Laboratory of Economics and Management
Sant'Anna School of Advanced Studies

Piazza dei Martiri della Libertà, 33 - I-56127 PISA (Italy)
Tel. +39-050-883-341 Fax +39-050-883-344
Email: lem@sssup.it Web Page: http://lem.sssup.it

LEM
Working Paper Series

Measuring and modelling Internet diffusion using second level domains: the case of Italy

Andrea Bonaccorsi [*]
Maurizio Martinelli [†]
Cristina Rossi [‡]
and
Irma Serrecchia[§]

[*] Sant'Anna School of Advances Studies
[†] IIT-CNR, Italy
[‡] Sant'Anna School of Advances Studies
[§] IIT-CNR, Italy

2002/17 **June 2002**

ISSN (online) 2284-0400

Measuring and modelling Internet diffusion using second level domains: the case of Italy

Andrea Bonaccorsi
Sant'Anna School of Advances Studies
P.zza Martiri della Libertà 33, 56124 Pisa, Italy
Phone: +39 050 883323; Fax: +39 050 883 344
bonaccorsi@sssup.it

Maurizio Martinelli
IIT-CNR
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy
Phone: +39 050 3152087; Fax: +39 050 315 2113
maurizio.martinelli@iit.cnr.it

Cristina Rossi
Sant'Anna School of Advances Studies
P.zza Martiri della Libertà 33, 56124 Pisa, Italy
Phone: +39 050 883591; Fax: +39 050 883 344
cris@sssup.it

Irma Serrecchia
IIT-CNR
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy
Phone: +39 050 3152086; Fax: +39 050 315 2113
irma.serrecchia@iit.cnr.it

ABSTRACT

The last 10 years witnessed an exponential growth of the Internet. According to Hobbes' *Internet Timeline*[1], the Internet hosts are about 93 million, while in 1989 they were 100,000. The same happens for second level domain names. In July 1989 the registered domains were about 3,900 while they were over 2 million in July 2000.

This paper reports about the construction of a database containing daily observations on registrations of second level domain names underneath the "it" ccTLD[2] in order to analyse the diffusion of Internet among families and businesses.

The section of the database referring to domains registered by individuals is analysed. The penetration rate over the relevant population of potential adopters is computed at highly disaggregated geographical level (province). A concentration analysis is carried out to investigate whether the geographical distribution of Internet is less concentrated than population and income suggesting a diffusive effect. Regression analysis is carried out using demographic, social, economic and infrastructure indicators. Finally we briefly describe the further developments of our research. At the present we are constructing a database containing domains registered by firms together with data about the registrants; the idea is to use this new database and the previous one in order to check for the existence of power laws both in the number of domains registered in each province and in the number of domains registered by each firm.

Keywords

Domain names, Internet metrics, Diffusion, Power laws, Zipf's law

JEL Classification: L8, O3

[1] URL: http://www.zakon.org/robert/internet/timeline/.

[2] TLD is the acronyms of Top Level Domain. A top-level domain name can either be an ISO country code (for example *.be* stands for Belgium) or one of the generic top-level domains (a so-called gTLD such as .com, .org, .net and so on). To register a second level domain name (e.g. oecd.org) a user needs to apply to the domain name registry with the delegated authority for the ccTLD or gTLD. (OECD, 1998).

1. INTRODUCTION

The Internet is getting bigger, and it's happening very fast, but very different figures are circulating about the extent of this phenomenon. One of the most cited results is that Internet traffic is doubling each three or four months displaying a sort of Moore's law for data traffic. This statement, which was true for the period 1995-1996, is no longer acceptable: it would have produced absurd traffic volumes (Coffman and Odlyzko, 2001).

The main difficulty in measuring the Internet is its distributed nature: it has no central authority in control and no directory of users exists. Moreover, it is not possible to give an unambiguous definition of an Internet user. A lot of different definitions are present in literature dealing with the time spent on line (Federcomin, 2000), the age of the users, the kind of activity performed (e-mail, surfing the Web, ftp and so on). In order to overcome this problem, several Internet metrics are available. The most suitable are the so-called *endogenous metrics* that are *"obtained in an automatic or semiautomatic way from the Internet itself"* (Diaz-Picazo, 1999). These metrics have the unquestionable advantage of the accuracy and among them the most used in the literature are Internet hosts and second level domain names (Naldi, 1997; Zook, 1999; Bauer, Berneand and Maitland, 2002). The widespread utilisation of Internet hosts is probably due to the easiness in obtaining data. The organisations that manage the different ccTLD and gTLD, perform the hostcount[3] under their TLD on a regular basis and provide these data on the Web or by ftp. For instance every six months Network Wizard publishes the results about all the TLD on its web site[4], whereas the RIPE[5] publishes the data about the ccTLD in its area (Europe, North Africa, Middle East) monthly.

