Do we live in a small world? Measuring the Spanish-speaking Blogosphere*

Fernando Tricas
(Departamento de Informática e Ingeniería de Sistemas, U. Zaragoza, Spain)

Víctor Ruiz (Blogalia.com)
Juan J. Merelo
(Depto. Arquitectura y Tecnología de Computadores, U. Granada, Spain)

30th April 2003

Abstract

The blogosphere is the community of bloggers, people or collectives who share information and opinions ordered chronologically. The Spanish-speaking blogosphere contains several thousand blogs; despite its small size, compared to the English-speaking (or maybe global) blogosphere, its characteristics are a bit different.

In general, it could be said that the Spanish blogosphere has not reached critical mass yet. Moreover, the main reference of the Spanish-speaking blogosphere is still the English-speaking web; most links found point outside the Spanish-speaking web.

In particular, it is still quite uncommon that news items seen or generated in the Spanish blogosphere become popular throughout it; when this happens, most of the time it's due to the reproduction of the English blogosphere. There is also an "increasing returns" phenomenon: most bloggers (and readers of blogs) concentrate in some blogging sites (such as Blogalia or BarraPunto), and so they dominate the link space of the whole blogosphere. Finally, there is a third characteristic: the Spanish-speaking blogosphere is slower than the English-speaking one: ideas, topics and links spread in a slower way.

This paper will show our experience in developing blogging tools, in particular, the "Blogómetro" (http://blogometro.blogalia.com), which is an open source program that checks on a daily basis the link space in the Spanish-speaking blogosphere, in a similar way to BlogDex or Daypop, which check the English-speaking blogosphere (and a small part of the global one). We will show and analyze data gathered from the end of the year 2002 to the beginning of 2003.

*Fernando Tricas work is partially supported by the Spanish research project CICYT TIC2001-1819 Authors want to thank Adela Torres for her English revision. Any fault you can find about the language has been added later by us.

1 Introduction

The Spanish blogosphere is, obviously, part of the global blogosphere, and is roughly defined as the set of blogs (or, sometimes, blog-looking web pages) that are written in Spanish (in any part of the wold) or in any other of the official languages in Spain (Catalan, Basque, Galician). We have also considered blogs written in other languages, if they are written by people living in countries that naturally fit in the Spanish blogosphere area of influence (for example, we heard recently about some Venezuelan bloggers that write in English to point the mainstream attention to their country, or the 'Trilingual blog', `http://trilingual.blogspot.com/`).

It is a quite active group, albeit often disregarded by the Spanish[1] traditional media, who, very often, when they write a generalist article on blogs, mention any popular English-speaking blog (like for instance Megnut, Meg Hourihan's blog) instead of the most popular Spanish blogs (for which there are several candidates, depending on the measure we use, as we will see).

Why would anyone want to undertake measurements on this? First of all, because we can. That is, the Spanish blogosphere has a size that is barely manageable by the analysis software we use; any bigger size would probably crash it, or the study would require other strategies. Second, because we want to understand it, understand its mechanisms, and see the position of each blog (notably our own) within this blogosphere. Third, because we want to look at its possible problems (fragmentation, disregard of itself, very oriented to the English blogosphere[2]), and point out possible solutions. Four, because we believe that this work will help to improve self–consciousness, it will bring more people to the activity, and it will help to establish relations. Furthermore, we hope that this knowledge will enable us to suggest a mechanism to improve communication within the Spanish blogosphere, and make it more visible to the non–blogging community.

The rest of the paper is organized as follows: next section is devoted to explain the framework we have used for this analysis; section 3 will be devoted to *big numbers*, overall macroscopic measures in the Spanish blogosphere; it will be followed by section 4 which will study the Spanish blogosphere as a social network. Finally, we will present our conclusions and some hints on how this work will follow in 5.

2 Methods and Context

In this section we are going to concentrate on our tools, we will make some comments about the Spanish blogosphere, and we will provide some hints about other projects of interest.

[1] In the context of this work we will use this term often not only in the sense of 'from Spain' or 'speaking in Spanish' but also to refer to things related to the considered blogs, as described above.

