Characterization of National Web Domains

RICARDO BAEZA-YATES[1]
Yahoo! Research
CARLOS CASTILLO[2]
Cátedra Telefónica
Universitat Pompeu Fabra
and
EFTHIMIS N. EFTHIMIADIS
Information School
University of Washington

During the last few years, several studies on the characterization of the public Web space of various national domains have been published. The pages of a country are an interesting set for studying the characteristics of the Web, because at the same time these are diverse (as they are written by several authors) and yet rather similar (as they share a common geographical, historical and cultural context).

This paper discusses the methodologies used for presenting the results of Web characterization studies, including the granularity at which different aspects are presented, and a separation of concerns between contents, links, and technologies. Based on this, we present a side-by-side comparison of the results of 12 Web characterization studies comprising over 120 million pages from 24 countries. The comparison unveils similarities and differences between the collections, and sheds light on how certain results of a single Web characterization study on a sample may be valid in the context of the full Web.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.5 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*

General Terms: Measurement

Additional Key Words and Phrases: Web Characterization, Web Measurement

1. INTRODUCTION

The World Wide Web Consortium[3] had a Web characterization activity from 1994 to 1999. The last summary of Web characterization studies of that working group was published by Pitkow [1999], and included both the characteristics of the Web pages and sites as well as the characteristics of the Web traffic generated by users.

One of the main difficulties involved in any attempt of Web characterization is how to obtain a representative sample. We have observed three types of sampling in the literature: complete crawls of a single Web site, random samples from the whole Web, and large samples from specific communities.

Complete crawls of a single Web site produce results that are biased by the choice of the Web site of study, typically of academic nature because it is easier to get

[1]This work was partially funded by ICREA and Universitat Pompeu Fabra
[2]Currently at Università di Roma "La Sapienza"
[3]Home page at <http://www.w3c.org/>.

access to the data. In this case, the whole set of pages always belongs to the same organization and therefore has not enough diversity to be representative. Random samples from the complete Web, on the other hand, include pages from different authors and organizations, but due to the large scale of the Web, are much less complete and usually they are not uniform.

Large samples from specific communities, such as national domains, have a good balance between diversity and completeness. They include pages that share a common geographical, historical and cultural context but are written by diverse authors in different organizations. Web domains also have a moderate size that allows good accuracy in the results; because of this, they have attracted the attention of several researchers.

Different methodologies have been applied to characterize several national Web domains, but to the best of our knowledge, in the last five years there is no study comparing their findings. In this paper, we:

—survey several reports on national Web domains;

—discuss a methodology to present these kinds of reports,

—present a side-by-side comparison of their results, and

—relates the results to socio-economic factors.

Besides surveying published results, we also summarize the characteristics of some collections that have not been reported in English (Brazil, Chile and Spain), have limited circulation (South Korea) or that have only produced data, but no analysis so far (Indochina, Italy and United Kingdom).

The rest of this paper is organized as follows: Section 2 introduces a methodology for presenting the results and summarizes general characteristics of the collections that are being studied. The next four sections compare the findings of the Web characterization studies according to contents and metadata (Section 3), links (Section 4), and technological aspects (Section 5). In Section 6 we compare the results of Web characterization studies with socio-economic factors. Finally, Section 7 presents our conclusions.

2. METHODOLOGY

This section explains how the results are presented in Web characterization studies, introduces the datasets used in this paper, and presents some general statistical properties of the Web.

2.1 Presentation of characterization results

The Web can be analyzed at several levels of granularity [Björneborn and Ingwersen 2004]. From a single byte through multi-byte sequences representing characters, to top-level domains and finally the entire corpus of digital information available in the Global Web, there is a series of possible levels of description. In Figure 1, we depict the ones that are most commonly found in Web characterization studies.