However, Internet hosts both under and over estimates the diffusion (Naldi, 1997). The under-estimation is due to the growing presence of firewalls and private networks (Intranet) together with the use of dynamic IP addresses for dial up accesses. Among factors of over-estimation the most important is the association multiple IP addresses to the same computer.

Among endogenous metrics, second level domain names represent a valid alternative to Internet hosts. This metric underestimate Internet diffusion: not all the users register a domain, nevertheless domains identify a *lower bound* in diffusion mainly capturing the proactive and interacting use of the network.

[3] For a definition of hostcount see OECD, 1998.
[4] URL: http://www.nw.com.
[5] URL: http://www.ripe.net/statistics/hostcoutn.htlm.

2. Methodology

In July 2001, the Italian Registration Authority (RA) for the ccTLD "it"[6] and the Sant'Anna School of Advanced Studies started a project for analysing the diffusion of the Internet network in Italy using domains registered underneath the ccTLD ".it".

Data were extracted from the databases of the registrations managed by the RA[7]. As a first step, registrants are grouped into several categories (individuals, firms[8], Universities and research centres, local public bodies, other public bodies and other registrants) in order to determine, for each category, the determinants of adoption and then of diffusion. A careful work of data cleaning was undertaken.

At the present, the classification of the registrations of individuals is complete while the classification for the other category is almost complete.

At July 18th 2001, WHOIS contained 52,401 entries referring to individuals: after correcting classification errors and eliminating registrants from other EU countries, 51356 entries for Italian registrant were included. This article illustrates for the analysis of the determinants of the adoption of an Internet domain name by individuals.

3. Results

3.1 Analysis of concentration

Only persons who are over 18 years old can register a domain under the ccTLD "it". The average age of registrants is quite low that is about 36 vs. about 49, which is the average age of the whole population. The class 28-37 is the most represented. About 85% of registrants are male while the total number of males in the population are about 48%. It seems that young people and males are more likely to register a domain. These data match the literature (Ingrassia, Comis and Mammana, 1995).

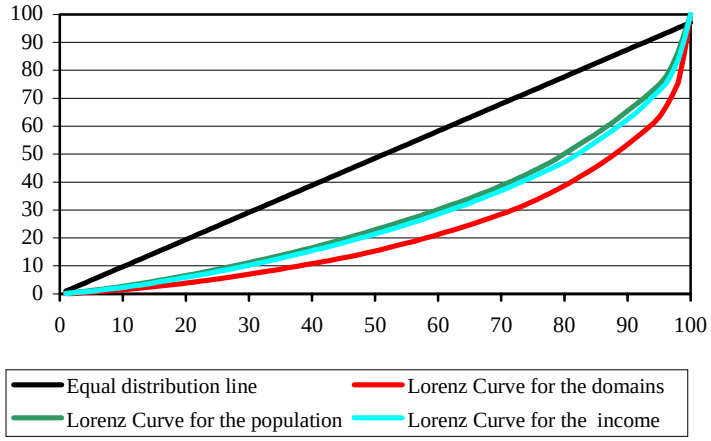
However, while a lot of surveys on samples of Internet users show the reduction of the *gender gap* for Web surfers, our data demonstrate that it is still present if we focus on an advanced and proactive use of the Internet network.

[6] The Italian Registration Authority is at the *Istituto di Informatica e Telematica* of *Consiglio Nazionale delle Ricerche* (CNR), Area di Pisa.
[7] Whois database, database of the state of the registrations, database of the letters of assumption of responsibility (LAR).
[8] In order to classify data we are using a database of Italian firms managed by *Infocamere*.