[2] Again, we are talking about blogs written in English, independently of location

2.1 The tools

Our project started at the beginning of the Summer of 2002, and it is in a 'beta' stage. All data in this study have been taken from the **Blogómetro**, a suite of tools whose main visible aspect is its blog (`http://blogometro.blogalia.com`), hosted in Blogalia (`http://www.blogalia.com/`). There, a list of fresh links taken from our list of blogs (ranked by the number of sites pointing to them) is published daily. The Blogómetro is an open–source collaborative project, offering our research to the community. It is open to the participation of interested people. In this sense, not only its source code is available at the project page (`http://sourceforge.net/projects/blogometro`), but also the list of sites scanned daily.

The bot that cralws the sites is written in Python and every day, early in the morning, checks all blogs in the list. From each raw HTML file, it scrapes the links, and stores them in a database if they have not been included before. In consequence, a link is considered *new* or *fresh* if its URL has not been seen before in that particular blog; that means that if a blog refers to another several times by its URL, it will count as a single reference; that also means that links included in the *blogroll* list are considered only once (during the lifetime of the data).

The DBMS is PostgresSQL, a free, open source program, that is easily interfaced via languages such as Perl or Python. The database contains only two tables, one for the blogs themselves, with URL and description, and another for the URLs. Data has been stored for approximately 4 months, from November, 15, 2002, to April, 15, 2003. The database was purged in two ways:

- First of all, as the amount of data grows very fast, we need to delete from time to time what we consider 'irrelevant links'; that is, links that only appear once in the database, and are old enough (two months at this moment), in order to keep the database manageable with our hardware.
- Second, and only for the analysis of the social network aspects, self–links have been excluded when possible. That is not always possible, except in the case that links to a blog include the blog's URL. People use their own customized systems of publishing (frames, iframes, skeleton in one domain and postings in other different one, ...) making it difficult to do this part of the work.

The addition of new blogs to our list is done by hand and, even though we have found moderate interest about the project, we have not received many submissions of sites from other people, so the list is mainly ours (with the pros and cons this may have).

2.2 About the Spanish blogosphere

As far as we know, the weblogging phenomenon started with BarraPunto (`http://barrapunto.com/`), a collective weblog oriented to provide news and discussion about free and open source software. In fact, it started as a Spanish clon of Slashdot (`http://slashdot.org/`). Some features make BarraPunto different from Slashdot (and more interesting from the point of view of blogging and, in particular, personal blogging); let us remark two of them.

Table 1: Number of blogs hosted in popular sites in the Spanish blogosphere

Site	Number of blogs
Blogging sites	
BarraPunto	648
Blogspot	315
Blogalia	164
Pitas.com	24
Antville	16
General hosting sites	
i! (España)	34
geocities.com	22
cjb.net	19

- The 'miBarrapuntos': an adaptation of other 'my' services that allow people to create their own BarraPunto-like sites, with support for comments, sections, topics, and even collaborative weblogs. It is a more advanced tool than the 'Diaries' you can find at the standard Slash code.

- Ecolutions: a mechanism to get any history posted in any part of the site, and copy it -maybe with some editing, and with a link to the original story- to our own blog.

In parallel, many people started blogging, using some of the 'standard' sites: mainly BlogSpot (http://blogspot.com/) and others, but also home-made blogs (using free or paid hostings, and free or commercial tools). Finally, let us comment the site that provides hosting to our tools. Blogalia (`http://www.blogalia.com/`) was created at the beginning of the last year, and it has become a populated site, with a very strong sense of community. It started as a personal project of one of us, Víctor Ruiz, to provide a simple and usable tool to Spanish–speaking people interested in starting a blog. It was also planned as a site where established bloggers could find a more friendly tool.

In Table 1 we can see the sites where most of the Spanish blogs are hosted. They are approximately half of the total number of blogs in our list, so there exist a wide set of blogs made with very different tools and hostings. They have been separated in sites oriented specifically to the hosting of blogs and other sites, oriented to general free–hostings. It is interesting to note that we are not aware of blogs hosted in some of the other popular blog–hosting sites. Further research will be needed in this area, we hope that people will start sending sites when our project will be more widely known. As a final remark, le us notice that there are some bloggers that use these same hosting sites hidden behind their own domain, so this table can show a slightly different picture of the reality.

2.3 Related initiatives

We want to talk here about other initiatives born in the Spanish blogosphere that partially overlap with our work, either in the community construction side, or in the mea-

suring aspects.