Three of these levels receive more attention by researchers: pages, sites and domains. A *Web page* is the unit of content that is described by the HTTP protocol, and is also the basic unit for showing results in Web search engines. A *Web site*, e.g., `www.mat.unb.br` is typically an ensemble of pages in the same topic, and is

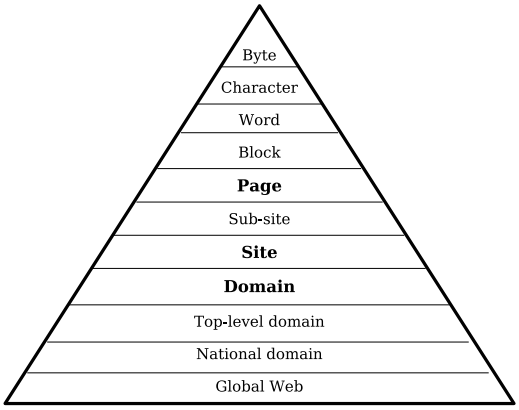

Fig. 1. Levels of granularity when describing a collection from the Web.

the basic unit used by most Web directories. A *domain* belongs to one organization and sometimes includes several Web sites, e.g., national domain such as `.br` or organizational domains, such as `unb.br`. Further, web pages can be divided into "first level" pages that are the homepages, and "second level" pages that are directly reachable from the home page. Throughout the paper we use the terms *Web sites* and *hosts* interchangeably.

Granularity is not the only axis for classifying the information that can be obtained from the Web. We can also divide the characteristics of the Web according to their type:

Content. This includes the actual contents of the objects, and their metadata or properties.

Links. This includes the relationships between objects, typically in the form of hyper-references.

Technologies. This includes the file formats, protocols and programming languages used for building the Web.

The two axes: granularity and type, can be combined to classify the properties appearing in the Web characterization studies presented in this paper, as shown in Table I. In our opinion, both axes should be used for presenting the results: in this paper, we order the results first by type, then by granularity.

2.2 Collections studied

We analyzed a total of 10 national domains plus the multi-national Web spaces of African and Indochinese Web sites. Below we list the sources used:

Africa. [Boldi et al. 2002; 2004] studied the domains of several African countries. The countries were: Egypt (`EG`), Libya (`LY`), Morocco (`MA`), Mozambique (`MZ`), Namibia (`NA`), Senegal (`SN`), South Africa (`ZA`), Tunisia (`TN`) and Zimbabwe (`ZW`).

Austria. (`AT` domain) [Rauber et al. 2002] presents an analysis of the Austrian Web using a data warehousing approach.

Table I. A list of properties that have been included in Web characterization studies, classified by granularity (G. = page, site, domain) and type (T. = content, links, technology).

G. \ T.	Contents	Links	Technologies
Pages	Word frequencies; Language; Text size; Page size; Age; Duplicates; HTML meta-tags	Indegree; Outdegree; PageRank; Hub score; Authority score	URLs; Response codes; Media and document formats; Dynamic pages; Scripting languages; HTML version
Sites	Sum of text sizes; Sum of page sizes	Indegree and outdegree in the hostgraph; Internal links; Distribution of strongly-connected components; Web structure	Types of Web sites with one indexable page; Technologies for dynamic pages
Domains	Sites per second and third level domain	Most referenced domains; Coverage of domain references	Software used as Web server; Prevalence of DNS wildcarding; Distribution of IP per address

Brazil. (`BR` domain) [Veloso et al. 2000] and [Modesto et al. 2005] are two analysis of this large country's Web using CobWeb [da Silva et al. 1999] and WIRE [Baeza-Yates and Castillo 2002], respectively.

Chile. (`CL` domain) [Baeza-Yates et al. 2000; 2003; 2005] have been carrying several analysis of the Chilean Web since the year 2000, using the WIRE crawler and data from the TodoCL[4] search engine, that also uses CobWeb.

Greece. (`GR` domain) [Efthimiadis and Castillo 2004] is a preliminary study using WIRE.

Indochina. is a collection obtained by the Laboratory of Web Algorithmics[5] in 2004. The countries included are Cambodia (`KH`), Laos (`LA`), Myanmar (`MM`), Thailand (`TH`) and Vietnam (`VN`). These collection was crawled using Ubicrawler [Boldi et al. 2004].

Italy. (`IT` domain) is a large collection obtained by the Laboratory of Web Algorithmics in 2004 with Ubicrawler.

Portugal. (`PT` domain) [Gomes and Silva 2005] is a study using the Viúva Negra crawler from the Tumba[6] search engine.

South Korea. (`KR` domain) [Baeza-Yates and Lalanne 2004] is a study using WIRE.

Spain. (`ES` domain) [Baeza-Yates et al. 2006] is a study using a modified version of CobWeb. We also use data from [Alonso et al. 2003], an in-depth study on 27 specific Web sites.

Thailand. (`TH` domain) [Sanguanpong et al. 2000] is a study using NontriSpider from the NontriSearch search engine [Sanguanpong and Warangrit 1998]. This is a more in-depth study that the corresponding part of the collection in the Indochinese sample.