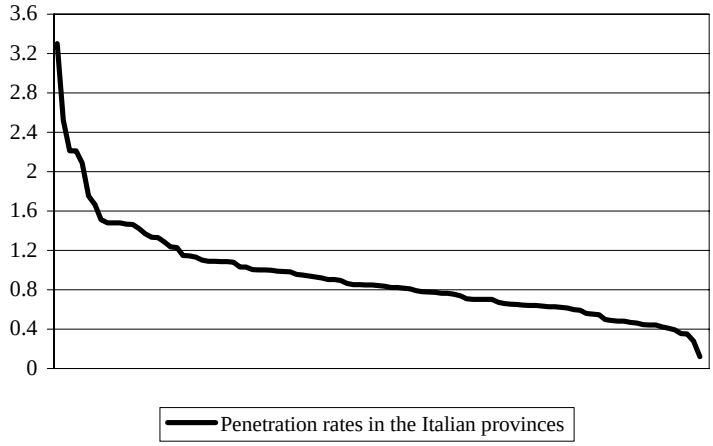
Italy is divided into 103 administrative units called provinces. The geographical distribution of domains is highly concentrated (Graph 1). The first three provinces, all including large cities (Rome, Milan, Naples) account for 29.1% of the total. The distribution of domains is much more concentrated than the distribution of the population: its Herfindahl index is 0.0418 vs. 0.0215, which represents the population concentration; the Gini index is 0.567 vs. 0.429. At the same time it is more concentrated than the distribution of the income: the Herfindahl is 0.025 and the Gini is 0.458. The penetration rate

$$\frac{Number\ of\ domains}{Population\ over\ 18}$$

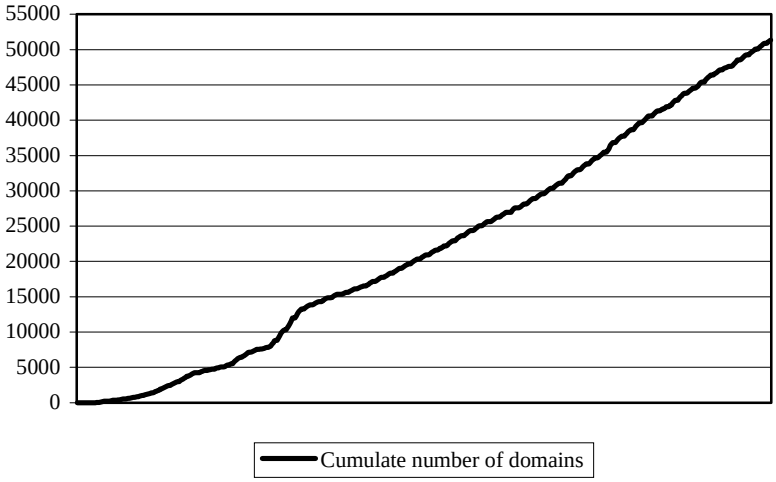
is between 0.3% and 0.03%. The three provinces having the higher penetration are of medium size.



Graph 1: Lorenz Curves for domains, income and population.



Graph 2: Penetration rates of domains in different provinces.



Graph 3: Cumulate number of domains registered by individuals.

These findings shed light on the diffusive effect of Internet, i.e. the potential of the technology to reduce differences in the concentration of population (provinces with large urban concentrations vs. provinces with small cities) and concentration of income (rich provinces vs. poor ones). A diffusive effect could be justified by the decentralized, non-hierarchical, immaterial nature of the Internet technology, which in principle should not have strong barriers to entry as it happens in manufacturing. Data show that this effect does not take place at all at the aggregate level. Domains are even more concentrated than population and income. A ranking of provinces by penetration rate, not reported here, clearly shows that the distribution of Internet follows large differences in the level of income. Before drawing conclusions, these data should be compared to those on the use of domains by business firms, and this comparison is currently in progress. Our preliminary conclusion is that, far from being an "equalizer", Internet technology follows and possibly sharpens existing differences in economic opportunities, not only across countries, but even within industrialized countries.

3.2. Determinants of adoption

In order to analyse the factors leading to the registration of domains by individuals, we run exploratory stepwise regressions using indicators at the province level. The dependent variable is the absolute number of domains in the province (Table 1) and the penetration rate (Table 2). Models are kept compact by only using explanatory variables of the same kind. Independent variables are not normalized.

