, the difficulty arises from the fact that there are many different ways of connecting in terms of quality, bandwidth and pricing schemes: a final user might pay for her own residential connection at home, or may connect from work or school where a high capacity link is paid for by her organization. Some of those companies or schools might provide a dial-in service for there members, thus serving as a free IAP. Sometimes, a third party such as a bank or a supermarket chain might give away Internet access as part of a marketing promotion.[14] Commercial IAPs may charge their users on flat rate or on a per usage (time or volume transferred) basis. The same applies to the telephone charges, that can be either flat or metered as local or long distance, depending on the regions' LECs tariffs and on the location of the IAP. As for high capacity connections, the prices charged by IAPs or ISPs may vary greatly, even from one customer to another for similar services. This complexity makes it difficult to come to quantitative conclusions.

A good possible approximation to evaluate the impact of access costs on the adoption of the Internet is to restrict the study to residential dial-in Internet service, for which it is possible to define a standard "Internet basket" reflecting the costs of a typical user that spends a certain number of hours on line every month.. The problem is that the number of residential dial-in users is not well known in general, neither is the amount of time they spend on-line or the type of services they use. Additionally, different people collect data according to incompatible methodologies.

[14] For example, Spain's mayor retail bank, Banco Santander Central Hispano, offers a free ISP service to its customers.

A first attempt to collect comprehensive data on Internet access pricing was conducted by OECD in 1996.[15] More recently, the ITU has published similar data for a series of major economies (Appendix D). These data are used in Section §4.2.3 to study the impact of access costs on Internet adoption. For the time series analysis of Section §4.4 only local telecommunications access cost are used, because of the unavailability of Internet access price data series.

Leased Lines

As seen in the previous section, leased lines are an essential part of Internet connectivity, as they are used in 1) high speed dedicated connections for large end-users to their IAP or ISPs, and 2) interconnecting links between ISPs distributed nodes and between the ISPs to the backbone. Therefore their cost should play an important role both on the use of high speed connections by end users and on the costs of the ISP itself.[16]

Determining the costs of leased lines may prove to be challenging. In most countries, leased lines are only provided by the incumbent operators at prices calculated by means of complex tariffs that vary with distance, capacity and use of the line, usually with little progressivity or clear relation to costs. Additionally, high volume users will get discounts that not always made public. Some organizations have tracked leased lines prices at different points in time, such as the OECD in its 1996 Internet pricing report,[17] the International Telecommunications Users Group in 1997[18] or the Eurodata Foundation in 1998.[19] Appendix D shows monthly costs of national 2 Mbps leased lines in different countries in 1997. This figure can be used as a proxy for a country's leased capacity costs to assess the relationship between the cost of leased capacity, the cost of Internet access and Internet penetration (see Section §4.2.3).

---

[15] [OECD96].
[16] [Leida98] p. 95 calculates that transport costs account for about 24% of the overall costs of an average ISP in the US.
[17] [OECD96], p. 32, Table 9.
[18] Survey available at http://www.intug.net [#INTUG].
[19] Survey published in Public Network Europe, March to September 1998 [#PNE].

International Connectivity

With most of the Internet backbone being controlled by U.S. corporations, non-U.S. ISPs are usually forced to pay for backbone connectivity by means of settlements. Settlements are payments made between networks or Internet service providers in return for interconnection and interoperability. The general tradition in the U.S. networks has been a "pass the parcel" approach so that payments migrate towards the center from the periphery but without any systematic settlement policy between large networks in the middle.[20] This means that the largest carriers –such as MCI-Worldcom, Sprint and GTE- typically consider each other peers and exchange traffic without charging fees, while smaller wholesale carriers or ISPs are required to pay for interconnection. This applies in particular to non-U.S. ISPs. In addition to interconnection settlements, non-U.S. ISPs are often required to pay for the whole international circuit from their country to the U.S. The original reason for this asymmetrical system was that as most Internet content used to be in the U.S., non-U.S. providers benefited more from interconnection than U.S. providers. This argument is being increasingly challenged by many non-U.S. ISPs who claim that the content imbalance does not longer apply.[21]

3.4. Social and Cultural Factors

In the end, Internet users are people, and their demand for Internet services will closely depend on how those services fit their cultural and social skills, habits and needs. It is beyond the scope of this thesis to investigate each of those factors quantitatively, but we present in this section a qualitative overview of the most relevant ones. A suggestion for further work is to incorporate data on those factors in the study of the cross-country differences in diffusion.[22]

[20] [Cawley96], p. 4.
[21] [ISPA98].
[22] For a more detailed discussion of the impact of social and cultural factors on Internet Diffusion, see [Petrazzini98].

Demographics

Demographics have been argued to play a role in Internet diffusion in various ways. First, geographical or political isolation seems to increase the use of the Internet as way to stay connected to the main work population centers: regions and countries with low population densities or geographically isolated, such as Canada, Iceland, Scandinavia, Australia and New Zealand seem to have a higher than average penetration level. Second, Internet access is often made first available in the population centers, then spread to the rest of the country. And third, age plays a role as younger people and in particular children are more likely to adopt more easily a new technology such as the Internet.

Education

Education also plays multiples roles in Internet adoption: first, a minimum level of "traditional" literacy is a prerequisite to use the Internet. Second, a minimum of computer and Internet literacy is also needed, and can be taught as part of the curriculum in middle or high school. But third, the Internet is most used by and useful for people with higher education, because it is becoming the locus of state-of-the-art knowledge in a growing number of fields. Therefore, there should be a link between different levels of education and Internet demand, that deserves further study.

Language

Language also plays an important role in Internet diffusion. For historical reasons, English has been the language of computing and internetworking since its origins. This means that anglophone populations or populations with a good command of English as a second language are in a somewhat more favorable position to adopt Information technologies. Also because the larger part of Internet early adopters have been English speakers, most of the content available is in English, and even content published in other languages is often coupled with an English version. In a survey published in 1997, the Internet Society estimated that around 82% of web content was published in English.[23] This compares to estimates of on-line population per language in which native English speakers represent

[23] [#ISOC].

only about 56% of Internet users.[24] There is clearly a deficit in local content in many countries that can act as a disincentive to adopt the Internet, especially in those where the knowledge of English is low. However there are examples of non-native English speaking countries, such as Iceland or the Scandinavian countries, that have higher penetration than the U.S., the U.K. or Australia. Therefore the language factor by itself can't explain everything.

[24] [#EUROMK99].

Chapter Three

Internet Metrics and Data Sources

In order to measure the relative “diffusion” of the Internet, we need to answer “how much Internet” there is in every particular country. This leads us to the issue of how to quantify its use. One of the most commonly asked questions about the Internet, and also one of the most difficult to answer, is “How big is it?” Despite the enormous amount of publications and media attention on the subject, there is not as of today a satisfactory answer. Several factors explain such difficulty:

- First, the ambiguous nature of “the Internet” leads to various possible definitions, each of them relevant for a particular use. For instance, a capacity planner would care about the amount of traffic carried over the physical networks, a manufacturer of routing equipment might be interested in the number of networks, domains or IP addresses in the Internet, while an electronic merchant would most likely want to know the demographics and purchasing habits of the human users of the Internet.

- Second, the nature of the Internet makes it extremely flexible in terms of networks, interfaces and terminals used, making any basis for statistics very short-lived. Should we be counting IP addresses, PCs, cellular phones, bytes transferred over the Internet backbone? Today, most IP addresses and Internet hosts correspond to individual computers used by one or a few people, but that might not be the case in the near future when for instance several home appliances are hooked to the network in order to allow remote control from the net. More in the short term, some of the metrics involving the counting of physical machines are likely to become less accurate as intranets, firewalls, and new appliances other than PCs develop.

The aim of this chapter is to provide some clarification over the complexity of Internet metrics and to present what data and data collection methodologies are available today. We will review some of the most interesting statistics available, comment on their strengths and limitations and suggest new approaches to the structuring of knowledge on the Internet.

1. Choice of Metrics for Study

Because of today's lack of reliability of Internet users' statistics, we have chosen to base our international comparisons of next chapter on the number of Internet hosts per capita rather than in users figures. Of all available metrics, hosts per capita presents several advantages: it allows automatic collection, it is mainly country-based, and it is related to some of the longer lasting elements of the Internet (IP addresses and the DNS). What follows is a categorization and discussion of available metrics to justify our choice (Section §3.2). Additionally, in Section §3.3 we show that the close relationship between Internet hosts and users allows for taking the first as a proxy for the second.

2. A Review of Metrics and Data Sources for Internet Diffusion

In an ideal world, we would know the evolution in the number of users along time, and the characteristics of those users (who they are, where they are). We would know the evolution of the use they make of the different available Internet services, how much time they spend on-line and what content do they access the most. We would know details about what service providers and terminals they use to get connected, whether they do it at home, at school or at work, if they pay for the service themselves, and if so how much. We would additionally want to understand how does the Internet change their way of working, of communicating, of living. And we would like to know how all those things change for people in different countries, regions and social, economic or educational backgrounds.

Understanding all this would allow us to better understand the technical, economic and social determinants and impacts of Internet development, and therefore to take better and earlier advantage of the forecasted benefits of this powerful tool.

Unfortunately, very little of this ideal information is available today. In addition, the available information presents many methodological problems that prevent its use for describing quantitative trends. It is difficult to get coherent comparable statistics along time and across countries.

2.1. Categorization of Internet Metrics

Among the many possible categorizations of Internet metrics, we find particularly appealing the one suggested by Michel Elie.[25] He suggests two main categories of metrics: endogenous and exogenous. Endogenous metrics are obtained in an automatic or semi-automatic way from the Internet itself and provide quantitative information such as number of IP addresses, hosts and traffic. Data concerning availability of physical components related with the Internet, such as computers and modems, also fall into this category. Exogenous metrics are defined and obtained from outside the network by surveying human users about their on-line and off-line behavior. To those two we add a third category, compound metrics, that are custom frameworks defined over various dimensions of the state of the Internet.

Endogenous metrics:

- Internet hosts
- Computers
- Modems
- Subscribers
- Content accessed
- Traffic flows

Exogenous metrics:

- Users
- Usage
- e-commerce revenues

Compound metrics

- Mosaic group

[25] [Elie98].

2.2. Endogenous Metrics

Despite the image of the Internet as a huge, somewhat anarchic and untraceable network, there are a number of technical means in the Internet protocol itself and closely related technologies that potentially allow for distinguishing and tracing different users and machines.[26] Those means can be used to produce very detailed metrics of the online behavior of Internet users.[27] The following are the most relevant of such metrics, with their strengths and limitations.

i) Internet Hosts

A first step in knowing the size and growth of the Internet is to use its definition as a "network of interconnected computers", and ask for the number of computers (or "hosts") connected to the network. This is in fact one of the most common indicators used to measure Internet development. The two most well known and broadly used Internet host surveys are performed on a regular basis by Network Wizards for the whole world and by RIPE (*Reseaux IP Européens*) for European Countries. Network Wizards has been collecting data every 6 months since 1981 (Table 3.1).

[26] Some of the tools allowing to identify users and machines in the Internet are:

- *IP address:* a unique number that identifies any single machine connected to the network, and allows information to be sent, routed and received over the net. The number is assigned in a hierarchical way from national registration authorities, to network managers and ISPs, and down to individual users. IP addresses are the basis for hostcount, the main metric used in that thesis.
- *Unique Ethernet address:* a unique number included in every network access device, such as a network card in a PC or peripheral, that allows information to be distributed over local area networks.
- *Unique code in computer's CPU:* Intel corporation recently cancelled its plans to include a unique code is all of its microprocessors that would allow tracking of the use of individual computers.
- *Unique code in operating system or program:* in a similar way, a unique watermark can be inserted in every copy of an operating system, software program or Internet browser, and included in documents created and transmitted over the Internet. Recently Microsoft corporation admitted having included one of such watermarks in is popular Office software –in fact, it was used to find the perpetrator of the "Melissa" virus.
- *Cookies:* cookies are unique identifiers sent by websites to keep track of individual users. They are kept in the user's computer. Informed users can disable cookies by changing the settings in their web browsers.

[27] Those means provide useful tools both to researchers interested in the evolution of the Internet and to market specialists willing to track customer behavior. Often though, those means raise important privacy concerns because they allow the construction of very large and detailed databases on individual users' on-line behavior, potentially without their explicit authorization. Regulation is being drafted to take into

Date	Hosts	Source
Aug-81	213	host table
May-82	235	
Aug-83	562	
Oct-84	1,024	
Oct-85	1,961	
Feb-86	2,308	
Nov-86	5,089	
Dec-87	28,174	old domain survey
Jul-88	33,000	
Oct-88	56,000	
Jan-89	80,000	
Jul-89	130,000	
Oct-89	159,000	
Oct-90	313,000	
Jan-91	376,000	
Jul-91	535,000	
Oct-91	617,000	
Jan-92	727,000	
Apr-92	890,000	
Jul-92	992,000	old domain survey
Oct-92	1,136,000	
Jan-93	1,313,000	
Apr-93	1,486,000	
Jul-93	1,776,000	
Oct-93	2,056,000	
Jan-94	2,217,000	
Jul-94	3,212,000	
Oct-94	3,864,000	adjusted counts
Jan-95	4,852,000	5,846,000
Jul-95	6,642,000	8,200,000
Jan-96	9,472,000	14,352,000
Jul-96	12,881,000	16,729,000
Jan-97	16,146,000	21,819,000
Jul-97	19,540,000	26,053,000
Jan-98	29,670,000	new domain survey
Jul-98	36,739,000	
Jan-99	43,230,000	

Table 3.1: Total Number of Internet Hosts, 1981-1999
(Source: Network Wizards [#NW])

These so-called "hostcount" surveys present a number of advantages that have made them become the most popular source for any Internet statistics. The surveys are based on two of the most long-lasting elements of the Internet, the IP addressing and the Domain Name Service –DNS–, which allows to study trends along time even though the use and available services in the net might have changed dramatically over very few years. Also, the fact that most of the DNS is geographically organized by countries is an easy way to compare Internet development across borders. For those reasons, hostcount are the most consistent, long-term time series data available on the Internet today, and despite its limitations it is the mostly used source for Internet diffusion studies.

Host count surveys are performed in an automated way by means of computer programs that query the domain hierarchy of the net using tools available from the IP and DNS

account those concerns and specify what information can and cannot be collected, by whom, and with or without the user's acknowledgment.

protocols to trace all Internet hosts connected to network. An Internet host is defined as a domain name with an associated IP address record, such as rpcp.mit.edu.[28]

Limitations

The hostcount used to be considered a good "lower boundary" for assessing the number of Internet users, as it was assumed that every assigned address would correspond to a computer, and there would be at least one human user per computer. But there are several reasons why this assumption is no longer true, and will probably become less accurate as time goes by:

1) Not all counted "Internet hosts" correspond to a single computers in the network: a computer might have more than one host name and IP address, and similarly a domain name might be just a "virtual" pointer to another address. Although the survey methodology tries to take into account such possible duplications, it is not always easy to identify them.

2) Additionally, not all machines with an assigned IP address correspond to different human users. Particularly in the near future one can image a multitude of devices such as pagers, cellular phones or even house appliances been assigned IP addresses allowing for communication and remote control. On the opposite, other machines, in particular servers, do not correspond to any particular human end user.

3) More importantly, there are a growing number of computers connected to the Internet that are not accessible to the automated survey. This is particularly the case of hosts in corporate Intranets that are located behind firewalls, computers that block access to the internal network.

4) Different machines and users might share the same IP number at different times, and therefore be counted as a single "host". This is the case, for example, of dial-up users served by ISPs: in many cases those users are allocated in each session different dynamic IP addresses chosen from a pool of addresses available to the ISP for that

[28] For details on hostcount methodology, refer to Network Wizard's web site [#NW].

purpose (e.g. using the Dynamic Host Configuration Protocol (DHCP), aimed at saving IP addresses).

5) The structure of the DNS does not guarantee that hosts under a particular domain are really located in a geographic area. In addition, hosts registered under generic Top Level Domains –gTLDs– (.com, .edu, .gov, .mil, .org) can be located anywhere in the world, although the majority of them are located in the United States (and many studies in fact assume that all gTLD domains are in the United States, which causes distortion in the data).

Limitation	Effect on hostcount results	Relative impact
Computers with more than one IP address	Causes hostcount to overestimate the number of connected computers	Low today, but growing
New non-human interface devices with assigned IP address	Increases the ratio of hosts/users	Negligible today, Important in the near future
Computers behind firewalls, Intranets	Causes hostcount to underestimate the number of connected computers	Important today (20%) (although new methodology accounts for this effect)
Dynamic IP allocation (IP address sharing)	Causes hostcount to underestimate the number of connected computers	Important (many dial-in users are assigned dynamic addresses)
DNS structure (TLDs and gTLDs)	Alters the results of hostcount for particular countries in either direction	Up to 80% for some countries

Table 3.2: Limitations and Bias of the Hostcount Methodology

ii) Computers

A second approach to the issue of quantifying the size of the Internet is to look at the number of computers in use and determine what fraction of them are connected to the network. This approach has the advantage of dealing with something more material, a computer, instead of the more ambiguous concept of an Internet host. Computers are also interesting because for most users they constitute the largest cost to connect to the

Internet (although the computer often has more uses than simply connecting to the Internet).

Limitations

A disadvantage of this method is that there is no automated way to conduct the survey, contrary to the case of Internet hosts. But the number of computers and of modems sold is a relatively well traced figure in most countries. It is then possible (but not obvious) to estimate the number of computers in use, based on the replacement rate, and to estimate the ratio of computers that are connected to the network.

Looking into the future, the relationship between computers and Internet user interfaces may become weaker as new appliances develop in the midst of TV, PDAs, cellular phones, pagers, all of them allowing Internet access. In Section §4.3.2 we take a closer look at the relationship between the two.

Data sources:

Computer Industry Almanac Inc. has been collecting data on computer and Internet statistics from various sources since 1980 and produces yearly estimates on computers in use and of Internet users per country, been the most cited source for such data (Table 3.3). The OECD and ITU also collect data on the number of modems and computers sold in each country.

Country	Computers-in-Use in Millions	% Share of Total
1 U.S.	164.1	28.32
2 Japan	49.9	8.62
3 Germany	30.6	5.28
4 United Kingdom	26.0	4.49
5 France	21.8	3.77
6 Italy	17.5	3.02
7 Canada	16.0	2.76
8 China	15.9	2.75
9 Australia	10.6	1.82
9 South Korea	10.6	1.82
11 Russia	9.2	1.59
12 Brazil	8.5	1.47
13 Spain	8.1	1.39
14 Netherlands	7.2	1.25
15 India	6.3	1.08
15 Mexico	6.3	1.08
Total Top 16 Countries	408.6	70.52
Worldwide Total	579.0	100.00

Table 3.3: Forecast of Top 16 Countries in Computers-in-Use in Year 2000 (Source: Computer Industry Almanac [#CIA])

iii) Modems

An alternative metric to counting computers is tracing the number of modems. Modems are in a sense more directly related to networking, since that is their specific function, while computers might be used in a standalone way. The ITU data does cover the number of modems in use since 1985. We consider however this metric might have been interesting in the past, but today has become less appropriate than the number of computers for two reasons: first, modems have become a very cheap commodity, are often included with new PCs and in many cases faster modems are bought to replace older ones. Additionally, a growing number of computers are connected to the net via different devices, such as corporate networks in businesses and alternative technologies at homes: ADSL, cable modems, etc.

iv) Subscribers

Internet subscribers are the people or businesses that are subscribed to an ISPs. In general, it is not easy to collect reliable data on their number: these data are proprietary to ISPs that in most places are not required to file them publicly, unlike the highly regulated environment of PSTN where carriers have to submit their penetration figures to the regulating authorities. In fact, the scarcity of homogenous, reliable data is one of the drawbacks of telecommunication de-regulation. The alternative is to take the data from voluntary disclosure from the ISPs themselves, or from market studies estimates. An exception to this scarcity of data happened for a while in countries with a near-monopoly ISP (usually linked to the incumbent public network operator), forced to disclose its user figures.

Limitations:

Even if we had good indications on the number of users and their access capacity, uncertainties would still remain regarding the usage of the network and the number of human users per subscription. We will not use subscriber figures in this study, mainly because of their limited availability. But we still think that it would be interesting to establish a collection mechanism for that metric that would be compatible with companies' non-disclosure rights.

v) Content Accessed

The statistics of content accessed are among the most developed in the Internet world and in particular for World Wide Web services. Most WWW servers gather statistics showing how many requests are made at what times and by whom, how many bytes are transferred, errors encountered, etc. These numbers are collected in "log files" and help system administrators and network planners to size and tune their systems.