First, let us concentrate on the community aspects:

- During last Christmas, a game that maybe many of you have played as children was proposed in the Spanish blogosphere: the 'Ciberamigo Invisible' (`http://www.awacate.com/amigo/`), a blogging version of this well-known game, also know as 'Secret Santa' in some places. The idea is to exchange gifts in a group of people in such a way that each person has to give something to another member of the group, with pairs selected in a random way. In our case the gifts were virtual things such as: graphics, banners, texts, ...

  It was an important social success with around one hundred participants (Anecdotally, a search in Google (`http://www.google.com/`) for 'Amigo invisible' gives the page of the event as the first one in the list at this moment).

- Concentrating on more technical aspects, and as an implementation of the 'Ridiculously Easy Group Forming' (`http://www.myelin.co.nz/cgi-bin/wcswiki.pl?RidiculouslyEasyGroupForming`), Philip Pearson created 'The internet Topic Exchange' (`http://topicexchange.com/`). At this site, anybody can create a 'channel' dedicated to any topic of his interest. Then, anybody blogging about this topic can send an adequately crafted request, and gets its contribution listed in the channel page. Surprisingly, the most active topics at this moment belong to the Spanish blogosphere: the 'bitacoras' channel (`http://topicexchange.com/t/bitacoras/`, whose main topic is blogging about blogs) and 'directorio de blogs hispanos' (`http://topicexchange.com/t/directorio_blogs_hispanos/`, whose objective is to be a way to publish newborn blogs).

- Finally, there are a number of directories that try to offer comprehensive listings of blogs with different classifications but, as far as we know, none of them has a list longer than ours.

Now, let us talk about some more technical approaches.

- The 'vecindario' (`http://www.pisotrece.com/vecindario/`) is an implementation of the 'Blogging Ecosystem' (`http://www.myelin.co.nz/ecosystem/`) that uses a lists of blogs smaller than ours (around 700 blogs) but that shows interesting results.

- Very recently, in the last days another project has been shown: the 'Mapa de la Blogosfera hispana' (`http://www.hiperespacio.com/blogosfera/`) which is a hand–made graphic representation of the relations of the Spanish blogosphere as perceived by the author.

In some of these projects and also as a personal feeling of our own work, the conclusion is that it is very difficult to get an idea about the size and extension of our blogosphere. One of the problems detected is that not many blogs ping sites like Weblogs.com (`http://www.weblogs.com/`) that would allow us (and others) to try some kind of auto–discovery. Most of the other blog–related tools rely on this site, so we are out of luck in this automation.

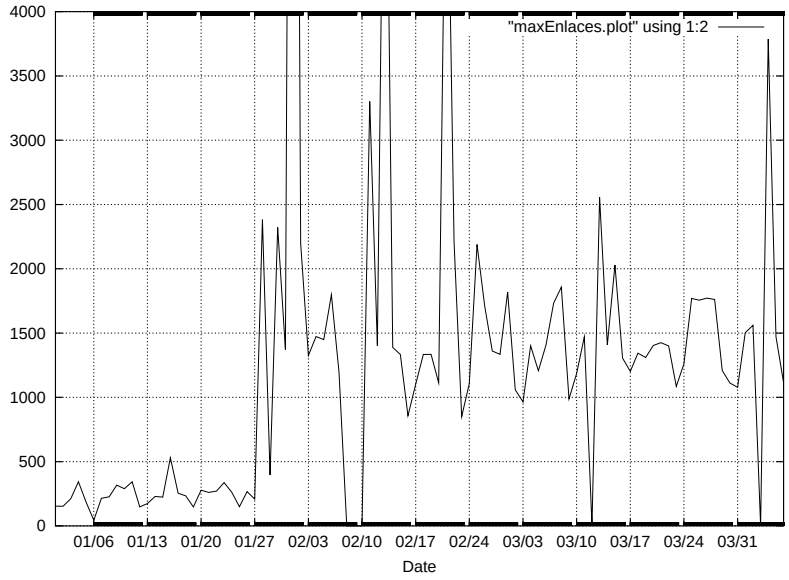

Figure 1: Evolution of the number of fresh links per day.