[4]TodoCL search engine, <`http://www.todocl.cl/`>.
[5]Laboratory of Web Algorithmics, *Dipartimento di Scienze dell'Informazione*, *Universitá degli studi di Milano*, <`http://law.dsi.unimi.it/`>.
[6]Tumba search engine's crawler, <`http://www.tumba.pt/english/crawler.html`>.

United Kingdom. (UK domain) is a large collection obtained by the Laboratory of Web Algorithmics in 2002 using Ubicrawler.

By observing the number of available hosts and the downloaded pages in each collection, we consider that most of them have a high coverage, of at least the home pages, that is, the first levels of the pages in their Web sites. The collections of Brazil and the United Kingdom are smaller samples in comparison with the others, but as we will see in the rest of the study, their sizes are large enough to show results that are consistent with the others.

For comparison, we also used information obtained from samples of the global Web [Broder et al. 2000; Dill et al. 2002], and from a study on Web graph compression [Suel and Yuan 2001].

Most of the national domain studies used different crawling software and hardware, but many of them are either with WIRE or Ubicrawler. In most cases, we observe that the similarities and differences are crawler independent. Besides that, the main factors that affect the obtained statistics are the following:

National domain boundaries. In some cases, the assigned top-level domain name is the most used for pages in the country; for instance, most of the Brazilian Web sites use the BR domain. In other cases Web sites are scattered across several domains as in the Spanish Web (which had a more restrictive policy of registrations under the country-code ES until 2005).

A possible choice for defining the Web of a country is considering all Web sites that are registered at a domain inside the assigned country-code, or that are hosted at an IP that belongs to a segment assigned to that country. In most of our own studies we use the union of both cases.

Crawling depth and coverage. Web sites are potentially infinite [Brin et al. 1998; Heydon and Najork 1999; Baeza-Yates and Castillo 2004; Eiron et al. 2004]. For example, dynamic pages can create groups of infinitely many pages, for instance, imagine a calendar on which you can click 'next year' forever. So, it is common to enforce some type of limit in the *depth* at which the crawl stops. Even this limit might not be enough for Web sites generating automatically many links, so also a per-site page limit is used by several crawlers.

When Web sites outside the main country domain are explored, it is typical to use some heuristic to avoid downloading too many unrelated pages, such as reducing the exploration depth or the number of pages downloaded, using trigger keywords or enforcing a lower limit on the number of links received by a page outside the country domain before crawling it.

The results presented in the paper are for the Web as collected by the crawlers. That is, for the Web before removal of spam. It should be noted that the crawlers in some cases included some obvious spam filters, but no post-filtering was done.

Static and dynamic pages. The handling of dynamic pages varies among crawlers. Some crawlers ignore them completely, others follow them but discard all the characters of the URL that follow the question mark (removing all the parameters); others try to remove parameters related to user tracking or session-ids (to reduce the presence of duplicates), and others simply follow links to dynamic pages without changing them.

Table II summarizes the characteristics of the studied collections. The number of unique host names was measured by the Internet Systems Consortium[7] in July 2005.

Table II. Characteristics of the studied collections. The host count is an estimation from the Internet System Consortium (2005). The collected pages is the number of pages that were downloaded and included in the collection. The maximum depth is sometimes different for static and dynamic pages.

Collection	Year	Available hosts [mill] (rank)	Collected pages [mill]	Limits	
				Depth	Pages per site
Africa	2002	0.4 (39^{th})	2.0	n/a	n/a
Austria	2002	1.6 (23^{th})	11.0	n/a	n/a
Brazil	2005	3.9 (11^{th})	4.7	5	10,000
Chile	2004	0.3 (42^{th})	3.3	5-15	5,000
Greece	2004	0.3 (40^{th})	3.7	5-15	25,000
Indochina	2004	0.5 (38^{th})	7.4	n/a	10,000
Italy	2004	9.3 (4^{th})	41.3	8	10,000
South Korea	2004	0.2 (47^{th})	8.9	5-15	5,000
Portugal	2003	0.6 (37^{th})	3.2	6	8,000
Spain	2004	1.3 (25^{th})	16.2	∞	400
Thailand	2000	0.5 (38^{th})	0.7	n/a	n/a
United Kingdom	2002	4.4 (10^{th})	18.5	n/a	n/a