However, the rising attention on usage measurement does not come from the technical side, but increasingly from the marketing side. As advertising revenue is becoming the dominant business model for many on-line services, knowing what content is "high use" or "low use" is an important information both for marketers and service developers. In many cases, the “page impression” or “hit” is becoming the unit for payments between advertisers and owners of web sites.

The so-called Internet or Web “rating services” collect and compile many of those log file data at many “neutral” sources such as proxy servers, exchange points and caching sites, and use this sample to produce cumulative statistics of the usage of web sites. Table 3.4 shows some of the main rating services available and the methodology they each follow to collect the data.

Service	Start Date	Ranking criteria	Sample Size	Update Frequency	Data Source
100hot	12/95	Home directory page views	100000+	Daily, Weekly, Monthly	Proxy server logs from strategic locations on the Internet backbone
Relevant Knowledge	10/97	Extrapolation of an estimate of "unique visitors"	6000-10000	Daily	Software installed on user's computer sends browser clickstream to data center.
Media Metrix	3/97	Page views, time spent	10000	Monthly	Software on user's computers that requires user's to mail in a diskette monthly; interviews
Net Ratings	7/97	Average usage and demographics	2000	Weekly, Monthly, Quarterly	Software on user's computers; interviews

Table 3.4: Comparison of Different Rating Services (Source: 100hot.com)

Limitations:

Rating services are potentially one of the richest endogenous sources of data on Internet users and usage. Conveniently analyzed, their data can provide an accurate idea of the quantitative and qualitative usage of the net. The problem is that it is still very complicated to translate raw log file data into user/usage data, because of the multiple caveats that exists: should different frames of a single page be counted only once, which of existing usage metrics are more relevant (“hits” vs. “pageviews” vs. sessions), how do we account for the increasing share of multimedia traffic (images, video, pushed content), how can we account for content accessed from caches at different points of the network or behind firewalls, how can we prevent webmasters to artificially try to increase their sites’ rating?

We believe that still a lot of research is needed in that field before content statistics can replace hostcount for the purpose of doing a quantitative assessment of the global Internet. They are however useful for estimating the sources and sinks of information on the web and the evolution of the most accessed content.[29]

vi) Traffic

Another relevant figure is the traffic carried on the Internet, or more precisely on the Internet "backbone". Until the early 90's, this was a somewhat easy data to collect as there was only one backbone, the NSFNET (created by the NSF in the mid 1980s). NSFNET statistics were collected made available via Merit Network, Inc. By the early 90's, alternative backbones began appearing, making the statistics less accurate. It is estimated that until September 1994 the NSFNET accounted for at least 75% of U.S. backbone traffic, and that after that its share fell rapidly until the NSFNET was shutdown in April 1995.[30] Since then, the number of private backbone networks has multiplied, making traffic figures much harder to obtain. Traffic metrics are however perceived as essential by many industry players, researchers and regulators, as they provide the basis for the agreements, settlements and contracts between the different players, and therefore facilitate the development of the Internet. A possible alternative is to measure the traffic going through the Networks Access Points (NAPs), points at which different backbone providers exchange their traffic. The Cooperative Association for Internet Data Analysis (CAIDA) is a government-supported organization created in 1997 to define and assure the utility of such metrics.[31]

[29] For a very comprehensive study of Internet content flows, based mainly of data from 100hot, see [OECD98].
[30] [McKnight97], p. 29.
[31] Refer to CAIDA's web site at: http://www.caida.org.

2.3. Exogenous Metrics

i) Users

The first, most straightforward statistic of the growth of the Internet, at least for the purpose of this thesis, is the number of people that are using the Internet. Knowing the number of *human* Internet users by country seems one of the most interesting metrics (together with their online behavior). Unfortunately, it is one of the most difficult to get, as human communication over the Internet is always computer-mediated and users usually do not identify themselves, neither is there anything like an Internet user registry. However a number of research and market study firms are collecting data on Internet users and usage. In doing so, several approaches are followed. Some follow classical statistical surveys for the overall population of a country. Others use on-line surveys completed by Internet users themselves. Others make derivations based on reported IAP's subscribers counts, or multiplying the number of hosts by an estimated multiplier. The problem with these estimates is that most of them are not performed in a very scientific way and their methodologies and assumptions vary widely, which leads to a very large spread in the estimates of different sources. Another cause of indetermination is the lack of standard definition of what an Internet user is: is it somebody who has "access" to the Internet? Who uses it on a regular basis? If so, how often/ how intensely? Who has an e-mail address? Should children be counted as users?

ii) Usage

Usage statistics try to address the online behavior of Internet users in terms of time spent on line, types of applications used, content accessed, etc. The methodology followed is to trace the behavior of a representative sample of users. One of the most important motivations for those studies is to evaluate how the usage patterns change in relation with changes in technical or economic changes in the service: broadband vs. narowband access, flat rate vs. usage-sensitive pricing, etc.

Usage statistics are starting to become prevalent especially as marketing organizations look at behavior patterns of users that may become potential consumers of products and services on an increasingly commercialized Internet. They will also become more interesting to researchers in order to distinguish the evolution of the many different on-line behavior of users. As of today, the number of available services is still limited to a few (mainly email, web browsing, and increasingly e-commerce, Internet broadcast, Internet telephony), but it is already a misleading simplification to consider all Internet users have a similar on-line behavior. Figure 3.1 illustrates the danger of such simplification by comparing Internet usage in two different countries, and shows how the share of different applications has kept changing over time. However this thesis does not consider usage variations among users, again because of the lack of consistent data for different countries.

Figure 3.1: Comparison of Internet Usage in the US and Bangladesh (Source: [ITU99])

Date	ftp	telnet	news	irc	gopher	email	web
Jul 93	42.9%	5.6%	9.3%	1.1%	1.6%	6.4%	0.5%
Dec 93	40.9%	5.3%	9.7%	1.3%	3.0%	6.0%	2.2%
Jul 94	35.2%	4.8%	10.9%	1.3%	3.7%	6.4%	6.1%
Dec 94	31.7%	3.9%	10.9%	1.4%	3.6%	5.6%	16.0%
Mar 95	24.2%	2.9%	8.3%	1.3%	2.5%	4.9%	23.9%

Table 3.5: Share of Traffic Generated by Different Internet Applications (Source: NSFNET)

iii) E-commerce revenues

We finally mention that an additional way to look at the Internet is to use a standard economic approach, in terms of generated cash flows, costs and revenues. This approach has not been very practical in the past, since the Internet has developed originally out of the commercial sector, and purposely mostly unregulated and untaxed. But as Internet service provision and electronic commerce become closer to mainstream economic activities, it should become easier to use economic magnitudes to trace the evolution of the network. Already today, the amount of revenue from e-commerce is being analyzed as another measure relevant to the development of the Internet. Still, some difficulties exist when translating e-commerce figures geographically: the location of the server/seller is not easy to determine (that's why some proposed regulation on taxation suggests to take into account the buyer's location instead of the seller's for tax purposes).

2.4. Compound Metrics

A completely different approach to understating the evolution of the Internet and in particular its impact in economic, social and cultural development is to *define* significant metrics based on a variety of quantitative and qualitative observations. We refer to such metrics as "compound" metrics. The main difference with the metrics defined above is that instead of looking directly at observable magnitudes, compound metrics artificially define measurement criteria and scales relevant to the Internet diffusion phenomenon. Although we did not use those metrics in our study, we though it worth mentioning because we believe that the need for this kind of approach will increase as the Internet becomes a spanning layer for an increasingly diverse and complex set of applications, making analysis based on observable metrics increasingly difficult to analyze.

i) Mosaic Group's Framework

One of the most widely publicized of such studies is the MOSAIC group's framework on global Internet diffusion. The group focuses in the state of the Internet in each country following six dimensions: pervasiveness, geographic dispersion, sectoral absorption, connectivity infrastructure, organizational infrastructure and sophistication of use. For each dimension, five ordinal values are defined ranging from 0 (non existent) to 4 (highly developed). The group claims the definitions for the different values are different enough to make it easy for informed Internet observers to agree on the values for particular countries and moments in time. The framework allows them to easily compare the status of the Internet across countries or the evolution of the status over time.

Level 0	Non-existent The Internet does not exist in a viable form in this country. No computers with international IP connections are located within the country. There may be some Internet users in the country; however, they obtain a connection via an international telephone call to a foreign ISP.
Level 1	Experimental The ratio of users per capita is on the order of magnitude of less than one in a thousand. There is limited availability, and use of the Internet is embryonic. Only one or a few networks are connected to the international IP network. The user community comprises principally networking technicians.
Level 2	Established The ratio of Internet users per capita is on the order of magnitude of at least one in a thousand. The user community has been expanded beyond networking technicians.
Level 3	Common The ratio of Internet users per capita is on the order of magnitude of at least one in a hundred. The infrastructure of supporting and related goods and services has become well-established, although is not necessarily extensive.
Level 4	Pervasive The Internet is pervasive. The ratio of Internet users per capita is on the order of magnitude of at least one in ten. Internet access is available as a commodity service.

Table 3.6: The Five Levels of the Pervasiveness Dimension in the Mosaic Framework (Source: [Press98])

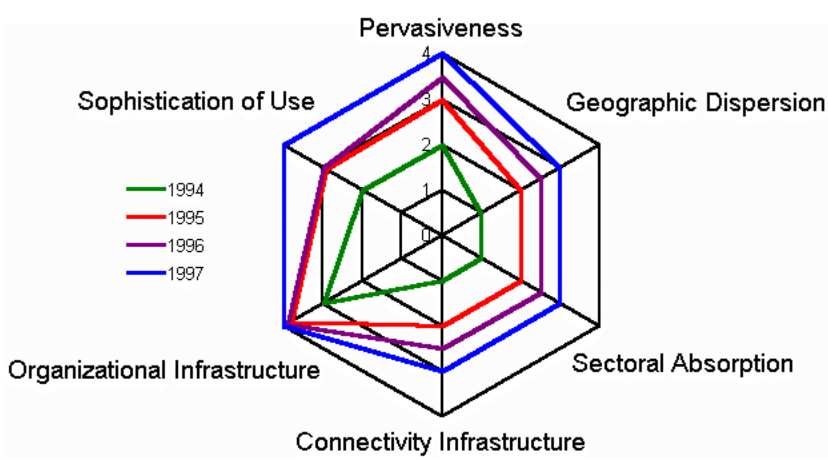


Figure 3.2: Changes in Dimension Values over Time for Finland (Source: [Press98])

3. Relating Internet Hosts to Users

It has been mentioned before how we can expect most of the metrics described above to be highly related with each other, as they all trace different manifestations of a single phenomenon, which is the expansion of the uses and users of the Internet. In Chapter 4, we base our study on international diffusion in hostcount figures rather than in user figures, because as was mentioned above there are no consistent user figures available. In this section we legitimate this approach by looking at the relationship between the limited user statistics that are available today and the hostcount figures. We find that there is a rather deterministic relationship between the two.

Figure 3.3 looks at the relationship between hosts per capita and users per capita for the top 15 countries in Internet users per capita in 1997.[32] The figure shows that there is a strong relationship between the two magnitudes, and that the number of users is initially much higher than that of hosts and progressively levels off (note that the number of users will never exceed 100%, while that of hosts is potentially unlimited). Another way to look at these results is presented in Figure 3.4, that represents the evolution of the ratio of users per hosts to the number of hosts per capita in those same countries: we see that countries with lower per capita penetration have a larger users-to-hosts ratio, and that as per-capita diffusion increases an asymptotic level is approached at around 1.5 users per host. This can be interpreted by thinking that in countries where there is "Internet scarcity" connected computers are used by many people, and that as diffusion progressed the relationship becomes closer to one user per host, or even multiple hosts per users as new and cheaper network appliances are developed .

[32] The user estimates are from the Computer Industry Almanac [#CIA].

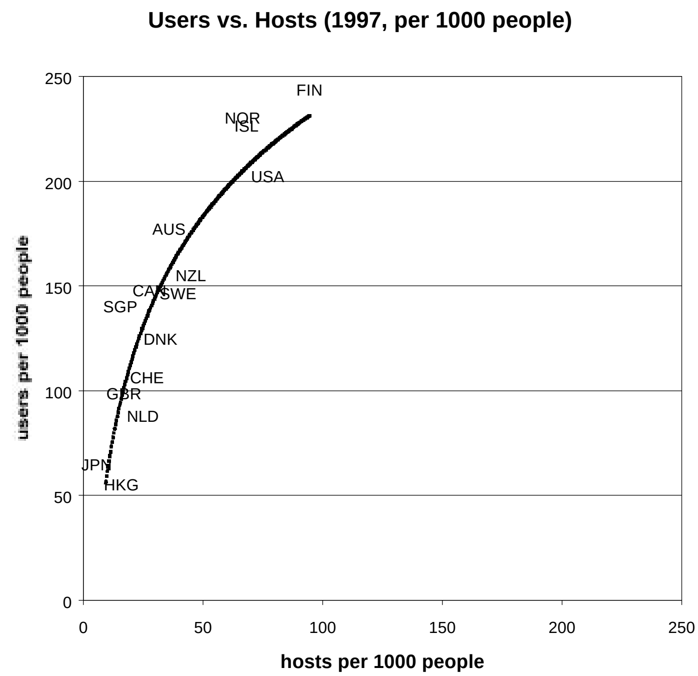


Figure 3.3: Users vs. Hosts in Selected Countries (Source: [#CIA])

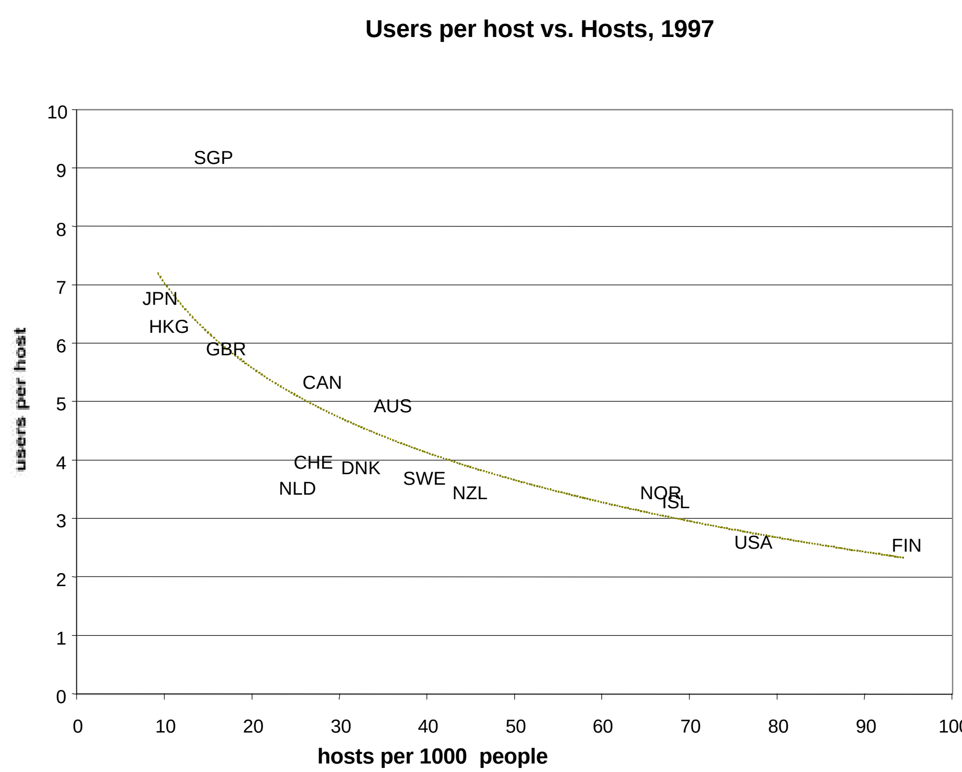


Figure 3.4: Relationship between Users per Host and Hosts per Capita

We conclude that hostcount are a good proxies for measuring Internet penetration. If we wanted to do a precise estimation of the number of users, we should make the appropriate corrections based on the fact that:

1) today, the number of users is higher than the number of hosts.

2) the ratio of users to hosts varies depending on the per capita host penetration, and can be approximated according to the curve in Figure 3.4.

3) the increase in the number of users is faster then that of host until the curve reaches the inflection point, at around 10% host per capita. For instance, going from 2.5% to 5% hosts per capita means on average an increase of 6% in users per capita (from 12% to 18%), while a similar shift in hosts per capita from 7.5% to 10% only adds about 2.5% new users.

In the remaining of this thesis, our concern is more with the relative values and growth rates across countries and across time, than on the absolute figures of users, therefore we will use directly the hostcount figures. However suggest taking into account the above relationship for any more detailed study.

4. Summary

For the rest of this thesis, Internet hosts per capita is the indicator chosen as the basis for our international comparisons. As discussed under Section §3.2.1, this metric has the advantages of allowing automatic collection, being mainly country-based, and being based on some of the longer lasting elements of the Internet (IP addresses and the DNS). Additionally the close relationship between Internet hosts and users makes it roughly equivalent in relative terms to concentrate in one or the other. Alternative metrics, in particular usage and/or compound metrics, would be more useful or more precise than hostcounts for specific purposes, but today’s availability and reliability of these data is still limited.

Chapter Four

Analysis of International Internet Development

The first part of this chapter describes how Internet diffusion can be modeled as a growth process (§4.1). Then the differences in the growth of the Internet between countries are described (§4.2) and analyzed in light of different causal factors (§4.3). Section §4.4 presents the results from a statistical model that combines time series and cross-section data between countries with various explanatory factors.

1. Internet Diffusion as a Growth Process

Figure 4.1 shows the well-known curve representing the growth in the number of Internet hosts worldwide. The curve shows an exponential growth pattern, characteristic of many processes of technology adoption and extensively studied in technical literature. In short, such diffusion models are applicable to processes in which important positive network externalities exist: the more new users adopt the technology (in the case of the Internet, the more people join the network), the more valuable the technology/network becomes to all users, both incumbent and new. Therefore, the rate of adoption tends to increase during what is called the growth stage, causing the number of users/adopters to increase at an exponential rate (Figure 4.2). The process is slowed down when an important share of the potential adopters have embraced the technology: this is the maturity stage. Finally, during the saturation stage the adoption rate drops as most potential users have adopted the technology. Additionally, the introduction of a substitutive technology can alter the general framework depending on the moment in which it is introduced. For most successful innovations, this decline stage will follow the saturation one. It may also happen that the introduction of new uses for an existing technology would cause new diffusion waves.

