INTERNET DIFFUSION & INTERNET DOMAINS: LOOKING FOR A NEW METRIC. THE CASE OF REGISTRATIONS BY ITALIAN INDIVIDUALS

Andrea Bonaccorsi, Cristina Rossi
Sant’Anna School of Advances Studies
Piazza Martiri della Libertà 33, 56124 Pisa, Italy

Maurizio Martinelli, Irma Serrecchia, Daniele Vannozzi
Institute for Informatics and Telematics-CNR
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy

ABSTRACT
This paper reports about the analysis of the Internet diffusion among families and business basing on daily observations on registrations of second level domain names underneath the “.it” ccTLD. In particular, domains registered by individuals are 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 building a database containing domains registered by firms in addition to domains registered by individuals. The idea is to use the database 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.

1. INTRODUCTION
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 under their TLD on a regular basis and provide these data on the Web or by ftp. For instance every six months Network

---

1 ccTLD is the acronym of country code Top Level Domain (for example .be stands for Belgium, .it stands for Italy). To register a second level domain name (e.g. cnr.it) a user needs to apply to the domain name registry with the delegated authority for the ccTLD or gTLD. (OECD, 1998).

2 For a definition of hostcount see OECD, 1998
Wizard publishes the results about all the TLD on its web site, whereas the RIPE publishes the data about the ccTLD in its area (Europe, North Africa, Middle East) monthly.

However, it is demonstrated that Internet hosts can both under or over-estimate the diffusion (Naldi, 1997). Among endogenous metrics, second level domain names represent a valid alternative to Internet hosts. This metric underestimates 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.

2. METHODOLOGY

In July 2001, the Institute for Informatics and Telematics of the Italian National Research Council (IIT-CNR), who is the “.it” ccTLD manager, and the Sant’Anna School of Advanced Studies started a project for analysing the Internet diffusion in Italy using domains registered underneath the ccTLD “.it”.

Data were extracted from the databases of the registrations managed by the IIT. As a first step, registrants are grouped into several categories (individuals, firms, 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 diffusion. A careful work of data cleaning was undertaken.

At present, the classification of the registrations of individuals is complete while the classification for the other categories is ongoing.

At July 18th 2001, WHOIS database contained 52,401 entries referring to individuals: after the classification and the elimination of registrants from other EU countries, 51,356 entries for Italian registrants were included. This article analyses the determinants of the adoption of an Internet domain 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 with respect to the average age of the whole population, which is around 36 vs. 49 years. The class 28-37 years is the most represented. About 85% of registrants are male while the total number of males in the population is about 48%. It seems that young people and males are more inclined to register a domain. These data agree with 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.

Italy is divided into 103 administrative units called provinces. The geographical distribution of domains is highly concentrated. 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

\[
\frac{\text{Number of domains}}{\text{Population over 18}}
\]

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

---

5 In order to classify data we are using a database of Italian firms managed by Infocamere.
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 and the penetration rate. Models are kept compact by only using explanatory variables of the same kind. Independent variables are not normalized.

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 scenario.

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 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. We will examine models having domains as a dependent variable.

6 Tolerance: A statistic used to determine how much the independent variables are linearly related to another (multicollinearity).
4. FURTHER DEVELOPMENTS OF THE RESEARCH

4.1 Analysis of registrations by firms

At present, we are populating the database with domains registered by firms. The naming rules of the “.it” ccTLD, state that individuals can only register one domain. This restriction does not apply to all the other registrants.

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 legal form. Particular attention is paid to the distinction between entrepreneurs and companies.

The 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. Moreover, given that we have the time series of the registrations starting from the early Ninety, we will estimate the hazard rates testing also for the presence of network externality phenomena. The idea is to test if the rate of diffusion in t+1 depends on the rate of diffusion in t.

4.2 The distribution of domain names

At 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 ordering cities according to their population and plotting in a graph the logarithm of the population against the logarithm of the rank, a straight line whose slope is about –1 is obtained. Therefore, it turns out that the population of the first city in a country is approximately twice of the population of the second city, three times of the third one 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 cities. If a power law is discovered, the challenge will be to analyse how it is generated. In fact, at present there is not any generally accepted theoretical foundation for this empirical regularity.

REFERENCES

INTERNET DIFFUSION & INTERNET DOMAINS: LOOKING FOR A NEW METRIC. THE CASE OF REGISTRATIONS BY ITALIAN INDIVIDUALS