2.3 Zipf's law and scale-free networks

The graph representing the connections between Web pages has a scale-free topology. Scale-free networks, as opposed to random networks, are characterized by an uneven distribution of links, and the distribution of the number of links to a page p follows a power law:

$$Pr(p \text{ has } k \text{ links}) \propto k^{-\theta}$$

We find this distribution on the Web in almost every aspect. It is the same distribution that was found by economist Vilfredo Pareto in 1896 for the distribution of wealth in large populations, i.e., 80% of the wealth is owned by 20% of the population. It is also the same distribution found by George Kingsley Zipf in 1932 for the frequency of words in texts, and that later turned out to be applicable to several domains [Zipf 1949], called by him the law of *minimal* or *least effort*.

One phenomenon that has appeared before in our own studies, and now is completely clear, is the smaller power law exponent at the beginning of several of the measures presented. In fact, this happens for file sizes up to 25Kb, pages per site up to 15 to 30, pages per domain up to 10 (except South Korea), number of out-links in a page up to 10 to 40, and average number of internal links per site up to 15 to 30, where a range is given to show the variability for different countries. We argue that this is due to another empirical power law that we call *maximal shame*[8] which forces people to work a bit more than the minimum, until they feel well about their

[7] Internet systems consortium's domain survey, <http://www.isc.org/ds/>

[8] Could also be called *minimal pride* but it counter reacts to minimal effort so we prefer the former.

work. Notice that this maximal shame can be for an individual or for a group (for example, in the case of a Web site).

3. CONTENTS

This section and the following two compare Web characterization results; as the way of reporting the data differs, for each observed characteristic we only include *comparable* data from the subset of countries from which it is available.

3.1 Languages

In the year 2000, it was estimated that around 70% [Grefenstette and Nioche 2000] of the pages were written in English, and that the numbers of words available in other languages was growing faster than the number of words in English. On January 2003, Google Zeitgeist[9] showed that around 50% of the queries to Google were using English, down from around 60% in 2001.

For language detection on the Web, two main techniques are applied: lists of stopwords in several languages are used, such as, in the studies of Chile and Brazil [Baeza-Yates and Castillo 2005; Modesto et al. 2005], and naïve Bayes over n-grams in the studies of Africa, Portugal and Spain [Boldi et al. 2002; Gomes and Silva 2005; Baeza-Yates et al. 2006]. The method used for the Web of Thailand was not specified in their paper [Sanguanpong et al. 2000]. In general, dictionary-based language detection works better with large texts and in the Web there are many pages that are very short; in these studies, when using list of stopwords many pages are not classified in any language, while n-grams-based techniques [Cavnar and Trenkle 1994] are able to classify accurately a larger subset of the collection.

The distribution of pages in English versus the pages in the local languages and other languages is shown in Figure 2. We also include Spanish and Portuguese as each of them is important in two of the studied samples. Note that English is a local language for some African countries, as well as Portuguese and other languages. We believe that Thailand has many pages in English (65%), as opposed to Thai (35%), because it is a major touristic destination and English is also the secondary official language of the elite class[10] in all cases there are large differences in the fraction of non-English languages across countries.

For example, in Portugal there are two official languages, Portuguese (official) and Mirandese (official - but locally used). In Spain the distribution of local languages is Castilian Spanish 52%, Catalan 8%, Galician 1%, and Basque 1%. Castilian is the official language, and the other languages are co-official and used regionally.

¿From Figure 2 we see that English ranges from about 8% in Chile to 65% in Thailand and 75% in Africa. Chile and Brazil have very similar ratios of English to their national languages, that is, 8% and 11% of English to about 90% and 88%, respectively. Similar patterns are observed in Portugal and Spain, where English is 18% and 30%, while Portuguese and Spanish are about 70% and 55%, respectively. One possible explanation of the low percentage of English language pages in Chile and Brazil might be that the English is spoken by a small percentage

[9]Online: <`http://www.google.com/press/zeitgeist.html`>, verified November 2005.
[10]U.S. Central Intelligence Agency, The World Factbook. Online: <`http://www.cia.gov/cia/publications/factbook/`>, verified November 2005.

of the population and that tourism is relatively low in both countries. For example, in 2002, Chile had 1.4 million tourists, Brazil had 3.8 million, and in contrast Thailand had 11 million.[11]