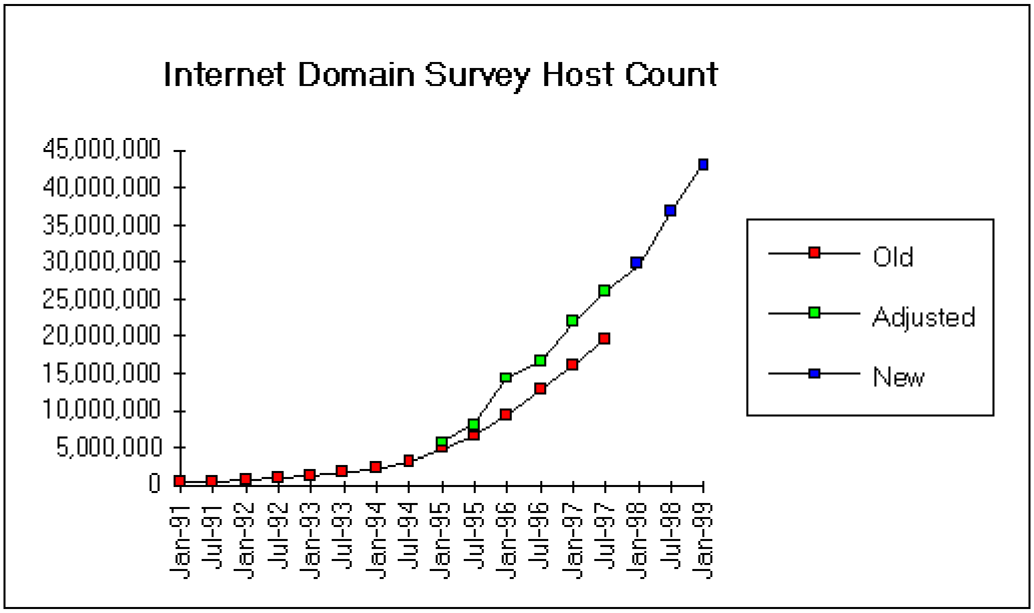


Figure 4.1: Internet Host Growth. Source: Network Wizards [#NW]

The case of the Internet is in a sense paradigmatic of such technological diffusion process favored by a strong network effect, and can therefore be modeled as an S-shape process that can be used to forecast the future growth. In its analysis, [Shuster98] finds the parameters of the growth process and predicts a saturation level of 80 million hosts worldwide, a take over time of 7.2 years (the takeover time is the time required for adoption to proceed from 10% to 90% of the saturation point) and a halfway point in October 1998 (the halfway point is the point at which an innovation reaches 50% of the saturation level). However, modeling growth process by adjusting growth curves leave important indeterminations, especially when done in the early stages of adoption: for instance, the forecast of the saturation level is often imprecise: Shuster estimated it to be 80 million hosts based on statistical projections, but this number seems too conservative today.[33] According to our discussion on the relationship of hosts to users, we could expect that the two numbers will become closer with time. In this case, is the entire world

[33] The latest survey results (January 1999) yielded over 43 million hosts, and no downward trend is observed.

population (circa 6 billion people/hosts) the roof of our "potential adopters", or are there large parts of the world that we can imagine are going to be indefinitely excluded from that category? The problem here is that the answer to that question is dynamic, and depends on the time horizon of our projections: with half the world population still waiting for their first phone call and 40% unable to read, it is very unlikely that in the short term, those uneducated and unconnected inhabitants will ever become Internet users. But what is special about the Internet is that if one believes in its potential to achieve cost effective connectivity and to be used as a powerful communication tool, then it is very important to include these potential users when forecasting.

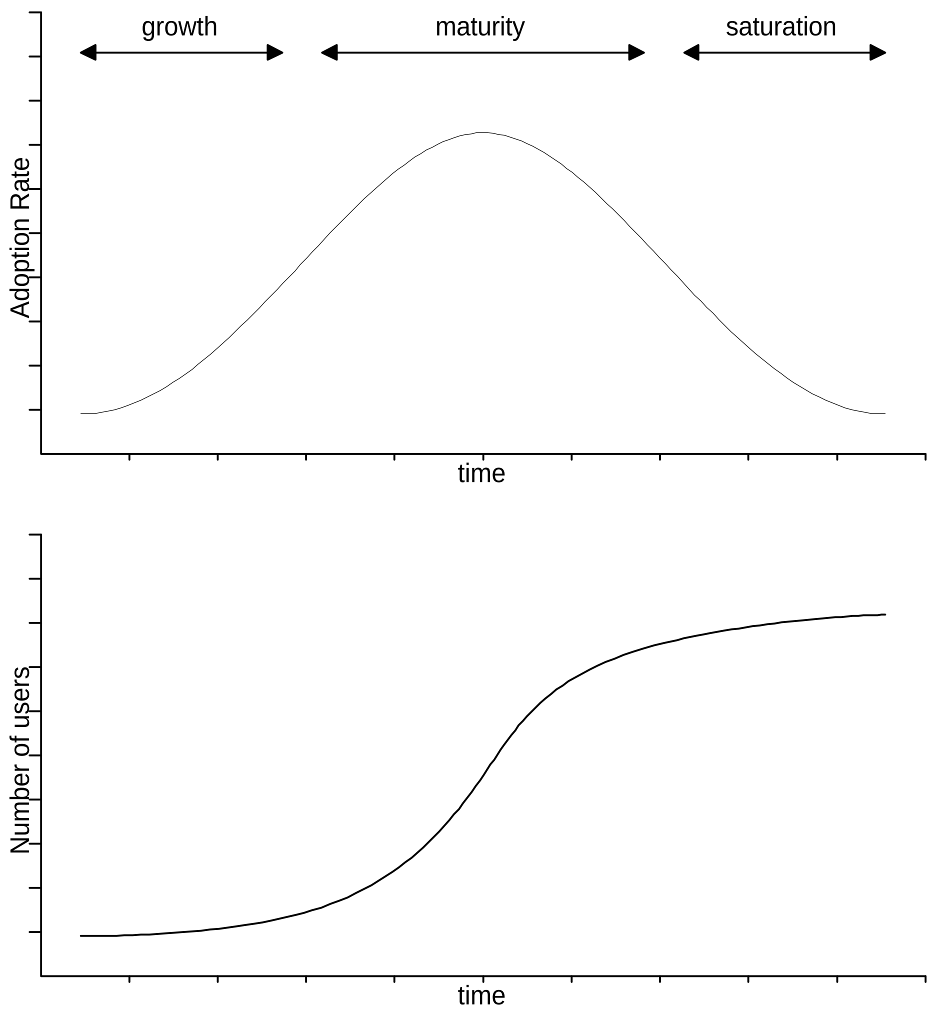


Figure 4.2: Adoption and Diffusion Curves

A second complexity comes from the continuously changing and expanding panoply of Internet services: as the Internet evolves from being a medium for exchanging email and computer files to a tool for commerce and entertainment, are we observing a single diffusion process or a series of superimposed trends caused by the development of new services over the Internet?

For the purpose of this thesis, which is the analysis of the cross country differences in Internet diffusion, we are not so much interested in entering in the theoretical details of the diffusion model, as in determining which parameter better represents the phenomenon we are analyzing: Internet diffusion over time. One approach might be to fit a diffusion curve such as the one in Figure 4.2 to each of the countries and then use its parameters as the proxies for that country. Two drawbacks of this methodology are that 1), the year-to-year growth rates for individual countries vary a lot, making it difficult to adjust a growth curve with enough statistical precision and 2), the relationship between the growth curve parameters and the process itself is not always transparent. We find that an almost equivalent way to look at the process is by measuring the compounded average growth rate (CAGR) over a period of years for each country. This, together with the initial level of diffusion provides a good representation of the growth process (Figure 4.3).

Figure 4.3 shows the relationship between the 94-98 compounded annual growth rate and the initial penetration level for the world's top 50 economies. As we would expect according to the diffusion model, the rate of adoption does decrease with adoption – i.e. the further along the country was in 94, the more slowly it grew during the 94-98 period. However, some comments can be made at this point: first, if the less developed economies were added to the chart, we would probably find a worse fit to the curve because of large swings in diffusion metrics at the very early stages of diffusion (for many of those countries 94-98 data are not available). Second, the slope of the curve is relatively flat (=-0.0695), meaning that the "catch-up" or rate at which laggard countries approach levels of more advanced countries is very slow.[34] Third, there remains a high variety between countries (the initial level of diffusion only explains about 50% of the subsequent growth). The remainder of this chapter is devoted to the explanation of those differences.

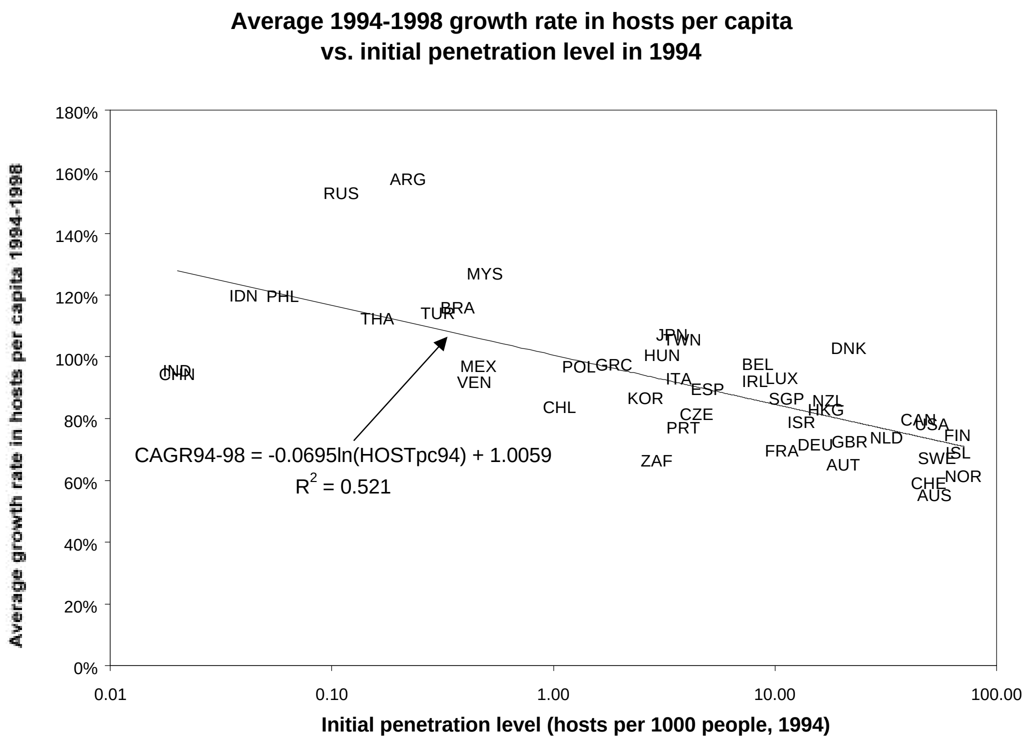


Figure 4.3: The Relationship between Average Growth Rate of Hosts per Capita and the Initial Penetration Level[35] (Source: [ITU99])

[34] At those rates, it will take for instance about 8 years for Thailand to get to the same level of diffusion the Netherlands had in 1994. And by that time, the Netherlands will be much ahead, meaning the catch-up will happen far away in the future, if ever.

[35] Each country is represented by a three letter acronym. A list of country acronyms can be found in Appendix C.

2. Disparities in Access: Geography, National Income and Internet Diffusion

This section first describes the disparities in Internet Diffusion worldwide. Next section will explore the reasons for such disparities.

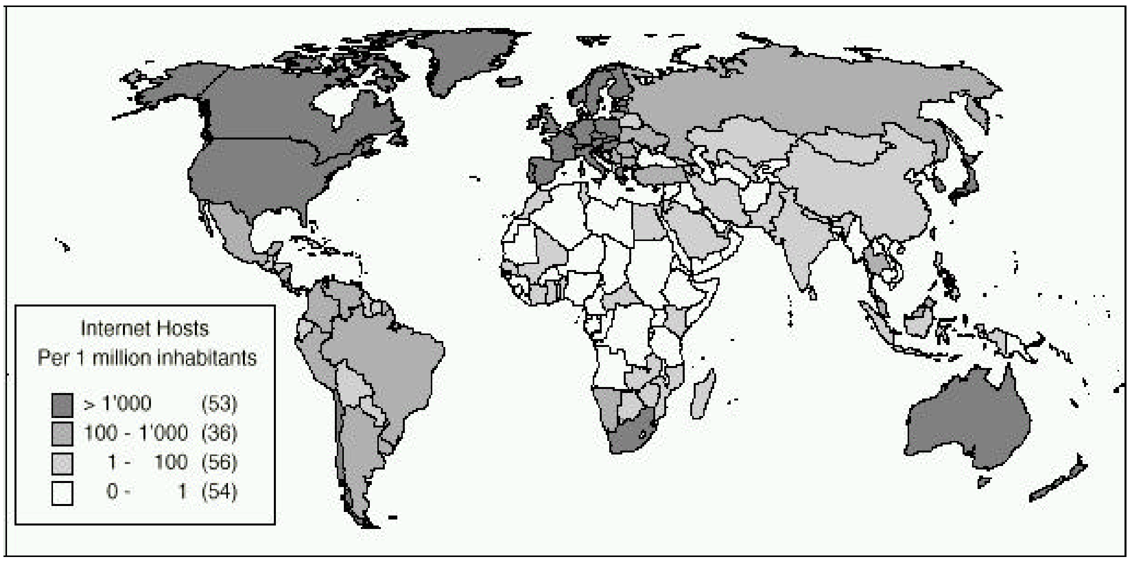


Figure 4.4: Internet Hosts per 1 million People, 1996 (Source: [ITU97])

Figure 4.4 gives an overall picture of the distribution of the Internet around the planet. Geographic and economic disparities in Internet diffusion are well known: in general, North America has had historically the largest portion of hosts and users, mainly due to the fact that this is where the Internet was developed. Europe continues to lag behind the U.S., followed by Asia and, much farther behind, Africa. Figure 4.5 shows the evolution of the total number of hosts per region in the past 5 years.

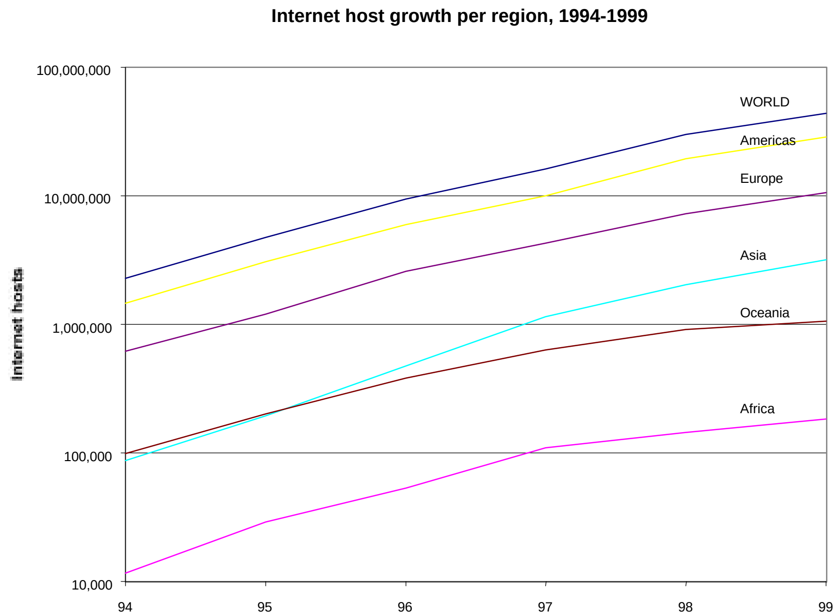


Figure 4.5

In terms of countries' economic wealth, the situation is even more skewed towards high income countries, as can be seen in Figure 4.6: today 95% of Internet hosts are located in high income countries, that represent only about 15% of the world's population.

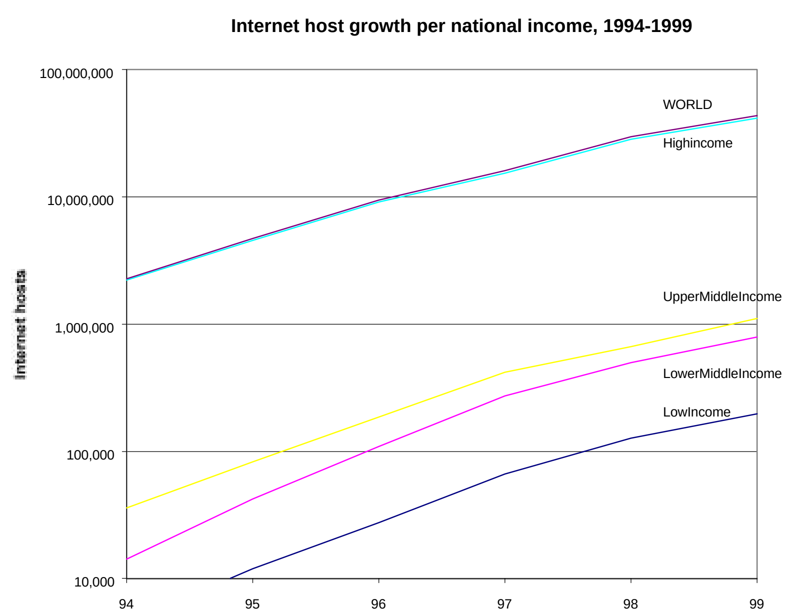


Figure 4.6

Although many envision the Internet as a tool with a very high development potential because of the relatively low cost and flexibility of its deployment, the truth is that today the use of Internet Services is much more skewed towards high income, highly developed countries than any other large-scale communication service. Figure 4.7 shows how 90% of Internet hosts are located in countries totaling only about 15% of the world population, while the same 90% cumulative level is only attained for Personal Computers at 25% of the population, for total GDP at 30%, and on the opposite end telephone lines are relatively more equally distributed (although 90% of the lines are still in countries with 50% of the population).

Note that this graph, although clarifying, would only reflect the real picture if diffusion within each country was homogenous: however, there are in fact important regional and other differences inside each country, meaning that the real distribution would be even more skewed than the country-averaged presented here shows.[36]

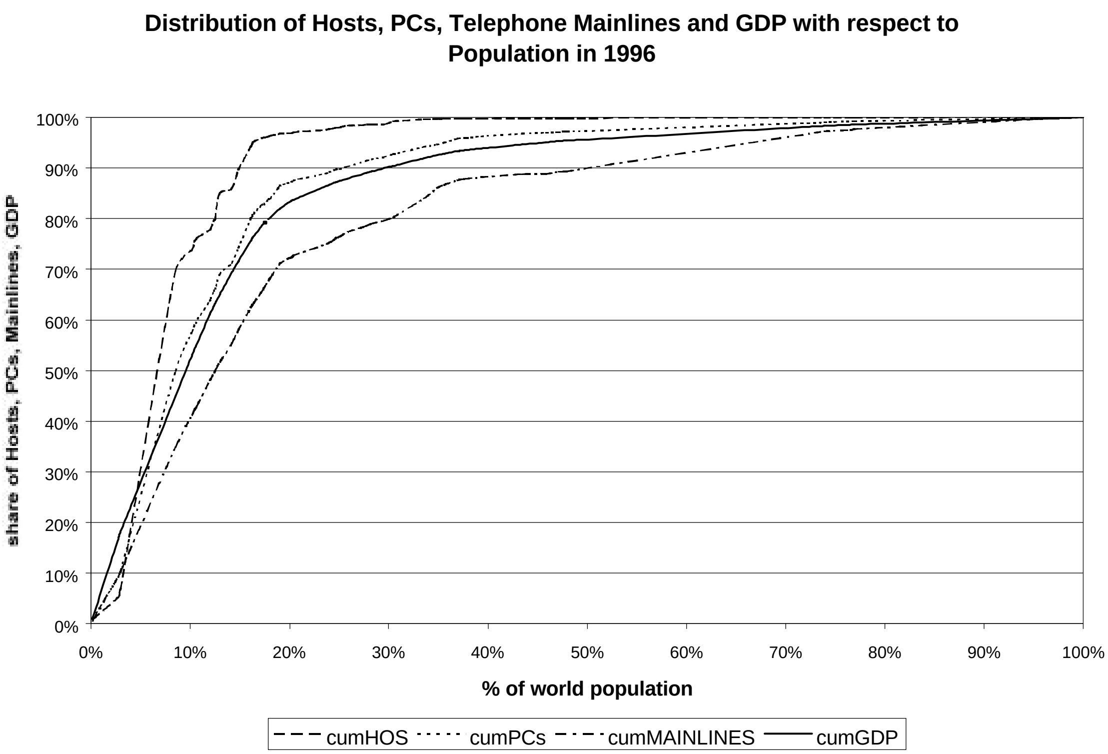


Figure 4.7

[36] See, for example, the distribution of Internet access in the U.S. studied in [Downes99].

One might think that these inequalities are due to the fact that the technology is still new and has still to diffuse from high-tech countries to the rest of the world. In that case, we would expect to see a higher rate of diffusion in laggard countries, but statistical evidence does not clearly indicate this effect: in fact, the growth rates of Internet hosts have consistently been decreasing and converging for most countries since 1994 (Figure 4.8 (a)). In 1998-99, the average growth rate for high and low income groups of countries of different income levels was very similar and close to 60%. Taking into account the very different initial levels, this means that at the current rates the time in which developing countries will catch up seems remote.

The important regional differences seen in Figure 4.5, although long lasting, do not seem as persistent as those related with national income: for instance, Figure 4.8 (b) shows how the Asian region achieved above-average growth during the period 1994-1998, leading to a significant change in the Asian share of the overall hosts and users.