Variables	M1: Economics Variables	M2: Skill Variables	M3: Social Variables	M4: Demographic Indicators	M5: Infrastructure Indicators	M6: Public Expenditure	M7: Industrial Demo-graphy
Employees in service sectors**	7.642 (.631)						
Employees in agricultural sectors**	-13.698 (3.076)						
Total disposable Income**	-1.735E-05 (.000)						
Number of patents**		1.644 (.141)					
Educational Infrastructure*		3.742 (1.342)					
Expenditure for movie theatres**			5.029E-02 (.005)				
Expenditure for sport events**			-5.845E-03 (.003)				
Expenditure for magazines**			-3.209E-02 (.010)				
Expenditure for newspapers*			8.468E-03 (.004)				
Density of population**				3.051E-02 (.001)			
Number of foreign residents**				.309 (.074)			
Telecommunication infrastructure**					12.915 (1.778)		
Port infrastructure**					-1.081 (.282)		
Cultural and recreational infrastructure**					3.378 (1.145)		
Energy and environmental Infrastructure**					-5.356 (1.863)		
Water and electrical systems**						1.399E-02 (.003)	
Public housing**						2.201E-02 (.004)	
Railroad and transportation works**						2.628E-03 (.001)	
Road and airport works**						6.981E-03 (.002)	
Housing**						-2.160E-02 (.008)	
Net increase in the number of join stock companies**							1.171 (.036)
R^2	0.958	0.641	0.953	0.946	0.517	0.748	0.912

*Table 1: Stepwise regressions with number of domains as dependent variable (** P<0.01, *P<0.05, Standard error in parenthesis)*

Variables	M8: Economics Variables	M9: Skill Variables	M10: Social Variables	M11: Demographic Indicators	M12: Infrastructure Indicators
Employees in service sectors**	3.374E-05 (.000)				
Per capita disposable income*	1.144 E-03 (.000)				
Per capita patents**		7.922E-04 (.000)			
Rate of enrolment to high schools (females)*		1.061E-05 (.000)			
Theatre and musical expenditure**			2.395E-06 (.000)		
Number of foreigners every 1,000 inhabitants**				2.971E-05 (.000)	
Banking and service infrastructure**					9.238E-06 (.000)
Energy and environmental infrastructure**					-3.691E-06 (.000)
Port infrastructure**					-4.086 E-07 (.000)
R^2	0.265	0.141	0.254	0.278	0.304

*Table 2: Stepwise regressions with penetration rate as dependent variable (** P<0.01, *P<0.05, Standard errors in parenthesis)*

Regressions in Table 1 show that the number of domains depends on the number of employees in the service sector. Total disposable income has a negative, but very weak effect.

Adoption depends on the skills available in the province, approximated by the patents and the educational infrastructure. Interestingly, while patents are a significant factor, the manufacturing sector, which produces most patents, is not relevant as such, suggesting that only a portion of it is important for Internet adoption.

Adoption also depends on cultural expenditure of higher quality (theatres and newspapers), while it seems to substitute for sports events and magazines. Internet is adopted more in densely populated provinces that are open to foreign residents. As expected, it is also more adopted in provinces with a larger telecommunication infrastructure, while other heavier infrastructures (energy and ports) have a negative impact. Again, cultural infrastructure is important.

Public expenditure in material infrastructures (water, road, railroad, airport, public housing) is also highly significant. Interestingly enough, Internet does not abolish the need for material

infrastructure at local level. Finally, the net increase in limited liability companies is also relevant, while the net increase in simpler types of firms is not significant.

Models using the penetration rate as a dependent variable basically confirm this picture (Table 2).

Provinces that are densely populated, tertiary, highly schooled and skilled, culturally open and internationalised, with adequate endowments of infrastructure and an advanced entrepreneurial environment are the best candidates for a more active and interactive use of the Internet.

The main problem of the models presented above is multicollinearity. Stepwise regression eliminates strong correlated variables: all variables must pass a tolerance[9] criterion (tolerance level 0.0001) to be entered in the model and a variable is not entered if it caused the tolerance of another variable already in the model to drop below the tolerance criterion. Anyway, an analysis of variance and covariance matrices is necessary, results are reported in the following tables. Let's start with the models having domains as a dependent variable.

4. Further developments of the research

4.1 Analysis of registrations by firms

At the present, we are constructing the database containing domains registered by firms. The naming rules established by the Italian Naming Authority[10], state that individuals can only register one domain name. This restriction does not apply to firms.

For each firm that has registered a domain, our database will report the date of registration, the name of the firm, its province of location and its juridical form. Particular attention is paid to the distinction between entrepreneurs and companies.

This database will allow to study the determinants of the adoption of domains by firms and then the pattern of diffusion of this technology in the Italian production system. On one hand, we will perform regressions by using indicators at the province level, on the other hand we will construct a sample of firms for which data on structural characteristics (with particular attention to the size) will be collected in order to test their influence on adoption.

[9] Tolerance: A statistic used to determine how much the independent variables are linearly related to another (multicollinear).

[10] Naming rules are available at http://www.nic.it/NA/index-engl.html.