3 Big Numbers

During the period of study 281648 links were observed (109687 excluding self-links and purged entries), which yields an average of 1573 links in a day; considering the number of blogs, this makes an average of 1.14 links per blog per day. present at If we assume that each new history posts a new link, we will have around 1500 new histories a day in the Spanish blogosphere. Not a big deal, but at least we have a ballpark estimate. Another measure that can be of interest is the activity of the blogs in our list: during the last month 1160 have posted at least one history, with a link. We can compare these data with the activity in January (1299 blogs), and in November (943 blogs). We can see the number of daily links in Figure 1. In this figure the grid shows weekly periods, allowing us to see some periodical behavior. There are some spurious peaks (too high, or zero) due to problems with our spider (network failure, bugs in the program, accidental deleting of some data that is recovered in the same scrapping, ...).

From the point of view of blogging, the week clearly starts at Mondays and grows until the weekend: at the end of these periods the number of links shows always a decreasing pattern; there is almost no life during the weekends. Our guess is that there are many techies blogging from work (even as part of their work, as a way to interchange information with others) and that during weekends they tend to be disconnected. If we pay more attention, we can discover a curious descending peak in most of the weeks, corresponding to Wednesdays (some weeks on Tuesdays): it seems that people start the week wanting to blog, and they needed to relax in the middle of the week, to continue later with the activity. Finally, let us point the difference of size in the weeks previous

to the one starting at Jan, 27 that corresponds to the last purge of the database. Anyway (and supporting our definition of uninteresting links) the weekly structure is also observable with sets of data that have been purged.

What was the most popular link during that period? Unsurprisingly, the first 20 links or so are taken by banners that have appeared in weblogs, such as popular weblogs such as Slashdot (number 2), and blog hosting sites and software such as Slash, Blogger, Blogspot and Blogalia. BarraPunto (`http://www.barrapunto.com`) starts to show up here, first, by itself, and then, by having its banners as the most pointed-to links). The first *real* link is `http://www.librodenotas.com/mt/prestige.html` (75 links), a (quite critical) page on the Prestige wreck, which was part of a campaign to Google–bomb the word *prestige*. It obviously succeeded. The daily newspapers "El País" and "El Mundo" show up roughly the same number of times, 48 and 45, with El País having a slight edge. However, this counts only references to the main page, not to particular news. Unsurprisingly, this edge is lost when we take into account all references: "El Mundo" is four times as popular as "El País", 887 vs 278; this is not surprising, since El País switched recently to a pay-per-content model. That is why it is almost reached by "Periodista Digital" (`http://www.periodistadigital.com`, 204 links), who usually makes some of El País content publicly available under the "fair quoting" provision of the copyright act. Looking at individual blogs, *Mini-D* (`http://www.minid.net`), a popular weblog on design and current events is the most popular one with 151 links, and *Cuaderno de Bitácora* (`http://rvr.blogalia.com/`), one of our own (Ruiz's), with 130 links, the second. The links from one point of the blogosphere to another represent roughly one tenth of the total, namely, 12533 for the studied period. Within these links, once again, a popularity contest takes place, with Libertonia (`http://libertonia.escomposlinux.org`) winning hands on. However, much of the links to Libertonia come from its own blog-like *diaries*, so maybe we should consider Libro de Notas. (`http://www.librodenotas.com`) the winner[3]

4 Spanish blogosphere as a social network

The data obtained from the Blogómetro have been analyzed using a number of software tools, most of which have crashed under the load (very specially GraphViz and UCINET (several times)). However, we have managed to obtain some useful information from them. The first analysis performed was to check if the Spanish blogosphere fits itself to a power law. We fitted a power law $f(x) = kx^a$ using the open-source tool GnuPlot, resulting $a = 0.58$ and $k = 328$. However, the chi-square test was around 6, much bigger than one, which means that the model does not fit itself well to the data. Data and function are shown in figure 2 Besides the fact that this data does not fit to the model, unlike data published by Kottke in `http://www.kottke.org/03/02/030212screw_the_po.html`, where it finds a perfect fit for the (global? English–speaking?) blogosphere, and an exponent of -0.83.

[3]We also live in the blogosphere, so let us share with you our ranks #7 (Ruiz's), 13 (Merelo's) and 14 (Tricas's), just in case you wanted to know.