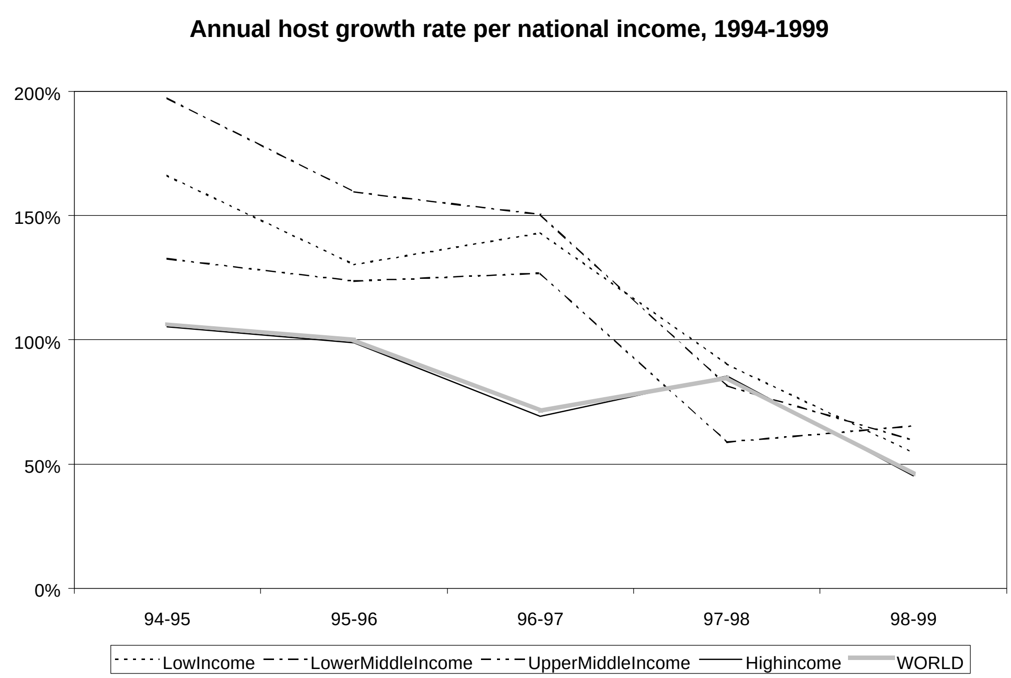
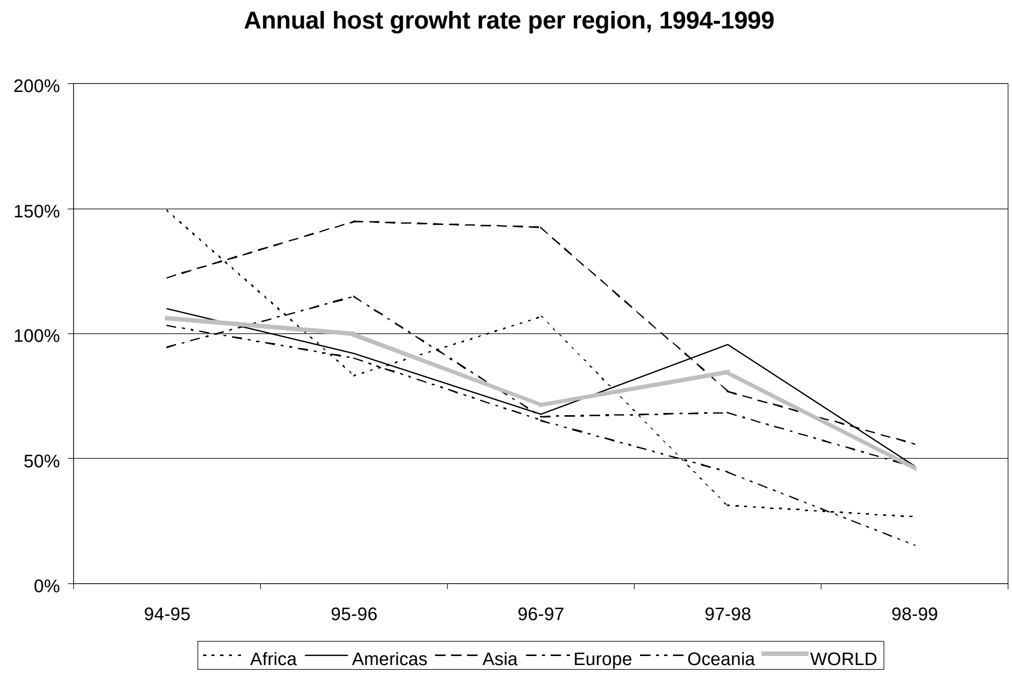


Figure 4.8 (a)



Annual host growht rate per region, 1994-1999
200%
150%
100%
50%
0%
94-95
95-96
96-97
97-98
98-99
Africa
Americas
Asia
Europe
Oceania
WORLD

Figure 4.8 (b)

3. Explaining Cross-Country Differences

Observing and describing the disparities between countries is one thing, but it is more interesting to explain the causes of such differences and even more to identify countries that do not follow the trend and analyze why. We have seen in the previous section that a country's wealth is strongly correlated with its level of Internet diffusion. Section §4.3.1 quantifies this effect. Also, as we discussed in Chapter 3, end to end connectivity relies on a number of network elements. The availability of those elements is likely to influence diffusion of Internet services. Such effect is traced in Section §4.3.2. Finally, Section §4.3.3 investigates the effect of costs.

3.1. Statistical Models

The statistical models used through this section are classical multiple regression models, using ordinary least squares –OLS– estimation (a more sophisticated model is used and explained in Section §4.4). The level of significance chosen is 5%, the standard in statistical analysis (the level of significance is related to the precision with which the confidence intervals of the regression estimates are likely to contain the true regression parameters. (See [Pindyck91] for a detailed discussion of multiple regression).

The regression output tables show the following information:

- ***R^2*** (coefficient of determination): the proportion of the total variation in the dependent variable explained by the regression on the independent variables, and lies between 0 and 1. Roughly speaking, a high value of R^2 is associated with a good fit of the regression and a low value of R^2 with a poor fit. The adjusted R^2 corrects for the degrees of freedom.
- **Coefficients**: estimated ’s for each variable and intercept (see equation [4.1]).
- **Standard Error**: standard error of the estimated coefficients ’s and (provides a measure of the dispersion of the estimates).
- **t-Stat** (t statistic): allows to test the null hypothesis of a coefficient (i.e., the hypothesis of the coefficient being 0) at a certain level of significance. For large

samples and a 5% level of significance, a t-value of around 2 or more allows to reject the null hypothesis.

- **P-value** (Probability value): Describes the exact level of significance associated with each coefficients. A P-value of over .05 indicates that the coefficient is not significant at the 5% significance level.

3.2. The Effect of Wealth

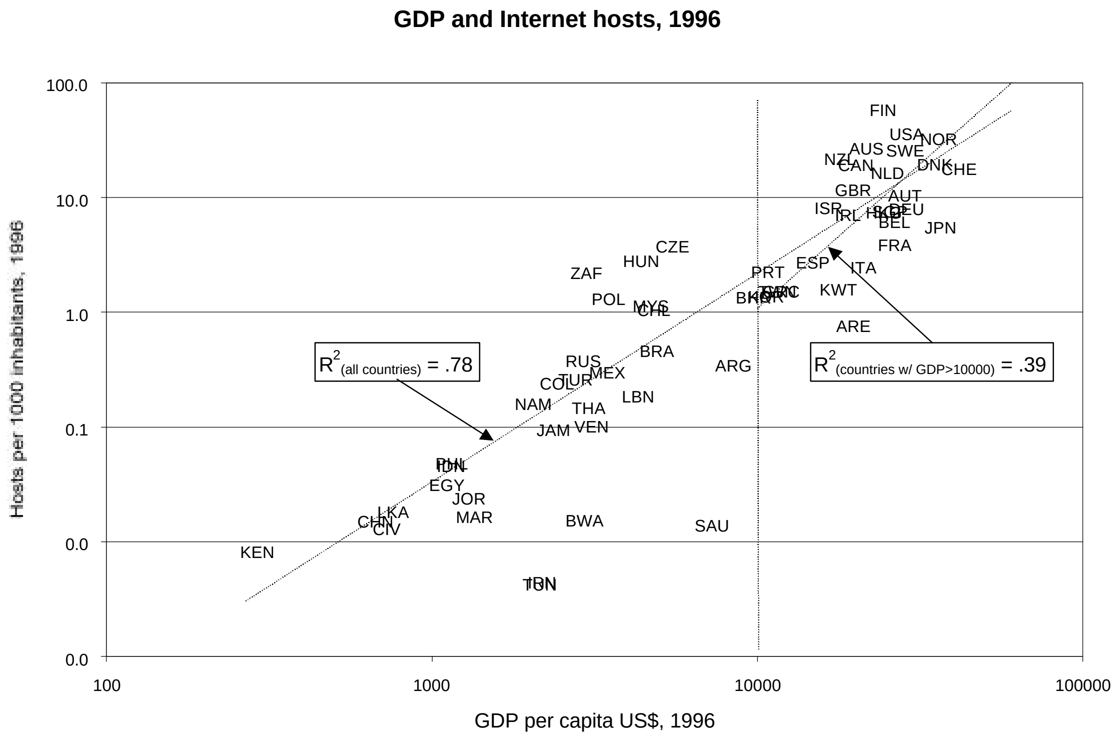


Figure 4.9

The first, almost obvious finding when studying Internet diffusion across countries is that wealth and Internet use are highly correlated: the wealthier a country, the higher the level of Internet Access. Such dependence can be investigated by studying the relationship between the Gross Domestic Product per capita and the number of Internet hosts per capita (Figure 4.9). This result is to be expected as most of the factors involved in Internet diffusion such as availability of infrastructure and education do themselves depend on wealth. But although per capita income does play an important role, the analysis shows that there remain very important disparities that cannot be explained just in terms of wealth. For instance, in 1996:

1) some countries' level of Internet development is much lower or higher than what could be predicted by its wealth level: For instance, Finland has a 40% less per capita income than Japan, yet it has roughly 10 times more Internet hosts per capita. Similarly, with similar income per capita, Italy and France have many fewer Internet hosts per capita than Sweden and Australia. Other "anomalies" are most Eastern European countries such as Hungary or the Czech Republic, with more Internet hosts than expected, and Japan and some North African countries (Algeria, Tunisia) with fewer users than expected.

2) for countries with similar levels of wealth, the relationship between wealth and Internet use becomes much weaker, and is not sufficient to explain the differences among countries. Table 4.1 shows the results from the regression for countries with a 1996 per capita income of over and under $10,000. We get coefficients of determination of R^2=.39 and R^2=.43 respectively, compared to R^2=.78 when all 59 countries are included.

Countries Included in Model	Fraction of variance explained (R^2)
all 59 countries in sample	.78
28 countries w/$GDPpc_{1996}$ >$10,000	.39
31 countries w/$GDPpc_{1996}$ <$10,000	.43

Table 4.1: Fractions of Variance Explained by GDP

An alternative statistic that could be used instead of the per capita GDP is the Human Development Index (HDI), a metric evaluated by the UN and based on three indicators: life expectancy, education and per capita GDP. The ITU has done this analysis and the results are similar (see [ITU99], p. 24).

3.3. The Effect of Infrastructure Availability : Teledensity and PC Penetration.

Teledensity, or number of lines per 1000 inhabitants, is one of the most used telecommunication infrastructure indicators. Although a very rough measure of the communication infrastructure in a country, this metric has the advantage of being clearly defined and regularly collected by the ITU and other international organizations. The problem is that at the fast pace of development of telecommunications technologies, there is a growing number of alternative means to achieve connectivity other than by means of telephone mainlines: cellular, high speed circuits serving multiple users, satellite links, etc. Therefore the adequacy of teledensity as a single infrastructure metric is limited today.[37]

PC penetration per country is another relatively well known, well traced metric. Here again, the use of this metric as a proxy for availability of equipment enabling Internet diffusion raises some caveats: as we described in Chapter 2, PCs are only one among various possible terminals that can be used to access the net. Although the share of PCs is dominant today, this share is likely to decrease in the future. Besides, the statistics available are usually obtained from the sales figures, but little is known about the average useful life of the equipment. Nor is a brand new 1999 Pentium III computer comparable to a 1994 intel-486 as an Internet access device.

Figure 4.10 and 4.11 show the strong relationship that holds between the per capita numbers of telephones lines, PCs and Internet hosts in a sample of 59 countries.[38]

[37] Even if we chose teledensity as a good proxy, some caveats still apply: is the relevant metric lines in use, or total installed lines (which includes idle lines)? And should multiple lines going into the same location be counted the same as individual lines? One seems to be able to find arguments in either direction.

[38] The countries in the sample are countries for which data was available for 1996 and include most World's mayor economies.

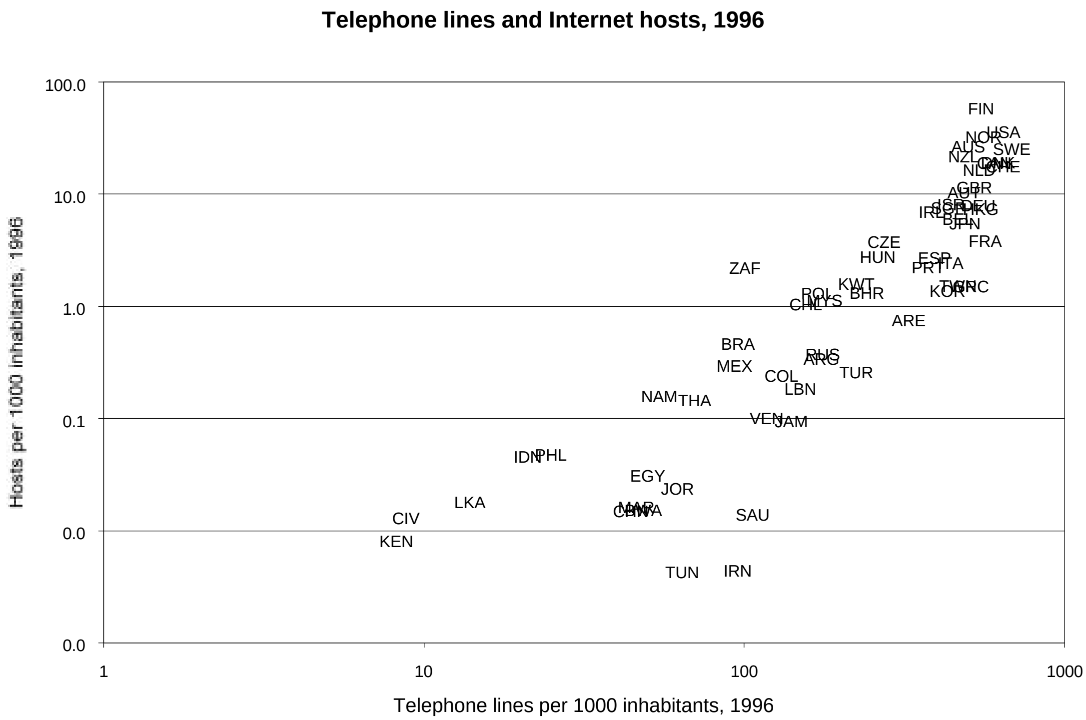


Figure 4.10: Internet Hosts Compared to Telephone Lines (Scales are Logarithmic) (Source : [#ITU])

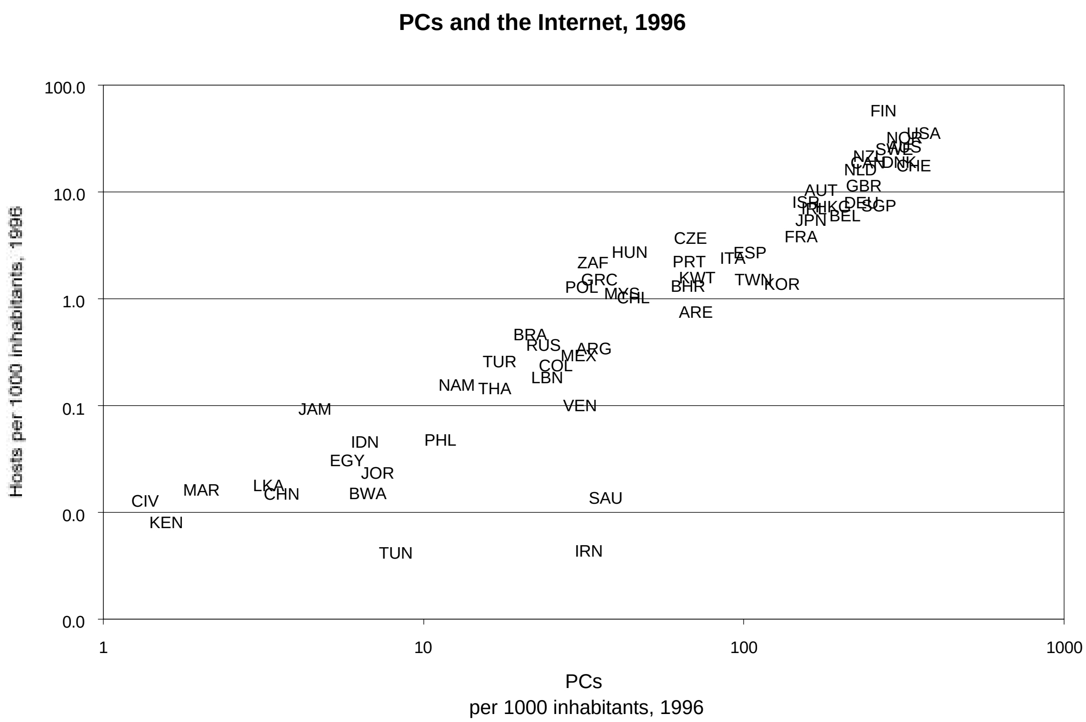


Figure 4.11: Internet Hosts Compared to PCs (Scales are Logarithmic) (Source : [#ITU])

Given the strong correlation found with each of the last three variables considered, it is tempting to go one step further and hypothesize that the combined values of a country's GDP per capita (*GDPpc*), Phone lines per capita (*LINpc*) and PCs per capita (*PCpc*) could be used to predict an accurate estimate of a country's Internet penetration measured in hosts per capita (*HOSpc*). Such hypothesis can be tested by running a multiple regression according to equation [4.1]. Table 4.2 shows the output of such regression.

$$\ln(HOSpc) = a + b_1 \ln(GDPpc) + b_2 \ln(LINpc) + b_3 \ln(PCpc) \qquad [4.1]$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	(8.02)	2.10	(3.82)	0.00
ln(LINp1000c)	0.48	0.38	1.27	0.21
ln(PCp1000c)	1.17	0.32	3.69	0.00
ln(GDPpc)	0.10	0.40	0.25	0.80
Adjusted R Square = 0.83				

Table 4.2: Multiple Regression Coefficients and Significance Levels

Unfortunately, the output from the multiple regression shows that all but one of the coefficients has a significant value, preventing us from affirming that *LINpc* and *GDPpc* are statistically significant in determining *HOSpc*. Moreover, the adjusted coefficient of determination of the overall regression is roughly the same as the one obtained by regressing only in the more strongly correlated variable, *PCpc* (Figure 4.11). In other words, we find that the best available predictor of Internet penetration in a country is the number of available PCs, and that almost nothing is added to this prediction by considering the country's teledensity or per capita GDP.

The origin of this limitation lies in the correlation between the three independent variables considered, shown in Table 4.3. The three variables are strongly correlated, in what is known as a high degree of multicollinearity. Multicollinearity arises when two or more variables are highly correlated, and makes the interpretation of the regression coefficients hazardous.[39] When multicollinearity is present, the standard errors of the coefficients

[39] for a discussion of multicollinearity see for instance [PINDYCK91], p. 83.

estimated by OLS tend to increase, and the reliance that can be placed in the coefficient values decreases. The remedy is to take some of the correlated variables out of the model, with the loss in precision that this implies. In our particular case, we find that we can not find more than one variable at a time to be a statistically significant determinant of Internet diffusion, with the best predictor being PC penetration.

	ln(HOSp1000c)	*ln(LINp1000c)*	*ln(PCp1000c)*	*ln(GDPpc)*
ln(HOSp1000c)	1.00			
ln(LINp1000c)	0.88	1.00		
ln(PCp1000c)	0.91	0.93	1.00	
ln(GDPpc)	0.88	0.93	0.95	1.00

Table 4.3: Correlation between Hosts, Teledensity, PCs and GDP per capita (in logarithms)

Summary of Findings

In the end, we are left with the result that PC availability is key for Internet diffusion, and we have also found that PC availability is highly correlated with a country's income level and teledensity. But what we cannot extract from those results is any idea of causality between variables: for instance, if we want to increase Internet penetration in a country, how useful is it to increase PC penetration (e.g., by subsidizing PC sales), without altering a country's wealth and teledensity? In our current data, those variables are too highly correlated to let us analyze all those effects.