4.2 The distribution of domain names

At the present a growing body of literature is devoted to discover and analyse the regularities displayed by the Internet network (Pitkow, 1998). In particular, the presence of power laws (Blank, Solomon, 2000) for Internet related phenomena is widely accepted.

Power laws are discovered in the number of in and out links of a web site (Barabasi, Albert, 1999; Adamic, Huberman, 2000), in the number of pages composing an Internet web site, in the behaviour of Internet surfers (Huberman, Pirollo, Pitkow, Lukose, 1998; Johansen, 2001).

Among the several types of power laws that have been detected, Zipf's law (Zipf, 1949) seems to play a central role. This law has been especially applied to the analysis of the distribution of the population of cities in a country or in a region (Gabaix, 1999; Krugman, 1996, 1998). Under this law if you order cities according to population and plot in a graph the logarithm of the population against the logarithm of the rank, you obtain a straight line whose slope is about -1. So you can conclude that the population of the first city in a country is approximately twice as much as the population of the second city, three times as much as that of the third city and so on.

Moreover, this seems to be a wide general law that has been applied in linguistics (for the frequency of words in a text, Mandelbrot, 1965; Alexander, Sidorov, 2001), in the study of the intensity of earthquakes (Sonette et al., 1996) and in several fields of biology and physiology (Jorgensen., Mejer and Nielsen, 2001). In economics it has been applied to the distribution of firm sizes (Axtell, 2001), measured in different ways (income, assets, number of employees) (Okuyama, Takayasu, Takayasu, 1999).

The idea is then to investigate the presence of power laws for domains registered by individuals in Italian provinces and later for domains registered by firms and other organisations. In particular, given that firms can register more than one domain, it is of interest to investigate if there is a power law in the number of domains registered by each firm.

If a power law is discovered, the challenge will be to analyse how it is generated. In fact, at the present there is not any generally accepted theoretical foundation for this empirical regularity.

4. REFERENCES

Adamic L., Huberman B. (1999) *Growth dynamics of the World Wide Web.* Nature, 401, page 131.

Albert L., Barabasi, R. A. (1999) *Emergence of scaling in random networks.* Science, 286(5439) pages 509-512.

Alexander G., Sidorov G. (2001). *Zipf and Heaps laws' coefficients depend on language.* Proceeding of Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2001), ed. Alexander Gelbukh, Lecture Notes in Computer Science, Vol. 2004 (Springer-Verlag), pages 332-335.

Axtel R. (2001) Zipf distribution of U.S. Firm Sizes. *Science, 293, pages 1818-1820*.

Blank A., Solomon S. (2000) *Power laws in cities population, financial markets and Internet site (scaling in systems with variable number of components).* Physica A, 287, pages 279-288.

Coffman K.G. and Odlyzko A.M. (2001). *Internet growth*. AT&T Labs report.

Diez-Picazo G.F. (1999) *An Analysis of International Internet Diffusion.* Ph.D. Thesis, MIT.

Fujita M., Krugman P., Venables A. (1998) *The spatial economy*. MIT press, Massachusetts.

Gabaix X. (1999) *Zipf's Law for cities: an explanation.* Quarterly Journal of Economics, pages 739-767.

Ingrassia S., Comis E., Mammana M. (1995). *Internet in Italia-Un'indagine statistica*. Università degli Studi di Catania

Johansen A. (2001) *Response time of Internauts*. Physica A, 296, pages 539-546

Jorgensen S., Mejer H., Nielsen S. (2001) Ecosystem as a self-organizing critical system. *Ecological Modelling, 111, pages 261-268.*

Krugman (1996) *The self-organizing economy.* Blackwell, Mandell

Mandelbrot B. (1965) *Information theory and psycholinguistics.* In Scientific Psychology: Principles and Approaches, eds. B. Wolman, E. Nagel (Basic, Books,1965), pages 550-562.

Naldi M. (1997). *Size estimation and growth forecast of the Internet*. Centro Volterra, Tor Vergata

OECD (1998) Intern*et infrastructure indicators.* OECD report.

Okuyama K., Takayasu M, Takayasu, H. (1999) Zipf law in income distribution of companies *Physica A, 269,pages 125-131*.

Sornette D., Knopoff L., Kagan Y., Vanneste C. (1996) Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. *Journal of Geophysical Research, 101, pages 13883-13894.*

Zipf K.G. (1949) *Human Behaviour and the principle of least effort*. Addison-Wesley, Cambridge, Massachusetts.

Zook M.A. (1999). *The Web of Consumption: Spatial Organisation of the Internet Industry in the United States*. American Behavioural Scientist (forthcoming).