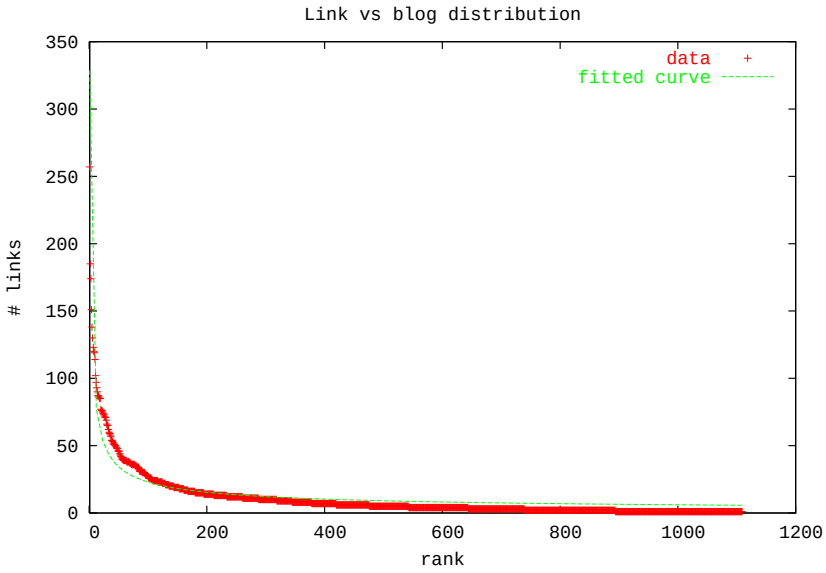

Figure 2: "Powerlaw" distribution of links per blog; x axis represents the blogs ordered by number of incoming links, and y axis the number of links. Data points have been fitted to $f(x) = 328x^{-0.58}$; however, the fit is not good enough.

Our working hypothesis here is that there is some fundamental *critical mass* that makes a certain community behave like a power law; unfortunately (or maybe fortunately), the community we belong to does not seem to have reached that level yet.

Other superstructures on the blogosphere were analyzed using Visone (available from `http://www.visone.de`). Visone is a tool that allows to plot maps of the social network under analysis, as well as perform some measurements on it. One of these measures is the *betweenness*, which roughly measures how often a particular blog is found when traveling using links from one blog to another. In that particular sense, as can be seen in Figure 3, *eCuaderno* (the blog of another BlogTalk speaker, José Luis Orihuela, #208, `http://orihuela.blogspot.com`) takes the central place among all Spanish blogs (with almost 6% centrality). Several others, including `http://bitacoras.net` (#8) and `http://www.gistain.net/` (#14) also are lying prominent places.

These "central" blogs play a prominent role within the blogosphere: they register memes in the community, and expose them, so that they can be picked up by other blogs, and thus, act as veritable "meme mills" that spread memes throughout the Spanish blogosphere. This fact is also reflected in other two measures: "hub" and "authority". The first *hub* is, once again, Libertonia, but taking into account that most of its links would be automatically generated, we will consider the second, 'Fernand0's barrapunto', a journal hosted by "BarraPunto" and authored by one of us (Tricas) the hub in the Spanish community; curiously enough, our other two BarraPuntos (Victor's and Merelo's) are placed 4th and 5th. This might be mainly due to the fact that we are editors within BarraPunto, and have automatic references when a person places one of the newsitems edited by us in their own journals. Maybe this only emphasizes the role that collective blogs such as BarraPunto, EsCompOsLinux of PuntBarra (the catalan-language equivalent) play as hubs in the community. Actually, they take 9 out of the first ten places, the other corresponding, you guessed it, to one of us: Tricas, whose manually edited blog is #9. Hubs' main feature is how often they quote other blogs; the opposite, being quoted, is measured by the authority quotient, which was also measured using Visone. Authorities are "quotables", in the sense that other blogs mention their histories very often. Most of the 10 first places are taken by collective blogs, notably Libertonia, but some others show up: PJorge (`http://www.pjorge.com`), which seems thus widely respected within the community (#7), simbiosis (`http://simbiosis.blogalia.com`), a blog that daily comments other blogs, and our own (#9, 10 and 11). Linux and open source-related blogs are the most prominent in these places, which shows its importance within the Spanish community, since maybe blogging in Spanish had its origin in it. Authority measures take into account not only the number of incoming links, but also their "quality", that is, the authority of the blog that posts them. Finally, another macroscopic measure is the "degree of separation", that is, the average number of links needed to reach a blog from another. It has been measured using UCINET, yielding a value of 3.761, that is, on average, approximately 4 links separate each Spanish blog from any other. That means that, quite matter-of-factly, the Spanish blogosphere is a small world.