The idea that we can not get any more insightful result than a single predictor at a time is not very satisfactory. What can we do in order to see the combined effects of some other of the determinants we discussed in Chapter 2? The problem is that with the data available we are faced with a tradeoff between generality and detail. We can build a model that will rather accurately work for all countries of the world, countries as diverse as Kenya, India, France or Finland. But the whole economic environment in those countries is so different that the model fails to capture determinants of Internet diffusion other than a basic correlation with wealth, teledensity or computer penetration. If we want to refine the analysis to take into account additional factors acting simultaneously, the only way to go is to restrict the sample to a group of countries with similar wealth and infrastructure characteristics (unless we could find a data set with better statistical characteristics – this is the intent of some of qualitative "compound" metrics described in Section §3.1.4).

The utility of such reduced-sample analysis is also raised by the fact that when we consider those countries with a more advanced level of wealth and Internet penetration, the relationship between both becomes weaker, leaving space for other explanatory variables.

3.4. The Effect of Telecommunication Costs

Internet Access Costs

In this section we aim to analyze the effect of costs on Internet diffusion, following the discussion on costs in Section §2.3.3. We start by focusing on the relationship between Internet diffusion and Internet access costs (the sum of IAP's access costs and telephone charges that the user has to pay in order to connect to the Internet). Do countries with lower costs have a higher penetration level? We begin by analyzing the plain relationship between access costs and Internet penetration (Figure 4.12). Surprisingly, the chart shows no pattern in the sense of a relationship between the two variables. But if we scale the access costs as a percentage of the countries' GDP, a clear downward trend emerges showing that countries with lower relative costs do achieve higher Internet penetration.

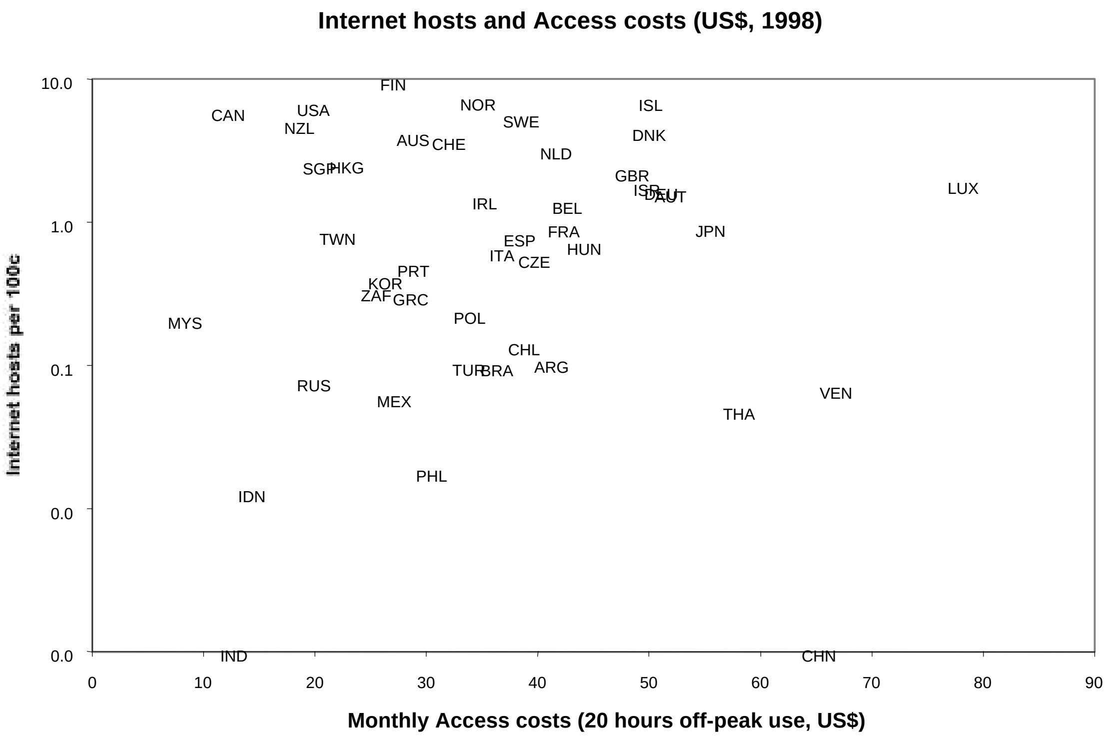


Figure 4.12

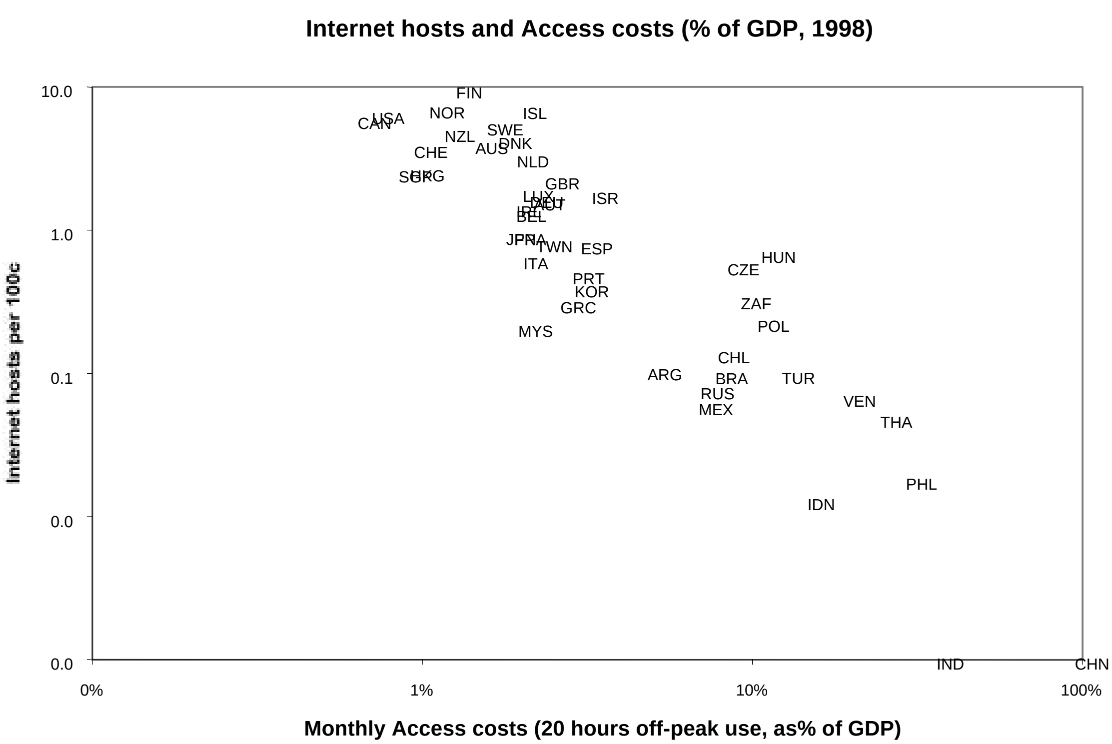


Figure 4.13

However, we face a similar problem as in §4.3.2 with respect to causality: are lower relative prices the *cause* of higher adoption of the Internet, or are both phenomena correlated but caused by another factor, such as wealthier, more productive economies? One way to address this issue would be to trace dynamically the evolution of costs and Internet penetration level across countries, and see if falling access prices lead to increased adoption rates, all other things being equal. Unfortunately, data on the historical evolution of Internet access costs for each country are not available. The analysis that follows is an attempt to overcome this obstacle and find a cross-country effect of costs on Internet diffusion.

In order to proceed, we make the following assumption: *the 1998 cost figures can be used as a proxy of the relative cost levels of Internet access in each of the countries of the sample for the period 1994-1998*. This is a strong assumption, as 1) very little information on historical costs is available, even for individual countries, and 2) important changes in access costs are likely to have happened during this period. However, given the scarcity of available cost data, making this hypothesis allows us to build a model in which we include the 1998 access cost ($ACCOS_{98}$) as a percentage of per capita GDP as one of the determinants of penetration growth ($HOS_{cagr94\text{-}98}$ – The compounded average growth rate of Internet hosts per capita between 1994 and 1998). The other factors included are the initial penetration level in 1994 (see §4.2.1), and 2 control variables $GDPpc_{97}$ (per capita GDP in 1997) and $PCpc_{97}$ (PCs per 1000 people in 1997), that we include based on the significant relationships found in §4.2.2. The resulting model has the following form:

$$\begin{aligned} HOScagr_{94-98} &= a + b_1 \ln(GDPpc_{97}) + b_2 \ln(HOSpc_{94}) \\ &\quad + b_3 \ln(PCpc_{97}) + b_4 \ln(ACCOS_{98}) \end{aligned} \qquad [4.2]$$

	Coefficients	Standard Error	t Stat	P-value	Signif at 95%
Intercept	(0.48)	0.40	(1.20)	0.24	
ln(GDPpc97)	0.05	0.07	0.82	0.42	
ln(HOSp(1000)c94	(0.18)	0.03	(7.14)	0.00	*
ln(PCp(1000)c97)	0.09	0.07	1.25	0.22	
ln(ACCOS as% of GDPpc)	(0.09)	0.04	(2.24)	0.03	
Adjusted R Square	0.70				

Table 4.4

	Coefficients	Standard Error	t Stat	P-value	Signif at 95%
Intercept	0.28	0.13	2.16	0.04	*
ln(HOSp(1000)c94	(0.14)	0.02	(8.04)	0.00	*
ln(ACCOS as% of GDPpc)	(0.15)	0.03	(4.54)	0.00	*
Adjusted R Square	0.66				

Table 4.5

The output from the regression is shown on Tables 4.4 and 4.5, before and after dropping the non significant variables (*GDPpc* and *PCpc*). The resulting model found is:

$$HOScagr_{94-98} = .28 - .14\ln(HOSpc_{94}) - .15\ln(ACCOS_{98}) \qquad [4.3]$$

The coefficient of $ACCOS_{98}$ ($_{4}$ = -.15) can be interpreted as the elasticity of Internet diffusion growth to access costs, i.e. by how much does the rate of growth vary when access costs vary.[40]

[40] The elasticity figure found is in fact relatively low: for instance, the effect of a 50% drop in access prices would only increase Internet growth by 15%, while a 100% increase will only cause a 15% drop in the growth rate (note the effect of logarithms in [4.3]). Given that the average growth rate for the period considered is close to 90%, growth is relatively inelastic to variations in access costs.

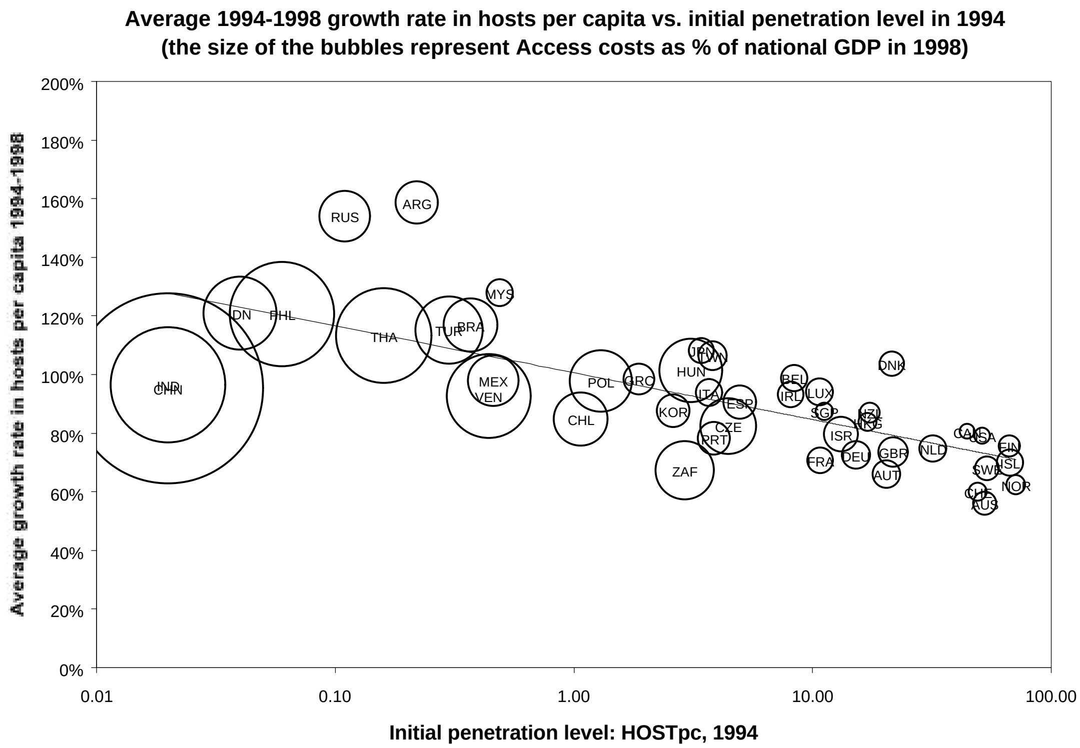


Figure 4.14

Figure 4.14 illustrates graphically the effect traced in this section: in it, each country's host compound average growth rate (CAGR) over the period 94-97 has been plotted against its initial penetration level, as in Figure 4.3. But we have represented the access costs relative to their GDP as the size of the bubble. The figure shows that countries with lower relative access costs such as Russia, Argentina or Malaysia achieved a significant higher growth over that period than China, India, Venezuela or South Africa, that have higher that average relative access costs.

The result is very rough, mainly because we use ITU's basket price in 1998 as a cost proxy for the whole period. Most importantly, because a "typical" usage time of 20 hours per week is chosen in order to construct the basket, the possible effect of metered vs. flat pricing on usage time is not reflected (i.e. for instance how does flat rate pricing encourage longer on-line use). We believe the limited availability of historical cross country pricing data, together with the variability in pricing schemes from one country to another make it very difficult to achieve any more precise quantitative conclusions today that would be valid across a large set of countries. However, it is possible to get a more precise measure of cost elasticity by analyzing one particular country or a smaller set of homogenous countries. In Section §4.4 we present an attempt to produce such a model, by limiting our sample to more a more homogenous and documented set of OECD countries.

Leased Line Costs

In Chapter 2 we discussed how leased lines are a key element of Internet infrastructure because they are used for providing connectivity both from large end-users to ISPs and between ISPs themselves and the Internet backbone. We now wish to investigate the relationship between the cost of leased lines and Internet penetration and get some insights about causality between the two by looking at the relationship with Internet Access costs. We restrict our sample to OECD countries because of data availability.

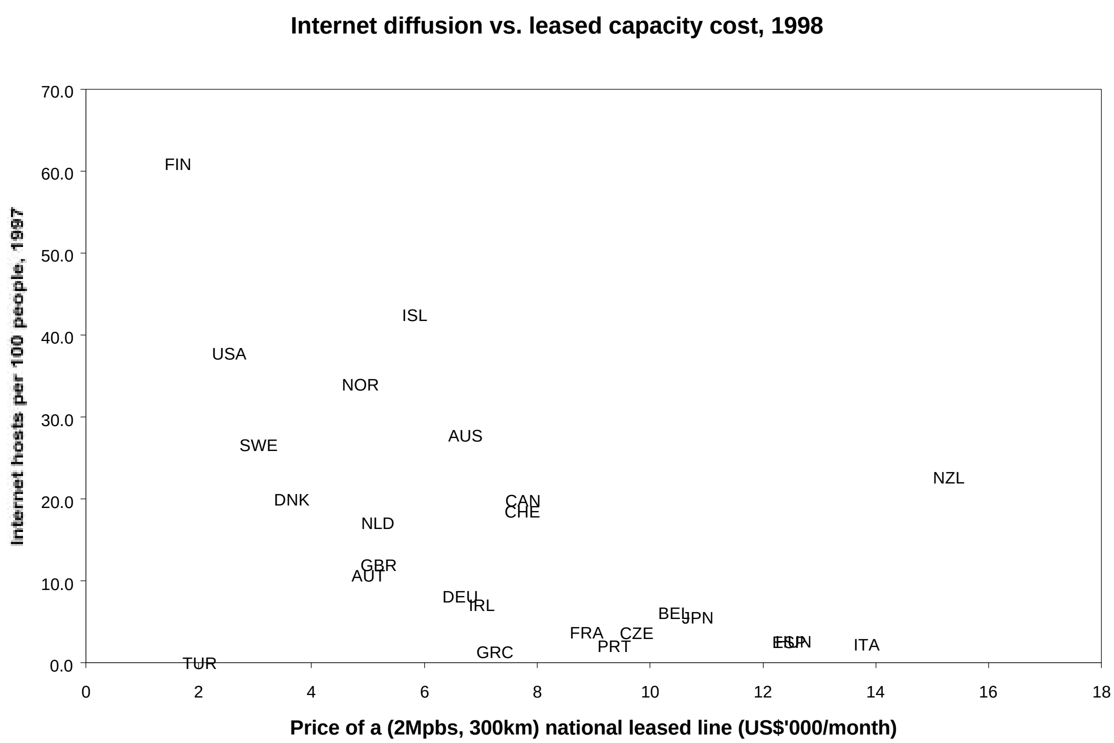


Figure 4.15

Is there a relationship between a country's cost of leased capacity and its level of Internet diffusion? Figure 4.15 shows the relationship between the two, using the price of a 2

Mbps, 300 km leased line as a proxy for leased capacity costs.[41] The figure shows a strong correlation between the two variable across the OECD, with the exceptions of Turkey (with a very low price of leased capacity) and New Zealand (with a very high cost, compared to countries with similar penetration).

	Coefficients	Standard Error	t Stat	P-value	sign. at 95%
Intercept	0.58	0.52	1.10	0.28	
ln(GDPpc97)	(0.05)	0.08	(0.59)	0.56	
ln(HOSp(1000)c94	(0.16)	0.04	(4.11)	0.00	*
ln(PCp(1000)c97)	0.13	0.09	1.39	0.18	
ln(ACCOS98)	(0.05)	0.06	(0.75)	0.46	
ln(LLCOS97)	(0.04)	0.04	(1.01)	0.32	
Adjusted R Square	0.59				

Table 4.6

It its more complicated to investigate the possible causality between leased capacity costs and Internet diffusion: are lower costs a facilitator of diffusion, or does it just happen that diffusion is higher in countries with a more developed and competitive telecommunications sector, in which costs are lower? Lacking reliable time series of leased capacity that we could use to compare to diffusion time series, we are drawn to a similar analysis to the one performed in the previous section with respect to access cost.

Table 4.6 shows the results of regressing the average growth rate of the Internet in the same determinant variables as in model [4.2], plus our proxy for leased capacity costs, the natural logarithm of the price of a 2 mbps, 300 km national link *ln(LLCOS)*. The sample is reduced to the 25 countries for which data on leased lines costs is available. The estimation results show coefficients with the expected signs (both higher leased capacity costs and access costs lower the diffusion rate), but the statistical significance of the relationship is weak and none of the coefficients is significant at the 95% confidence level.

---

[41] Using a single price as a proxy for leased capacity cost is a rough approximation: as seen in Chapter 2, leased line prices are often highly variable and arbitrary. One could try to construct a "leased capacity cost basket" based on the most used leased links in each particular country, that would depend on geography and ISPs market structure; this lies beyond the scope of this thesis. For our purpose, we will assume that the price chosen gives at least a good indication of the relative prices of leased capacity across countries.

This means that we cannot conclude that lower costs have had a significant impact on the rate of diffusion of the Internet over the past 5 years.