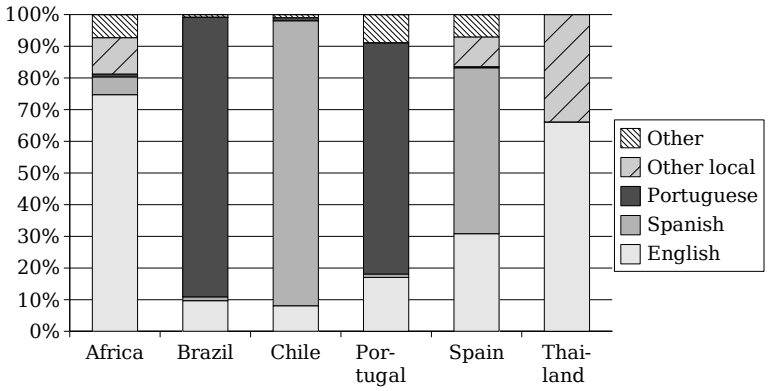

Fig. 2. Distribution of the number of pages in different languages.

3.2 Page size

The average file size of HTML pages were 13 KB for the African sample, 24 KB for Brazil, 21 KB for Chile, 22 KB for Greece, 14 KB for South Korea, 21 KB for Portugal and 10 KB for Thailand. The distribution of page sizes is very skewed, as shown in Figure 3, and can be modeled by a double-pareto distribution [Mitzenmacher 2003].

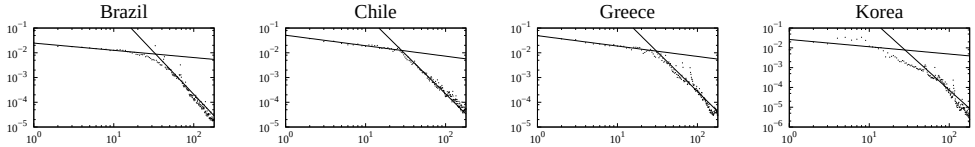

Fig. 3. Histograms of file sizes. The x-axis is the file size in Kilobytes and the y-axis the relative frequency. The average size $\bar{x}$ and two parameters for the power-law are given: θ_1 for smaller sizes, and θ_2 for larger sizes.

We observed two different exponents, one for smaller pages (less than 20 KB) and another for larger pages. The observed power-law parameters (θ_1 and θ_2) vary among samples, and are roughly 0.4 for the smaller sizes and 3.5 for the larger sizes. In a previous study [Arlitt et al. 1999], a power-law was also observed, and the exponent for the larger sizes was 1.5. The difference may be due to two reasons.

[11] World Bank, World Development Indicators database (WDI). Online: <`http://devdata.worldbank.org/wdi2005/`>, verified November 2005.

First, there are differences in the usage of HTML coding for writing Web pages; nowadays, pages tend to be more complex. Second, their study used data from traces from Web page users, who probably do not tolerate large page sizes as a Web crawler does. For a discussion on models for Web page sizes, see [Downey 2001; Mitzenmacher 2003; Baeza-Yates and Navarro 2004].

3.3 Page age

Page age information was obtained by reading the `last-modified` header in the HTTP responses that contained this information. Though the header information is not fully reliable, it is the best available. Days or months are grouped together, so little variation does not matter. The crawler ignores dates that occur in the future, as well as dates prior to 1990. The distribution of the age of pages exhibits an exponential distribution, which can be explained by modeling page changes as a Poisson process [Brewington et al. 2000]. Figure 4 shows that the data is consistent with an exponential distribution, except for the South Korean sample that shows more pages than expected having less than one year of age.

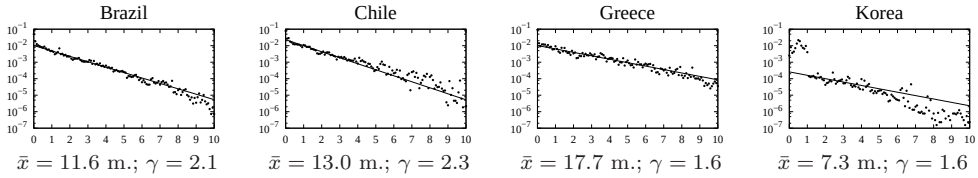

Fig. 4. Histograms of page ages. The x-axis is the page age in years and the y-axis the relative frequency. The average age $\bar{x}$ (in months for clarity) and the parameter of a fitted exponential distribution γ are given. Note that unlike all other histograms in this paper, only the y-axis is in logarithmic scale.