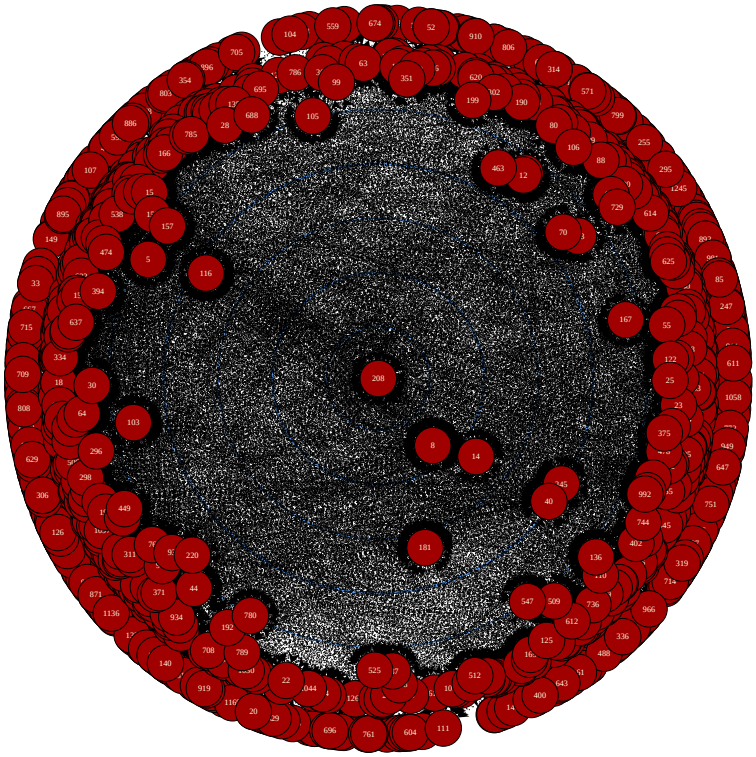

Figure 3: *Betweenness* plot for the Spanish blogosphere. Each blog is situated on a filled circle according to its betweenness measure. The “most betweened”, that is, the one with that particular value highest, is placed on the center.

5 Conclusions and Future Work

This paper shows the still immature state of a community, the Spanish speaking blogosphere, which is growing very quickly. Its immaturity is shown by the fact that it has not reached yet the state where incoming links follow a power law, but it is probably in the good path, since it is already well connected, and shows "small world" features. This fact is not connected with the power law distribution, as we had initially expected; probably it only depends on size.

The *blogómetro* is an ongoing project, and its measures will be periodically taken to show the state of the Spanish blogosphere, and measure its evolution. We will try to improve the software in several ways, including public access to the database using web interfaces, addition of self–discovery features, and improvement of other technical details (detecting links that are equal, even if they look different, among others). Until now, we have concentrated on having a tool for studying the blogosphere and helping others to discover it. Maybe we should do more work on the audience aspects, to do the tool known and useful for others. We expect that by this time, next year (or the next-to-next) we will be able to show a power law.

We have the feeling that Spanish bloggers do not link so often and frequently as they should and, maybe, we are loosing topics because no links are provided. We are thinking about trying to measure words or phrases, in order to detect topics without links in a similar way to the recently implemented 'Word Burst' of DayPop (`http://www.daypop.com/burst/`) or 'Memeufacture' (`http://memeufacture.com/`). A more refined work, separating links to blogs items from links to other general media will be the subject of our research. Another phenomenon that we would like to measure and detect would be what we could call 'background histories'; that is, histories that appear in the blogosphere and are linked slowly during long periods of time accumulating an important number of links, but that do not appear in daily rankings because of this slowness: for sure some of them will be interesting topics for reading.

Other projects we intend to undertake in the future include: cluster formation in the blogosphere, interactive visualization, and, if data is available, take measures on other blog communities (such as, for instance, the Portuguese-speaking, which is probably very similar to ours).

Finally, it would be interesting to compare these results with the ones of the global blogosphere, to detect similarities and differences.