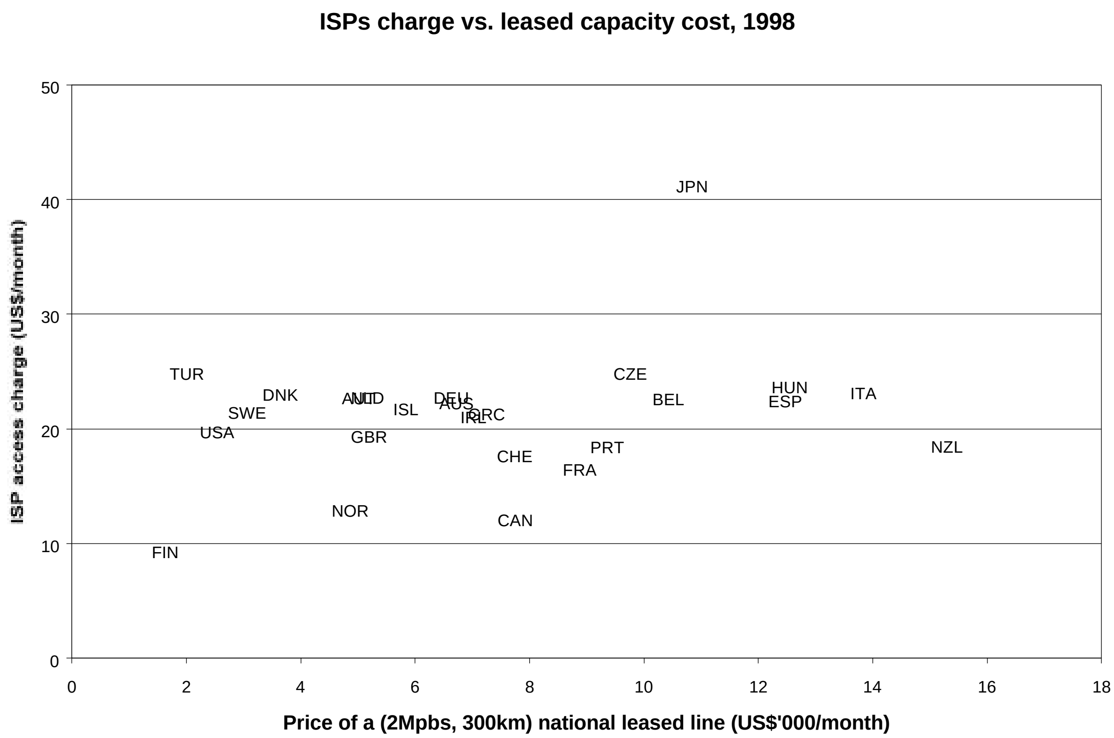


Figure 4.16

If this result is counter-intuitive, we can try to illustrate some of the apparent paradoxes in Internet access pricing by looking at the relationship between leased capacity and Internet Access pricing. With transport cost accounting for a large share of ISPs' costs, we would expect to see some correlation between the two. Figure 4.16 shows none of this pattern: ISPs' charge appear to lay in a rather narrow range of $15 to $25 per month for most countries in the sample, while leased capacity costs are widely distributed over almost a whole order of magnitude (with costs in New Zealand, Italy or Spain 6 to 8 times as high as in Finland, the U.S. or Sweden).

The fact that capacity costs do not seem to translate into access charges is surprising and calls for further cost analysis beyond our scope. However, if we believe the results of Leida's ISPs cost model [Leida98], ISPs' prices are very close to costs in the U.S. The

U.S. is one of the countries with lower capacity costs. Therefore, how can most OECD countries afford similar ISP's charges when their transport costs are so much higher? A possible explanation is underpricing: ISPs' prices in most countries might not reflect real costs, and could be cross-subsidized. This is more likely to be the case in countries in which the main ISPs are companies with large interests in other areas of the telecommunication sector, such as incumbent operators or new players in the fixed, mobile or data networks field. In the case of local access providers, subsidizing may happen through usage: ISPs' prices are cheap but costs are recovered via metered access calls. This would be a barrier to pure ISPs' entrants to compete in such a market, and deserves further study, for which a possible starting point would be to know the share of users in each country whose ISP is closely related with some of the larger players in the industry.

4. Pooled Time Series and Cross-Section Model

In the previous section we have identified the broad factors determining Internet diffusion across countries. We have shown how national wealth, infrastructure development and costs play a mayor role, but also that there remain unexplained disparities in the diffusion process between countries, in particular in those countries with a higher level of penetration. In this section we wish to refine the analysis by looking more closely at how the diffusion process takes place over time. We do so by combining cross-country and time-series data in a statistical technique know as pooling.

We restrict the analysis to countries in the OECD. The main reason for that restriction is that those countries have a longer history of Internet use and also a better availability of historical data. Those countries happen to be also the most advanced in terms of Internet use. It is also interesting to apply the analysis to those countries because those are the ones that show higher unexplained variability in our models in Section §4.3. After presenting the model and its results, we comment on the limitations we face and possible ways to overcome them.

4.1. Model Specifications

The model we use embodies the theoretical argument that growth in Internet diffusion is a function of initial penetration, related infrastructure availability and costs. As discussed extensively in Chapter 2, those are not the only likely determinants of Internet diffusion; they are however those for which we could gather consistent time-series data.

Data

The model uses the following national-level panel data on the following variables for the period 1990-1997:

- *HOSTSPC* = Number of Internet hosts per capita
- *LINESPC* = Main Telephone Lines per capita
- *PCSPC* = Number of Personal Computers per capita
- *RESCOSTUSD* = Residential Monthly Telephone Subscription in US$
- *GDPUSDPC* = GDP per capita in US$

The data set covers these variables when available for the 29 OECD countries from 1990 to 1997 (Figure 4.17). The data has been obtained from the International Telecommunication Union's World Telecommunication Database 1997. Most variables are available for all countries and years. The dependent variable, HOSTSPC, is missing in the early 1990s in some countries. Some data for 1997 had not yet been included in the database.

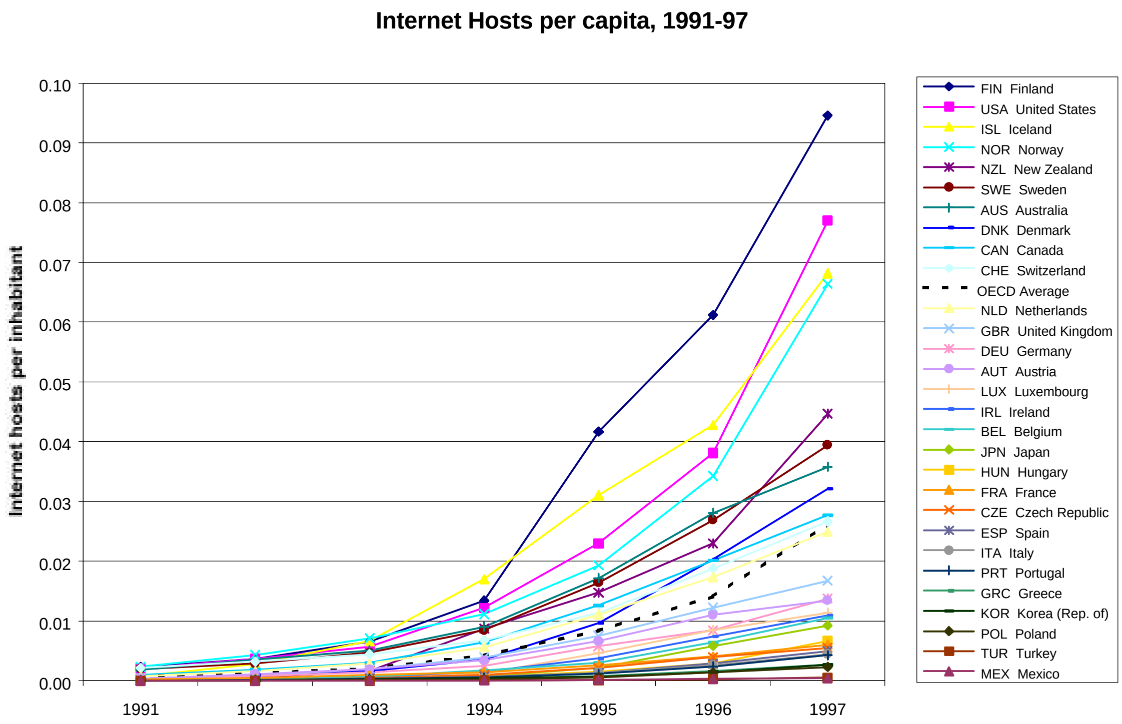
General Regression Equation

We use *i* to refer to countries and *t* to refer to years. The general form of the model employed is the following:

$$ln(HOSTSPC)_{it} = a_i + b(t\text{-}1990) + c_1 ln(HOSTSPC)_{it-1} + c_2 ln(PCSPC)_{it-1} + c_3 ln(LINESPC)_{i\,t-1} + c_4 RESCOSTUSD_{it} + c_5 GDPPCit + e_{it} \qquad [4.\ 4]$$

Where a_i is the country fixed effect, *b* is a time trend accounting for the overall diffusion phenomenon, c_1 is the growth rate with respect to the previous year, c_2 and c_3 indicate the effect of existing infrastructure, c_4 indicates the effect of cost, c_5 accounts for GDP effects and e_{it} is the error term. We employ per-capita quantities because the nature of Internet use is individual-based and there is no reason why national population should affect individual behavior: the data show that countries with large differences in population and similar other characteristics, such as the U.S. and the Nordic countries, show similar trends. The use of log-linear specifications is determined by the exponential growth process that we are examining, and for consistency we use log-linear specifications for all infrastructure penetration indicators. The data for the independent variables accounting for infrastructure (*PCSPC* and *LINESPC*) are lagged one year because it is likely that there is some delay between the infrastructure availability and consumer adoption, and also because this facilitates the use of this model as a forecasting tool. We believe the effect of price on adoption to be faster and thus we use the current value for a particular year.

We have chosen to model the year fixed effects as a time trend rather than as independent effects for each year because this allows an easier interpretation of the growth process.



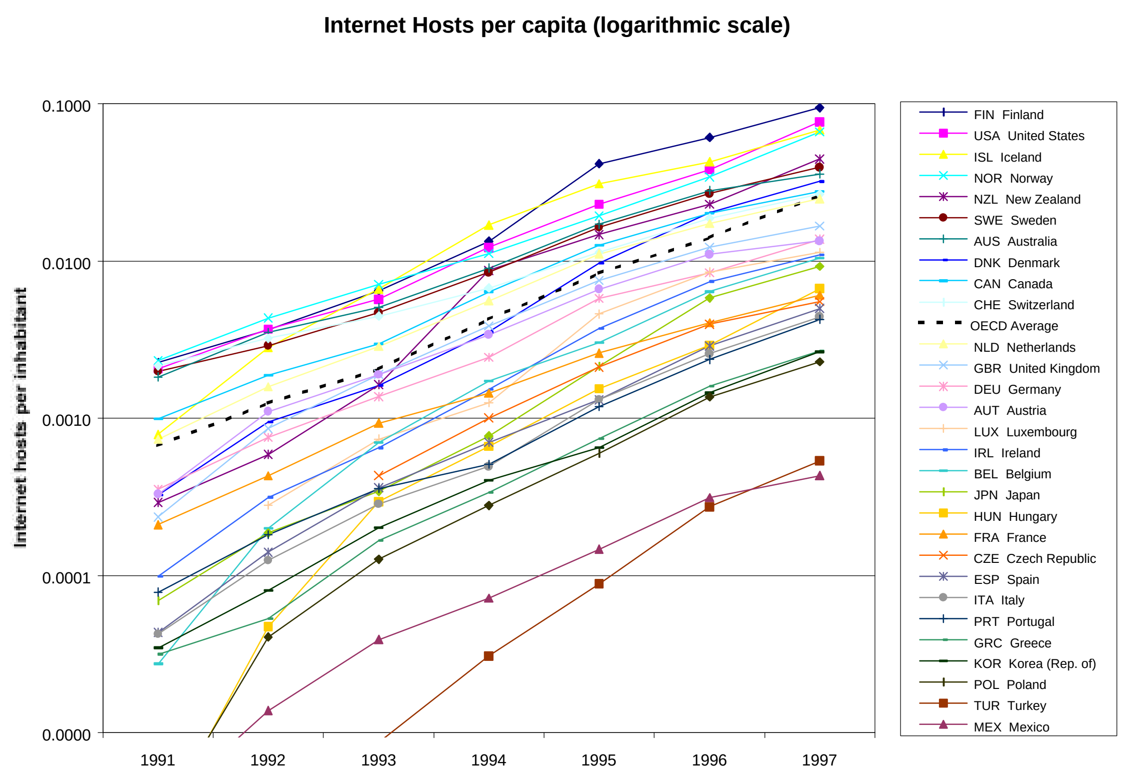


Figure 4.17: (a) and (b): Internet Hosts per capita in OECD Countries, 1991-1997

4.2. Estimation Results

We begin by estimating the overall growth trend for the complete sample of 29 countries, by regressing on (1) the 1-year lagged value of the dependent variable, on (2) a time trend starting in 1990, or on (3) a combination of both, and allowing for different intercepts for each country. Table 4.7 shows the coefficient of determination from each of those regressions.[42] The fact that all three regressions explain a large fraction of the variance indicate the high importance of the growth trend in the data.

	Model	Equation	Fraction of variance explained (R^2)
1	Lagged values only	$ln(HOSTSPC)_{it} = a_i + c_1 ln(HOSTSPC)_{it-1} + e_{it}$	.994
2	Time Trend only	$ln(HOSTSPC)_{it} = a_i + b(t-1990) + e_{it}$	.996
3	Lagged values and Time trend	$ln(HOSTSPC)_{it} = a_i) + c_1 ln(HOSTSPC)_{it-1} + b(t-1990) + e_{it}$	.998

Table 4.7: Fractions of Variance Explained

The second regression also allows us to derive a metric of the relative position of each country relative to Internet diffusion, and to translate it into a time delay. If we assume, as we will argue later, that the growth rate is the same for all countries and constant over time, this means that the U.S. is lagging approximately 4 months behind Finland (the leader in terms of Internet penetration), while for instance France and Spain are 3 and 4.5 years behind, respectively. The results from this estimation are shown in Figure 4.18.

[42] The R^2 shown in the table is the weighted (or heteroskedasticity consistent) coefficient of determination. Weighted statistics account for cross section heteroskedasticity, i.e. variability in average distribution of errors between sections (in our case, between countries).
The complete outputs from all the regressions mentioned in this section can be found in appendix E, with both weighted and unweighted statistics.

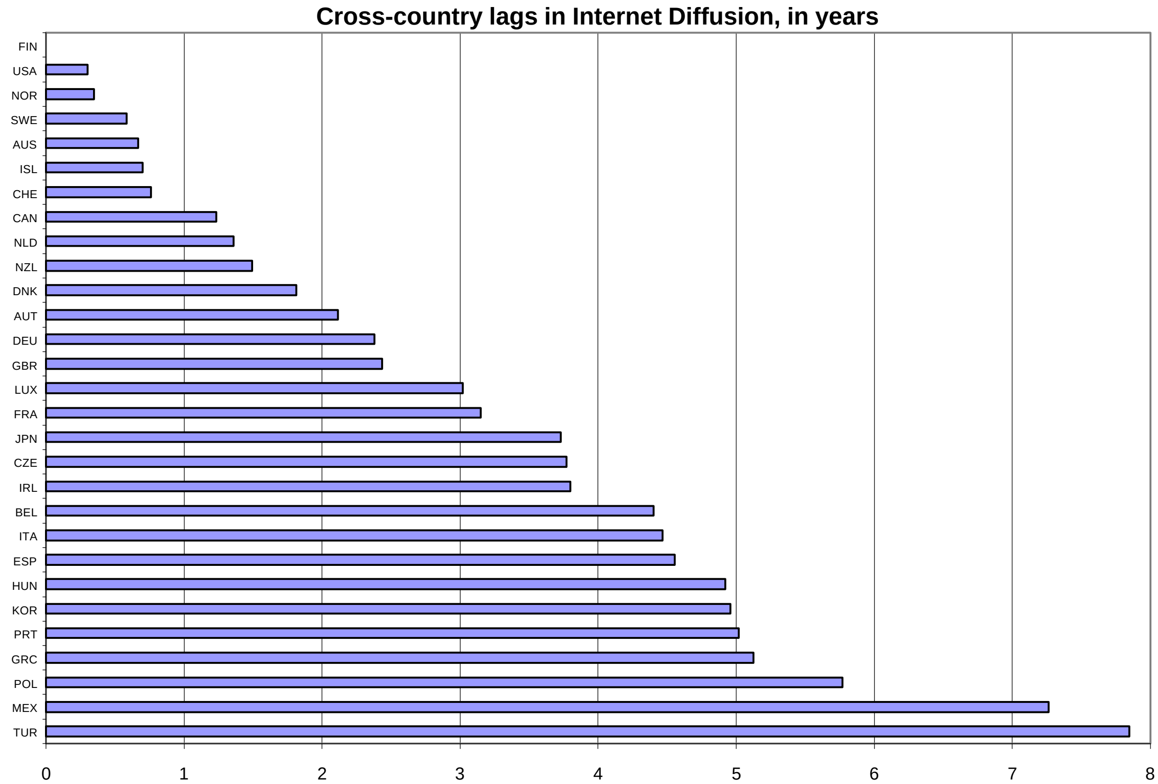


Figure 4.18: Cross-Country Differences in Internet Penetration, Measured as a Time Lag (in years)

Our next step was to find patterns for the remaining cross country and cross time variability. This proved to be a harder task as most of the variance is already explained with the general time trend. Our first attempt with our full model (with all variables from equation [4.4] included, but a common intercept) explains 99% of the variance but the coefficients for *RESCOSTUSD*, *GDPUSDPC* and *LINESPC* are not significant.[43] After a few attempts we arrive at model 5 in which 99% of the variance is explained by the time trend, and the lagged values of *LOG(HOSTSPC)* and *LOG(PCSPC)*, with a common intercept for all countries. If we allow the intercepts to vary for each country with the same regressors we get model 6, in which *LOG(PCSPC)* is no longer significant. This means that LOG(PCSPC) is a good proxy for the country fixed effect.

[43] For details, refer to Appendix E, Model 4.

	Model	Equation	Fraction of variance explained (R^2)
4	All variables w/ common intercept	$ln(HOSTSPC)_{it} = a + b(t\text{-}1990) + c_1 ln(HOSTSPC)_{it-1} + c_2 ln(PCSPC)_{it-1} + c_3 ln(LINESPC)_{i\,t-1} + c_4 RESCOSTUSD_{it} + c_5 GDPPCit + e_{it}$	.991
5	Lagged values + Time trend + log(PCs) w/ common intercept	$ln(HOSTSPC)_{it} = a + b(t\text{-}1990) + c_1 ln(HOSTSPC)_{it-1} + c_2 ln(PCSPC)_{it-1} + e_{it}$	.992
6	Lagged values + Time trend +log(PCs) w/ fixed effects	$ln(HOSTSPC)_{it} = a_i + b(t\text{-}1990) + c_1 ln(HOSTSPC)_{it-1} + c_2 ln(PCSPC)_{it-1} + e_{it}$	.998

Table 4.8: Fractions of Variance Explained

To summarize, the two models that best help us explaining and forecasting Internet diffusion are:

Model 5) (common intercept for all countries)

$$\ln(\text{HOSTSPC})_{it} = -.992 + .085(t\text{-}1990) + .620\ln(\text{HOSTSPC})_{it-1} + .508\ln(\text{PCSPC})_{it-1}$$

Model 6) (fixed effects - different a_i coefficients for each country)

$$\ln(\text{HOSTSPC})_{it} = a_i + .360(t\text{-}1990) + .354\ln(\text{HOSTSPC})_{it-1} + .137\ln(\text{PCSPC})_{it-1}$$

In plain words, the simplest way of explaining (and forecasting) a country’s level of Internet penetration as measured by Internet hosts per capita for a particular year is by looking just at the level of diffusion of the previous period and applying a growth rate that is common across country. A slightly better fit is achieved by *i)* taking into account also the penetration of Personal Computers in the previous period, and *ii)* accounting for different intercepts for each country. Both improvements are statistically significant, but add little to the explanation in terms of how much additional variability they explain.

4.3. Key Findings

The key findings from our analysis can be summarized as follows:

Growth Trend

The process of Internet diffusion in OECD countries between 1990 and 1997 is a very regular and steady exponential growth, at a rate that is almost constant (in logarithms) along time and across countries.

Network Effect

The growth process can be decomposed into two different drivers: one, represented by the lagged value of *ln*(*HOSTSPC*), accounts for the network effect (the more penetration in period *t-1*, the more growth in period *t*); two, a general time trend. The two effects explain roughly half of the overall trend each.

Few Cross-Country Differences

The countries we have studied differ in orders of magnitude in terms of the range of the dependent variable (hosts per capita) over time. However, we observe very few differences in the growth process itself, given the initial differences. This lower than expected variability made it difficult to study the effect of other variables that we presumed important.