3.4 Pages per site

The distribution of Web pages onto sites follows a power-law with parameters between 1.3 and 1.7 (except for the South Korean sample) as shown in Figure 5. For large samples of the whole Web, a power-law has also been observed, with parameter between 1.78 and 1.91 [Huberman and Adamic 1999]. Note that in the different studies, different limits for the number of pages per Web site were used, as shown in Table II.

The average number of Web pages per site varies widely across collections: Brazil has 66, Chile 58, Indochina 549, Italy 410, Greece 150, South Korea 224, Spain 52 and United Kingdom 248 (for Portugal, the exact average is not specified but it is said to be below 100 pages per Web site). For calculating these averages, we do not take into account single-page Web sites, which are analyzed in the next section.

In the case of Indochina, Italy and the U.K., by manual inspection we observed that there is a significant amount of pages including a session-id, or links to some of the dynamic pages that CobWeb and WIRE discard by using patterns. For instance, there are many links to applications such as “post” or “edit postings” in Blogs, that can be avoided during crawling by filtering those URLs using regular

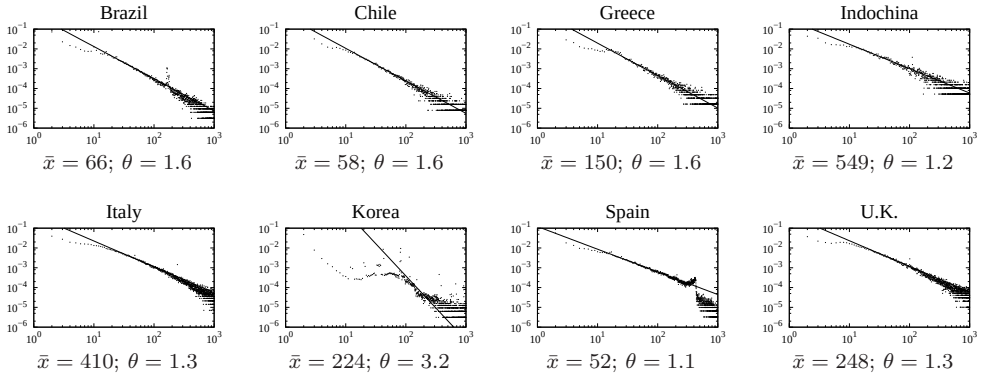

Fig. 5. Histograms of the number of pages per Web site, with relative frequencies in the y-axis. The average number of pages per Web site $\bar{x}$ and the power-law parameter θ are given.

expressions. Depending on the crawl the number of links may be reduced between 20-40%.

In the case of the South Korean sample, as shown in Figure 5, for Web sites with less than 50 pages the distribution is not a power-law. There are a large number of Web sites with very few pages, mostly built for spamming search engines. We observe that these are sites in the same domain, each hostname with a single or very few documents inside. (For a detailed account of spam see work by Fetterly, Manasse, and Najork [2004] and Gyöngyi and Garcia-Molina [2005]). The differences among the distributions disappear if the page sizes are considered, and the power-law exponents are closer to each other, as shown in Figure 6 (as in the other graphs, we include here only the collections from which we have data about page sizes).

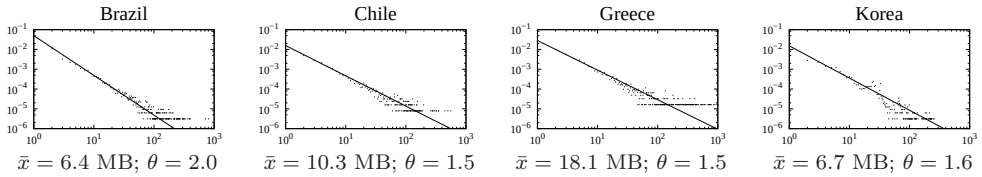

Fig. 6. Histograms of the total size of Web pages per site. The x-axis is the sum of the sizes of the pages in Megabytes, and the y-axis is the relative frequency. The average size $\bar{x}$ and the parameter for the power-law θ are given.

3.5 Sites and pages per domain

In the studied collections domains, e.g., `xxx.gr`, have on average between 1.1 and 2.5 sites per domain, and over 95% of the domains have only a single Web site, e.g., `yyy.xxx.gr`. In the case of South Korea, the average is much larger (26.1) due to the presence of several spam Web sites. In this collection, over 20% of the domains

have more than 10 Web sites, which is quite different than other countries. The distribution of sites into domains is shown in Figure 7.