Almost no "Catch up"

A consequence from the above is that there seems to be in practice little opportunity for "catch up" for countries that are lagging behind and would like to foster Internet penetration. In other words there seems to be little that they can do to decrease the time gap represented in figure 4.17.

Personal Computer Penetration is a Key Lever

The only clear exception to the former is PC penetration that clearly has a positive impact on penetration rates.

Effect of Telecommunication Cost is Found not Significant

The cost of local communications chosen does not prove to be a good proxy for capturing the effect of access cost, although we know that it is an important component of Internet access. A better metric of access costs to the end user is needed before we can conclude that Internet demand is insensitive to access price. An approach based on the cost of "baskets", as exposed in Section §2.3.3, can be a solution but those baskets have not been consistently defined nor recorded in the past.

4.4. Model Limitations

We devote the following paragraphs to discuss some of the issues we have faced during the development of the study and also how we think it could be expanded and/or refined.

Data

Probably the most challenging task in performing a quantitative assessment of Internet diffusion is linked to the difficulty in gathering comprehensive and coherent indicators, especially on the dependent variable side. The variable chosen, Internet hosts per capita, has several drawbacks that limit its accuracy, as was discussed in Section §3.1.2.

Other issues arise when choosing the appropriate data for the independent variables. The infrastructure indicators are basically reliable: the number of mainlines has been a basic telecommunication indicator for years and the number of PCs can also be obtained from yearly sales data, although the number of PCs in operation at a given time might be more difficult to estimate because of obsolescence, replacement, etc.

A more difficult issue rises with respect to price. It seems that the price of Internet access should play a role in Internet diffusion, as we showed in §4.2.3 for our sample of 59 major economies. However, such impact could not be derived from the actual data used for OECD countries. There are at least two possible reasons for such result: the first is related to our comment on the dependent variable: we are measuring the number of hosts, instead of the usage of the Internet, and the former might be less sensitive in the short run than the latter to shifts in price. And second, it is very difficult to define a measurement of the price of Internet access that would be consistent across countries. The costs of Internet access for a final user are usually a combination of equipment, telecommunications and ISPs (Internet Service Provider) costs. In the absence of an Internet price basket measured in each country over the period 1990-97, we chose instead to study the effect of costs that should be correlated with the price of Internet access, in particular the cost of local calls and the price of residential telephone access, but none proved significant in explaining cross country differences.

Model

The cross-section time series model used for the analysis was appropriate for the study and allowed us to use powerful statistical tools to gain insights on the data. However it seemed to be difficult to really observe many cross-section differences, probably due to a lack of heterogeneity in the sample: the OECD countries chosen have a lot of commonalties and the study could have been refined by including data from other countries such as developing countries, but unfortunately there is almost no data of that kind available outside the OECD.

We could also observe two particular methodological issues that should deserve further attention in an extension of this study.

i) Heteroskedasticity

The data for the dependent and independent variables are heteroskedastic both across countries and along time, with the data of the last years and the countries with higher Internet penetration tending to have lower variance than the ones from earlier years and/or countries with lower penetration. For this reason we used Generalized Least Squares (GLS) with cross section weights and White heteroskedasticity consistent covariance.[44]

[44] One of the basic assumptions of the linear regression model is that the error term has zero expected value and constant variance for all observations. An error term that has constant variance is called homosketdastic. If the variance of the error term changes, than the error is heteroskedastic and correction techniques have to be used. In the case of cross-sectional models, it is often the case that the variance of the error terms is different between sections. See [Pindyck91] for a discussion on heteroskedasticity and related correction techniques.

ii) Multicollinearity

As in our previous analysis in Section §4.3.2, some of our dependent variables have relatively high degrees of correlation, which has the effect of lowering the significance level of the coefficients of the regressions. There is similarly a high correlation between the time trend and the series of lagged values of our dependent variable.

Chapter Five

Policy Recommendations and Conclusions

Results from this thesis include findings on: the commonalties and differences of Internet diffusion worldwide; the relationship between Internet diffusion and some key determinants; and the complexity and limited availability of metrics for its measurement. From those findings we derive recommendations for improved data collection, for policy aimed at pushing Internet diffusion, and for further research.

1. Main Results

1.1. Commonalties and Differences of Internet Diffusion Worldwide

The Internet is growing at exponential rates in almost all countries of the world. However, today's distribution of Internet hosts and users is quite heterogeneous: most of the Internet is concentrated in high-income countries. Moreover, the distribution of Internet hosts is much more skewed than the distribution of national income or even other related technologies, such as personal computers or telephone lines (see figure §4.7). Some of the differences can be explained by historical circumstances. That is, the Internet remains concentrated in the countries where it was first developed (e.g., the U.S.) and it is still too early in the evolution of the Internet for the effect of these initial conditions to have been overcome. We observe, however, that current growth rates are not closing the gap: that is, unless the trend is reversed, today's big gap between Internet-wealthy and Internet-poor nations will persist for the foreseeable future.

1.2. Relationship between Internet Diffusion and some Key Determinants

Even within nations with similar socioeconomic and cultural conditions, the observed level of growth varies widely. We found that some of those variations can be explained by factors such as the availability of Internet-friendly infrastructure or by elements of costs, which indirectly account for market structure.[45] However we also saw that as we look at countries with higher penetration levels the relationship between diffusion and those explanatory factors becomes weaker. In particular, when our analysis of the growth of the Internet is restricted to OECD countries between 1990 and 1997 (Section §4.4), we could explain variations in Internet growth by the combined effect of a growth trend and the availability of personal computers, but were unsuccessful in finding a statistically significant impact due to our measure of access costs. This could be due either to missing variables (e.g., we have neglected to account for important cultural determinants such as availability of local content or other factors that drive Internet diffusion) or to measurement problems (e.g., measuring diffusion in terms of hosts per capita provides only an indirect measure of the true size of the Internet in a country, as we discuss in Section §2.2). Below, we provide suggestions for future data collection and analysis that should improve our understanding of the diffusion process.

1.3. Complexity and Limited Availability of Metrics

As described in detail in Chapter 3, probably the most challenging task in performing a quantitative assessment of Internet diffusion is linked to the difficulty of gathering comprehensive and coherent metrics. We identified several reasons for this difficulty. First, metrics of Internet infrastructure and usage are often not defined clearly or consistently, and even when this is not the case, they are often not applied correctly. Second, there are no generally accepted quantitative definitions and metrics of Internet usage. Third, the

[45] Specifically, we examined the impact of Internet access pricing (including relevant telephone charges for dial-up access) on Internet diffusion and the impact of other proxies for Internet costs such as leased line rates. Because leased line rates are in some cases set in competitive markets and in other cases by regulatory tariff, these rates provide an indirect measure of industry structure. Typically, leased line prices are higher in markets with less competition and more direct regulatory oversight.

existing metrics are often related to a rapidly changing technology and therefore may become inconsistent over time. Fourth, because of the self-governance of the Internet there are almost no “official” data on the industry. And fifth, the novelty of the Internet limits the availability of historical data to a few years, especially in some countries.

For the above reasons, almost all available studies on Internet diffusion refer to a single data source, which is the survey of Internet hosts undertaken by Network Wizards and the RIPE. This survey uses the Internet Protocol itself to count the number of computers connected to the Internet with a domain name associated to it. While this is the most widely used and appreciated survey available, it has several important limitations (see Section §3.2). The most important one is that the measured host data do not indicate the total number of *users* who can access the Internet. In fact, a growing number of users located behind companies’ firewalls or users that are assigned dynamic IP addresses by their ISPs aren’t properly accounted for (although the survey methodology was recently changed to address this limitation). Furthermore, host counts do not provide any information on network usage.

Limitations also apply to data concerning the determinants of diffusion. For example, because the Internet is not regulated in most countries, there is limited data about ISP subscribers, users and usage.

2. Recommendations for Data Collection

The Internet is likely to be an important component of the global information infrastructure, and hence, national authorities would benefit from a better understanding of Internet diffusion patterns. To facilitate this, national authorities ought to encourage the collection of better data. Such data would lead to a better understanding of Internet diffusion and its determinants, and therefore enable researchers, policy makers and regulators to better focus policies aimed at increasing the benefits of Internet use around the world. Such data gathering should be consistent with the corporate non-disclosure rights of ISPs and other Internet actors. A good framework that could be used as a guideline for which indicators to track was suggested by the Committee on Indicators of

Internet Impacts on Development of the U.S. National Research Council and is presented in Appendix F. The main categories of relevant indicators that apply to the Internet are:

- **Indicators Related to the Environment for Internet Use:**
  - Supportive Economy and Infrastructure
  - Policy and regulatory Environment
- **Indicators of Internet Supply**
  - Quantity of Internet Service
  - Sustainability
- **Indicators of Internet Use**
  - Costs of Internet Use
- **Indicators of Impact on Formal Organizations**
  - Perceived Benefits of the Internet
  - Organizational Decision Making
- **Indicators of Sectoral Impacts**
  - Sectoral Use and Diffusion of the Internet
- **Internet Impacts on Sectors and their related development goals**
  - Education
  - Private Sector
  - Government and Civil Society

We would like to encourage national agencies and international organizations to increase their gathering of systematic data in those dimensions, in order to allow researchers to better comprehend the phenomenon of Internet diffusion.

3. Recommendations for Policy

There is still a long way to go for the Internet to make the world become the "global village" anticipated by Marshal McLuhan in 1965. Today, the Internet is much less equally distributed than most other communication technologies, including the telephone, TV or personal computers. Moreover, current growth rates do not indicate that less developed countries are narrowing the gap. This thesis shows that some policies aimed at increasing the availability of supporting infrastructure, as well as lowering the costs of access and transport, can have an impact on the pace of diffusion. However, it also shows that much more data collection and research is needed before the accuracy of those and other policies can be fully understood. Table 5.1 provides a summary of the policy issues linked to Internet diffusion.

1) Internet diffusion is important for the growth of national economies
2) Internet diffusion is influenced by a variety of factors, in a way that is often complex to understand, and calls for improved data collection and socioeconomic analysis.
3) However, there are factors that do encourage or limit diffusion, upon which governments and regulators can have an impact: ➢ Availability of supportive infrastructure (telephone lines, PCs, connectivity to the international backbone, development of a national backbone) ➢ Initial diffusion (via “seeding” in universities, school, government centers) ➢ Competitive Internet service provision (translates in lower access costs) ➢ Competitive access and transport networks (translates in lower transport costs for providers)
4) The right policy depends on the national context and development level.

Table 5.1: Internet Diffusion: Policy implications

4. Areas for Further Research

Three main areas have been identified in this thesis in which we believe a large amount of research and clarification is still needed:

1) definition, standardization and collection of metrics that would account better for the phenomenon of Internet diffusion and its impact in national economies and societies, as well as for the determinants of diffusion (demand and supply and costs of Internet-related goods and services).

2) study of the relationship between additional endogenous and exogenous variables in Internet diffusion. In particular, develop compound metrics to account for the multiple dimensions of Internet impact in a country’s economy and society.

3) distinguish different phases in diffusion and sketch which policies might be more effective depending on the phase. Table 5.2 provides an example of what might be the possible outcome of such an analysis.

	Country's Level of Internet Diffusion			
	Experimental	Established	Common	Pervasive
International connectivity	+++			
Wealth/Income	++	++	+	
Basic Infrastructure	+	++	++	+
Language/ English	+	++	++	+
Competitive ISP market		+	++	++
Competitive data networks		+	++	+
Metered local calls		+	++	+++
ISP prices		+	++	+++

Table 5.2: Key Diffusion Factors by Phases of Internet Diffusion (Estimate)
The four levels of diffusion correspond to the classification
of the MOSAIC group (see Table 3.6)

5. Final Remarks

We believe that the Internet is more than a mere technological innovation. Rather, it is a powerful tool that is able to empower humans all over the world who can use it to learn, trade, communicate and share information in new, easy and flexible ways. It is important to understand how the use of this tool is growing and being spread over the planet, in order to know how different policies might be used to encourage its use or to prevent the rise of new inequalities between the information rich and poor. We hope this thesis will have provided some light on the topic, showing what assertions on Internet diffusion can be made as of today, and that it will encourage others to carry on the additional data gathering and analysis needed to improve our understanding of Internet diffusion.

Appendix A

References

[Cawley96] Cawley, Richard A. Policies to support the scaling and extensibility of Internet: Principle to guide settlements policy consistent with technical and pricing developments of Internet, 1996. http://ksgwww.harvard.edu/iip/cai/cawley.html.

[DOC98] *The Emerging Digital Economy*, U.S. Department of Commerce, 1998. http://www.ecommerce.gov.

[Downes99] Downes, Thomas A. and Greenstein, Shane M. *Do Commercial ISPs Provide Universal Access?* included in *Competition, Regulation and Convergence: Selected Papers from the 1998 Telecommunications Policy Research Conference* (Gillett and Vogelsang, editors). Lawrence Erlbaum Assoc., 1999.

[DTI98] Set of three reports on *Benchmarking the UK in the Information Age*, U.K Department of Trade and Industry ,1996, 1997 and1998. http://www.isi.gov.uk/isi/bench/index.html.

[Elie98] Elie, Michel. *The Internet and Global Development*, Internet Society, 1998. http://www.isoc.org/inet98/proceedings/5d/5d_3.htm.

[Gromov99] Gromov, Gregory R. *The Roads and Crossroads of Internet History*, 1999. http:// www.netvalley.com/netvalley/intvalstat.html.

[Hargittai98] Hargittai, Eszter *Holes in the Net: The Internet and International Stratification*, Internet Society, 1998. http://www.isoc.org/inet98/proceedings/5d/5d_1.htm.

[Henisz98] Henisz, Witold J. and Zelner, Bennet A. *The Institutional Investment for Telecommunications Investment*, proceedings of the Telecommunication Policy Research Conference 1998.

[ISPA98] *Good Intentions: the Effects of Telecoms Pricing Policies on the European Internet*, EuroISPA, 1998. http://www.euroispa.org/papers/telecoms1.html.

[ITU97] *Challenges to the network: Telecoms and the Internet*, ITU, 1997.

[ITU99] *Challenges to the network: Internet for Development*, ITU, 1999.

[Kennedy92] Kennedy, Peter. *A guide to Econometrics*, MIT Press, 1992.

[Laursen97] Laursen, J. V. *The Internet: Past, Present and Future – Internet and WWW History*, updated 1997. http://www.vissing.dk/inthist.html.

[Leida98] Leida, Brett. *A Cost Model of Internet Service Providers: Implications for Internet Telephony and Yield Management*, MIT S.M. Technology and Policy Thesis, 1998.

[Leiner97] Leiner, Barry M. et al. *A Brief History of the Internet, version 3.1*, 1997. http:// www.k12.hi.us/~telecom/inethistoru.html.

[McKnight97] McKnight, Lee W. and Bailey, Joseph P. (eds.). *Internet Economics*, MIT Press, 1997.

[NAS98] National Academy of Sciences. *Internet Counts: Measuring the Impacts of the Internet*, National Academy Press, 1998. http:// www.bsos.umd.edu/cidcm/wilson/xnasrep2.htm

[OECD96] OECD. *Information Infrastructure Convergence and Pricing: The Internet*, OECD, 1996.

[OECD98] OECD. *Internet Infrastructure Indicators*, OECD, 1998.

[Petrazzini98] Petrazzini, Ben A. *Explaining Variations Among Countries in Internet Development*, Internet Society, 1998.

[Press97] Press, Larry. *Tracking the Global Diffusion of the Internet*, Communications of the ACM, Vol. 40, No. 11, November 1997. http://som.csudh.edu/cis/lpress/articles/worldacm.htm.

[Press98] Press, Larry. *An Internet Diffusion Framework*, Communications of the ACM, Vol. 41, No. 10, October 1998.

[Pindyck91] Pindyck, Robert S. and Rubinfeld, Daniel L. *Econometric Models & Economic Forecasts*, McGraw Hill, 1991.

[Shuster98] Shuster, Marc S. Diffusion of Network Innovation*: Implications for Adoption of Internet Services*. MIT S.M. EECS Thesis, 1998.

[Tanenbaum96] Tanenbaum, Andrew S. *Computer Networks*, Prentice Hall, 1996.

Data Sources

[#CIA] Computer Industry Almanac *(Various)*.
http:// www.c-i-a.com.

[#EUROMK] Euro-Marketing, *Global Internet Statistics*.
http://www.euromktg.com/globstats.

[#INTUG] International Telecommunication Users Group *Annual International Tariff Surveys*.
http://www.intug.net.

[#ITU] ITU *World Telecommunications Indicator Database* 1997.

[#ISOC] Internet Society. *Web Languages Hit Parade*, June 1996.
http://www.isoc.org:8080/palmares.html.

[#NW] Network Wizards *Internet Domain Survey*, 1995-99.
http://www.nw.com.

[#OECD] OECD Communications Outlook 1997.

[#PNE] Public Network Europe, March to September 1998.

Appendix B

List of Acronyms

ADSL	Asymmetric Digital Subscriber Line
CAGR	Compound Annual Growth Rate
DNS	Domain Name Service
DHCP	Dynamic Host Configuration Protocol
GLS	Generalized Least Squares
IAP	Internet Access Provider
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISP	Internet Service Provider
LEC	Local Exchange Carrier
NAP	Network Access Point
OLS	Ordinary Least Squares
PC	Personal Computer
PDA	Personal Digital Assistant
PNO	Public Network Operator
POP	Point-of-Presence
PSTN	Public Switched Telephone Network
TCP	Transmission Control Protocol
TV	Television

Appendix C

Country Acronyms

ARG	Argentina
AUS	Australia
AUT	Austria
BEL	Belgium
BRA	Brazil
CAN	Canada
CHE	Switzerland
CHL	Chile
CHN	China
CZE	Czech Republic
DEU	Germany
DNK	Denmark
ESP	Spain
FIN	Finland
FRA	France
GBR	UK
GRC	Greece
HKG	Hongkong,
HUN	Hungary
IDN	Indonesia
IND	India
IRL	Ireland
ISL	Iceland
ISR	Israel
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
MEX	Mexico
MYS	Malaysia
NLD	Netherlands
NOR	Norway
NZL	New Zealand
PHL	Philippines
POL	Poland
PRT	Portugal
RUS	Russia
SGP	Singapore
SWE	Sweden
THA	Thailand
TUR	Turkey
TWN	Taiwan-China
USA	USA
VEN	Venezuela
ZAF	South Africa

Appendix D

Data Tables

Internet Access Costs and Leased Lines Costs, 1998.

Sources: [ITU99], p. A-29 (access costs), [#INTUG] and [#PNE] for leased line costs.

country	country code	flat rate	connection charge	ISP charge US$	Call Charge US$	Total US$	Tot as% of GDPpc	ln Tot as%	Total w/ line rental US$	2mbps monthly leased line cost, US$'000
Argentina	ARG		*	30	11	41	5%	-1.26	54	N/A
Australia	AUS			22	6	29	2%	-1.79	36	6.7
Austria	AUT	*		23	29	52	2%	-1.61	65	5.0
Belgium	BEL	*	*	23	20	43	2%	-1.67	61	10.4
Brazil	BRA		*	33	3	36	9%	-1.06	50	N/A
Canada	CAN		*	12	0	12	1%	-2.14	27	7.7
Chile	CHL	*		32	7	39	9%	-1.05	53	N/A
China	CHN			39	26	65	108%	0.03		N/A
Czech Republic	CZE	*		25	15	40	9%	-1.02	43	9.8
Denmark	DNK	*		23	27	50	2%	-1.72	63	3.6
Finland	FIN	*	*	9	18	27	1%	-1.86	45	1.6
France	FRA	*		17	26	42	2%	-1.67	54	8.9
Germany	DEU	*		23	28	51	2%	-1.62	65	6.6
Greece	GRC	*		21	7	29	3%	-1.52	36	7.3
Hongkong	HKG	*	*	18	5	23	1%	-1.98	32	N/A
Hungary	HUN			24	20	44	12%	-0.92	49	12.5
Iceland	ISL			22	28	50	2%	-1.66	56	5.8
India	IND			13	0	13	40%	-0.40	17	N/A
Indonesia	IDN			9	6	14	16%	-0.79	17	N/A
Ireland	IRL	*	*	21	14	35	2%	-1.68	53	7.0
Israel	ISR			25	25	50	4%	-1.44	56	N/A
Italy	ITA			23	14	37	2%	-1.65	48	13.8
Japan	JPN		*	41	14	56	2%	-1.70	70	10.8
Korea	KOR		*	12	14	26	3%	-1.48	28	N/A
Luxembourg	LUX			22	56	78	2%	-1.65	94	N/A
Malaysia	MYS	*	*	1	8	8	2%	-1.65	14	N/A

country	country code	flat rate	connection charge	ISP charge US$	Call Charge US$	Total US$	Tot as% of GDPpc	ln Tot as%	Total w/ line rental US$	2mbps monthly leased line cost, US$'000
Mexico	MEX	*		24	3	27	8%	-1.11	40	N/A
Netherlands	NLD			23	19	42	2%	-1.66	60	5.2
New Zealand	NZL		*	19	0	19	1%	-1.88	38	15.3
Norway	NOR	*		13	22	35	1%	-1.92	47	4.9
Philippines	PHL			31	0	31	33%	-0.48	41	N/A
Poland	POL	*	*	14	20	34	12%	-0.93	37	N/A
Portugal	PRT			19	10	29	3%	-1.49	42	9.4
Russia	RUS			20	0	20	8%	-1.10	23	N/A
Singapore	SGP			15	5	21	1%	-2.02	26	#N/A
South Africa	ZAF	*		17	8	26	10%	-0.99	35	#N/A
Spain	ESP			23	16	38	3%	-1.47	50	12.5
Sweden	SWE			22	17	39	2%	-1.74	57	3.1
Switzerland	CHE		*	18	14	32	1%	-1.97	50	7.7
Taiwan-China	TWN			21	1	22	3%	-1.60	24	#N/A
Thailand	THA		*	25	33	58	27%	-0.56	61	#N/A
Turkey	TUR			25	9	34	14%	-0.86	36	2.0
UK	GBR	*		20	29	49	3%	-1.57	60	5.2
USA	USA			20	0	20	1%	-2.10	38	2.5
Venezuela	VEN			28	39	67	21%	-0.67	74	#N/A

Appendix E

Regression Outputs

Model 1

```
============================================================
Dependent Variable: LOG(HOSTSPC_?)
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:56
Sample: 1991 1997
Included observations: 7
Total panel (unbalanced) observations 185
============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
============================================================
LOG(HOSTSPC_?(-1))   0.761025   0.012119   62.79638   0.0000
   Fixed Effects
      AUS--C        -0.682662
      AUT--C        -0.888148
      BEL--C        -0.459509
      CAN--C        -0.733169
      CZE--C        -0.936848
      DNK--C        -0.768948
      FIN--C        -0.473526
      FRA--C        -1.100564
      DEU--C        -0.901667
      GRC--C        -1.331950
      HUN--C        -0.696474
      ISL--C        -0.336436
      IRL--C        -0.577888
      ITA--C        -1.193539
      JPN--C        -0.960296
      KOR--C        -1.264581
      LUX--C        -0.799826
      MEX--C        -1.590190
      NLD--C        -0.802494
      NZL--C        -0.405006
      NOR--C        -0.509469
      POL--C        -1.032065
      PRT--C        -0.687616
      ESP--C        -0.984882
      SWE--C        -0.721165
      CHE--C        -0.820599
      TUR--C        -1.334639
      GBR--C        -0.653429
      USA--C        -0.527835
============================================================
Weighted Statistics
============================================================
R-squared            0.994669    Mean dependent var-8.774344
Adjusted R-squared   0.993672    S.D. dependent var 5.064004
S.E. of regression   0.402848    Sum squared resid  25.15443
Durbin-Watson stat   1.401957
============================================================
Unweighted Statistics
============================================================
R-squared            0.957230    Mean dependent var-6.125491
```

```
Adjusted R-squared   0.949228    S.D. dependent var 1.922283
S.E. of regression   0.433142    Sum squared resid  29.07988
Durbin-Watson stat   1.408250
============================================================
```

Model 2

```
============================================================
Dependent Variable: LNHOSTSPC_?
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:58
Sample: 1990 1997
Included observations: 8
Total panel (unbalanced) observations 214
============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
============================================================
   @TREND(1990)      0.763180   0.009966   76.57471   0.0000
   Fixed Effects
      AUS--C        -7.754643
      AUT--C        -8.859224
      BEL--C        -10.60579
      CAN--C        -8.185753
      CZE--C        -10.12350
      DNK--C        -8.628163
      FIN--C        -7.246089
      FRA--C        -9.647531
      DEU--C        -9.060714
      GRC--C        -11.15613
      HUN--C        -11.00357
      ISL--C        -7.780034
      IRL--C        -10.14639
      ITA--C        -10.65570
      JPN--C        -10.09298
      KOR--C        -11.02849
      LUX--C        -9.550637
      MEX--C        -12.78936
      NLD--C        -8.280126
      NZL--C        -8.383112
      NOR--C        -7.512194
      POL--C        -11.64924
      PRT--C        -11.07488
      ESP--C        -10.72267
      SWE--C        -7.692151
      CHE--C        -7.824678
      TUR--C        -13.23582
      GBR--C        -9.102278
      USA--C        -7.475285
============================================================
Weighted Statistics
============================================================
R-squared            0.996200    Mean dependent var-13.44772
Adjusted R-squared   0.995601    S.D. dependent var 11.36670
S.E. of regression   0.753910    Sum squared resid  104.5819
Durbin-Watson stat   1.044412
============================================================
Unweighted Statistics
============================================================
R-squared            0.909158    Mean dependent var-6.629047
Adjusted R-squared   0.894841    S.D. dependent var 2.400164
S.E. of regression   0.778331    Sum squared resid  111.4670
Durbin-Watson stat   0.937916
============================================================
```

Model 3

```
============================================================
Dependent Variable: LOG(HOSTSPC_?)
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:51
Sample: 1991 1997
Included observations: 7
Total panel (unbalanced) observations 185
============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
============================================================
   @TREND(1990)      0.375283   0.016379   22.91293   0.0000
LOG(HOSTSPC_?(-1))   0.360707   0.018042   19.99251   0.0000
   Fixed Effects
      AUS--C        -4.345243
      AUT--C        -4.973920
      BEL--C        -5.330140
      CAN--C        -4.580003
      CZE--C        -5.636885
      DNK--C        -4.798811
      FIN--C        -3.932867
      FRA--C        -5.501534
      DEU--C        -5.081019
      GRC--C        -6.376223
      HUN--C        -5.764350
      ISL--C        -4.021335
      IRL--C        -5.240737
      ITA--C        -6.037626
      JPN--C        -5.624736
      KOR--C        -6.282498
      LUX--C        -5.255768
      MEX--C        -7.309486
      NLD--C        -4.658634
      NZL--C        -4.409875
      NOR--C        -4.070313
      POL--C        -6.330097
      PRT--C        -5.721467
      ESP--C        -5.866581
      SWE--C        -4.334505
      CHE--C        -4.472249
      TUR--C        -7.358745
      GBR--C        -4.862934
      USA--C        -4.111036
============================================================
Weighted Statistics
============================================================
R-squared            0.997850    Mean dependent var-8.080509
Adjusted R-squared   0.997431    S.D. dependent var 4.415086
S.E. of regression   0.223791    Sum squared resid  7.712689
F-statistic          71462.15    Durbin-Watson stat 1.701909
Prob(F-statistic)    0.000000
============================================================
Unweighted Statistics
============================================================
R-squared            0.987387    Mean dependent var-6.125491
Adjusted R-squared   0.984930    S.D. dependent var 1.922283
S.E. of regression   0.235979    Sum squared resid  8.575639
Durbin-Watson stat   1.510709
============================================================
```

Model 4

```
============================================================
Dependent Variable: LOG(HOSTSPC_?)
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:23
Sample: 1991 1997
Included observations: 7
Total panel (unbalanced) observations 152
Cross sections without valid observations dropped
============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
============================================================
         C          -1.210073   0.195787  -6.180558   0.0000
   @TREND(1990)      0.100228   0.017957   5.581630   0.0000
LOG(HOSTSPC_?(-1))   0.582987   0.024484   23.81111   0.0000
 LOG(PCSPC_?(-1))    0.620221   0.066127   9.379267   0.0000
LOG(LINESPC_?(-1))  -0.208837   0.115944  -1.801188   0.0738
   RESCOSTUSD_?     -0.009542   0.008298  -1.149814   0.2521
    GDPUSDPC_?       5.90E-06   5.12E-06   1.152425   0.2510
============================================================
Weighted Statistics
============================================================
R-squared            0.991996    Mean dependent var-9.392223
Adjusted R-squared   0.991665    S.D. dependent var 4.576492
S.E. of regression   0.417826    Sum squared resid  25.31385
F-statistic          2995.096    Durbin-Watson stat 0.970021
Prob(F-statistic)    0.000000
============================================================
Unweighted Statistics
============================================================
R-squared            0.950112    Mean dependent var-6.360691
Adjusted R-squared   0.948048    S.D. dependent var 1.892605
S.E. of regression   0.431383    Sum squared resid  26.98318
Durbin-Watson stat   0.886188
============================================================
```

Model 5

```
============================================================
Dependent Variable: LOG(HOSTSPC_?)
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:37
Sample: 1991 1997
Included observations: 7
Total panel (unbalanced) observations 178
Cross sections without valid observations dropped
============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
============================================================
         C          -0.992147   0.121783  -8.146825   0.0000
   @TREND(1990)      0.084559   0.013315   6.350511   0.0000
LOG(HOSTSPC_?(-1))   0.619826   0.023121   26.80845   0.0000
 LOG(PCSPC_?(-1))    0.507651   0.050130   10.12676   0.0000
============================================================
Weighted Statistics
============================================================
R-squared            0.992535    Mean dependent var-9.147067
Adjusted R-squared   0.992406    S.D. dependent var 4.720263
S.E. of regression   0.411331    Sum squared resid  29.43963
F-statistic          7711.649    Durbin-Watson stat 0.992689
Prob(F-statistic)    0.000000
============================================================
Unweighted Statistics
============================================================
R-squared            0.953020    Mean dependent var-6.157031
Adjusted R-squared   0.952210    S.D. dependent var 1.932908
S.E. of regression   0.422553    Sum squared resid  31.06783
Durbin-Watson stat   0.875668
============================================================
```

Model 6

```
==============================================================
Dependent Variable: LOG(HOSTSPC_?)
Method: GLS (Cross Section Weights)
Date: 05/04/99   Time: 11:40
Sample: 1991 1997
Included observations: 7
Total panel (unbalanced) observations 178
Cross sections without valid observations dropped
==============================================================
     Variable      CoefficientStd. Errort-Statistic  Prob.
==============================================================
   @TREND(1990)      0.359580   0.020535   17.51021   0.0000
LOG(HOSTSPC_?(-1))   0.354185   0.018890   18.75024   0.0000
 LOG(PCSPC_?(-1))    0.136849   0.112496   1.216483   0.2258
   Fixed Effects
      AUS--C        -4.102368
      AUT--C        -4.644578
      BEL--C        -5.048851
      CAN--C        -4.308754
      CZE--C        -5.172739
      DNK--C        -4.544597
      FIN--C        -3.646146
      FRA--C        -5.170724
      DEU--C        -4.788421
      GRC--C        -5.867544
      HUN--C        -5.252565
      ISL--C        -3.673978
      IRL--C        -4.949785

      ITA--C        -5.646190
      JPN--C        -5.276853
      KOR--C        -5.915943
      MEX--C        -6.761721
      NLD--C        -4.372823
      NZL--C        -4.127854
      NOR--C        -3.889607
      POL--C        -5.774085
      PRT--C        -5.278321
      ESP--C        -5.469057
      SWE--C        -4.061612
      CHE--C        -4.201889
      TUR--C        -6.751573
      GBR--C        -4.580523
      USA--C        -3.900848
==============================================================
Weighted Statistics
==============================================================
R-squared            0.998238    Mean dependent var-8.424153
Adjusted R-squared   0.997879    S.D. dependent var 4.786142
S.E. of regression   0.220436    Sum squared resid  7.143053
F-statistic          41646.84    Durbin-Watson stat 1.717844
Prob(F-statistic)    0.000000
==============================================================
Unweighted Statistics
==============================================================
R-squared            0.988028    Mean dependent var-6.157031
Adjusted R-squared   0.985585    S.D. dependent var 1.932908
S.E. of regression   0.232070    Sum squared resid  7.916927
Durbin-Watson stat   1.483334
==============================================================
```

Appendix F

Indicators of Internet Impacts

The following list of indicators is extracted from the report *Internet Counts: Measuring the Impacts of the Internet*, published by the U.S. National Academy Press in 1998 [NAS98]. The report is the result of a project led by the U.S. National Research Council's Office of International Affairs. The purpose of the project was "to conduct a series of activities to highlight applications of information and communications technologies to development and to examine ways in which those technologies can help USAID [the U.S. Agency for International Development] and other development assistance organizations better achieve their goals". We have chosen to reproduce the list of indicators because it provides a very sound framework for categorizing data that it would be useful to collect for the purpose of studying Internet diffusion.

INDICATORS RELATED TO THE ENVIRONMENT FOR INTERNET USE

Supportive Economy and Infrastructure

- GNP per capita
- number of telephones
- number of telephones per capita
- indicators of penetration of telephone service in rural areas
- indicators of penetration of electrical power in rural areas
- density of population in rural areas
- percentage of population in urban areas
- indicators of the strength of markets for personal computers, modems, and related technologies

Policy and Regulatory Environment

- estimated cost, time, and rate of success in establishing an ISP
- estimated cost, time, and rate of success in establishing an ISP account
- nondiscriminatory access to Internet service
- modem and/or computer tariff
- waiting time for a telephone line
- cost for installation of a telephone line
- waiting time for a leased line
- cost per minute to access points of presence (POPs)
- commercial availability of modems and computers
- local service for modems and computers

INDICATORS OF INTERNET SUPPLY

Quantity of Internet Service

- total number of ISPs
- total bandwidth to outside country (kilobytes/second)
- total number of modems connected to ISP servers for dial-up access
- total number of leased lines to customers
- total number of POPs
- total number of secondary-city POPs
- percentage of population within local calling area of POP

Quality of Internet Service

- percentage of send failures (messages that fail to reach their destination)
- average delivery time of e-mail/data transfer from each ISP to every ISP
- average delivery time of messages
- average time to check an empty mailbox
- mean connect speed of subscribers
- call failure rates for ISPs (the percentage of calls that fail to connect to the Web)
- number of members in an information industry association or ISP association
- number of ISPs offering full Internet service
- percentage of nonprofit ISPs
- number and percentage of profitable ISPs
- prices charged by ISPs for Internet access
- total funds invested by ISPs in expansion
- total ISP revenue

Sustainability

- number of foreign- and domestic-owned ISPs
- number of local technical staff
- number of ISPs offering user training
- number of institutions that monitor their own traffic, use, and number of hits on pages
- average number of years of schooling of adult population
- literacy rate
- number of information technology courses offered in universities
- average salary of Web designers and other ISP employees
- percentage of ISPs offering Web hosting, Web design, and other services
- ratio of national, regional, and international traffic to total traffic (both coming into and going out of the country)
- number of home pages on domestic servers
- ratio of national, regional, and international participation in listserves and news groups

INDICATORS OF INTERNET USE

- total number of subscribers by category of user
- average number of workstations per subscriber
- average number of people with access per workstation
- rate of change in the number of subscribers

- turnover rate
- total traffic (kilobytes per day)
- total connect time per day
- total number of e-mails per day
- average subscriber connect time
- average subscriber connections per day
- number of subscribers using leased lines
- Internet use for:
  - communication
  - downloading software
  - interactive discussions
  - noninteractive discussions
  - use of another computer
  - real-time audio or video
  - searches for product/service information
  - purchases based on Web information
  - searches for company/organization information
  - searches for other information on the Web
  - browsing/exploring
  - seeing what is new at a favorite Web site
  - business purposes
- percentage of users who connect from their own homes
- percentage of users who connect from an office
- percentage of users who connect from both home and office
- percentage of users who connect from an Internet cafe or business center

Costs of Internet Use

- price elasticity of demand
- fees paid to an ISP for leased-line, dial-up service, and other services
- installation fee(s)
- fees (fixed and/or usage dependent) paid to the telephone company
- price of a phone call per minute to connect with the ISP
- costs to the organization of Internet training courses and staff salaries paid during training

INDICATORS OF IMPACTS ON FORMAL ORGANIZATIONS

Perceived Benefits of the Internet

- number of messages/transactions to/from/by an organization per day that are domestic versus regional versus international in source or destination
- reported relative importance of the Internet versus other means
- cost savings on communications

- time savings on communications
- ratio of Internet to other channels in obtaining information
- percentage of an institution's dissemination through the Internet
- percentage of total public information made available through the Internet
- number of Web server hits or requests fulfilled per month from domestic versus regional versus international sources
- number of electronic newsletters or bulletins produced
- number of subscribers to newsletters and/or bulletins
- number and percentage of people trained in using the Internet
- number and percentage of subscribers with a LAN
- number of top/middle/lower-level users in an organization with access to the Internet
- relative importance placed on the Internet by top/middle/lower-level staff
- number of networks and "virtual organizations" of which an institution is a member
- investments in computer and other telecommunications facilities
- approximate number of users who (1) use e-mail, (2) "surf," (3) maintain own (individual or organization) home page, and (4) use an Intranet
- presence of a distinct information strategy as part of an organization's overall organizational strategies and plans

Organizational Decision Making

- change in number of people involved in an institution's decision making
- relative importance of the Internet versus other means of gathering data and information in decision making

INDICATORS OF SECTORAL IMPACTS

Sectoral Use and Diffusion of the Internet

- total number of subscribers per sector
- increase in the number of subscribers per sector
- percentage of Internet use per sector for (1) e-mail, (2) "surfing," (3) maintaining a (individual or organization) home page, and (4) use of an Intranet
- number of subscribers in primary city and secondary cities

INTERNET IMPACTS ON SECTORS AND THEIR RELATED DEVELOPMENT GOALS

Education

- number of schools/universities with Internet access
- number of students with Internet access

- average time of student access
- number of teachers with Internet access
- number of training courses on the Internet offered to teachers
- quality of training courses on the Internet offered to teachers (accreditation)
- number of new courses offered since the Internet was introduced
- number of schools/universities utilizing distance education via the Internet
- number of courses that supplement conventional teaching methods with distance education or other Internet-dependent technologies
- number of students enrolled in distance education
- number of nonuniversity institutions offering distance education
- ratio of job placement of students with Internet experience/training in school to overall placement
- ratio of average starting salaries of individuals with Internet experience/training in school to overall starting salaries
- number of scholars/researchers attracted to a university/country (in part) because of Internet access

Private Sector

- rates of participation of African firms in international markets
- rates of participation of foreign firms in African markets
- numbers of Web pages providing information on a market
- numbers of persons communicating about a market on the Internet
- volume of transactions in a market using the Internet
- number (percentage) of chambers of commerce with Internet access
- number (percentage) of other business organizations with Internet access
- number of small and medium-sized enterprises with Internet access
- number of small and medium-sized enterprises posting products and prices on the Internet
- rate of change in the value of an enterprise's exports (imports) since acquiring Internet access
- rate of change in the value of a country's exports (imports) since acquiring Internet access
- rate of change in the value of a firm's exports since acquiring Internet access
- number of companies reporting growth since availability of the Internet
- number of firms engaged in electronic commerce
- value of sales via the Internet
- funds allocated by private companies to Internet-related training
- growth rates of private telecenters that provide Internet services

Government and Civil Society

- number of ministries/departments with a presence on the Web
- number of ministries/departments with e-mail reply addresses on the Web
- quality of Web site content in the above-described classes of Web sites

- percentage of ministries/departments who use it for dissemination of information about governmental actions or policies
- number of political parties with a presence on the Web
- Internet access to government policy papers and pending and existing legislation and regulations
- number of organizations using Internet networks, user groups, etc., to influence government
- number of list servers, news groups and conferences holding on-line discussions of public policy issues
- number of NGOs with Internet access
- number of publicly-available sites with free or low-cost Internet access, such as kiosks, post offices, community centers, or libraries
- number of independent sources of information and news provided via the Internet
- number of newspapers, radio stations, and TV stations using the Internet to collect news
- number of newspapers, radio stations, TV stations, and other media with Web sites
- percentage of domestic and foreign readers