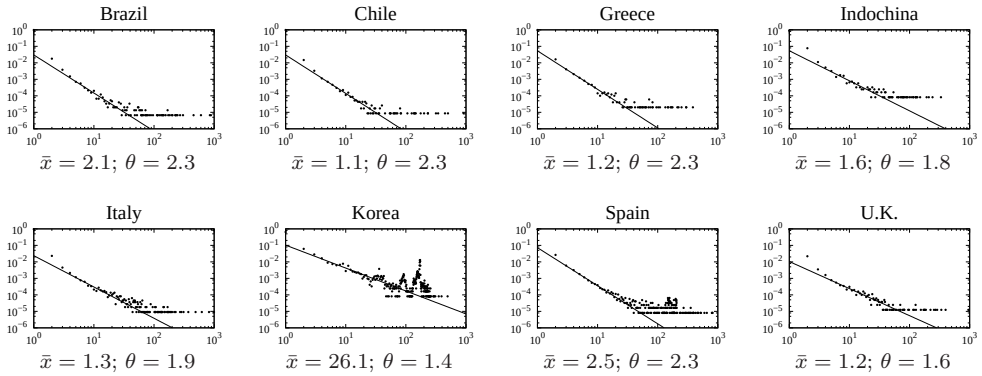

Fig. 7. Histograms of the number of sites per domain, with the relative frequency in the y-axis. The average $\bar{x}$ and the power-law exponent θ are given.

Another anomaly can also be observed in the collection of pages from Spain, and it is also due to groups of spam Web sites. The differences between the collections tend to be smaller when the number of pages per domain is analyzed, as is shown in Figure 8.

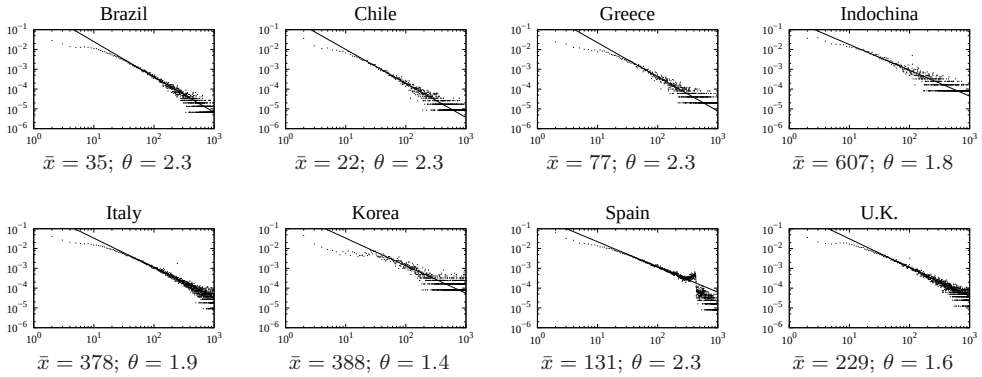

Fig. 8. Histograms of the number of pages per domain, with the relative frequency in the y-axis. The average $\bar{x}$ and the power-law exponent θ are given.

3.6 Second-level domains

In the United Kingdom and several countries of Indochina, the country code cannot be used directly, and only third level domains can be registered (for instance, under `.co.uk` or `.ac.th`). In other countries such as Brazil, the policy is hybrid, and educational and governmental entities can apply for a domain directly under `BR`, for

example, `www.ufmg.br` while companies and individuals have to use a third-level domain, for example, `sbc.org.br` or `petrobras.com.br`. Finally, there are countries, such as Spain or Greece, where there is no policy regulating the use of second level domains. Consequently, domains are registered directly under the country domain, e.g., `.es` or `.gr`, which makes difficult the identification of the subdomains. Figure 9 shows the distribution of second-level domains in those countries where we were able to differentiate between subdomains. To be able to compare data, we have grouped them in Commercial (`COM`, `CO`, `LTD` and `PLC`), Organization (`ORG`, `OR` and `ART`), Educational (`EDU`, `AC` and `SCH`), Government (`GOV`, `GO`, `NHS` and `POLICE`), Individuals (`PE`, `ME`, `ADV` and `IN`) and Networks (`NET` and `NE`).

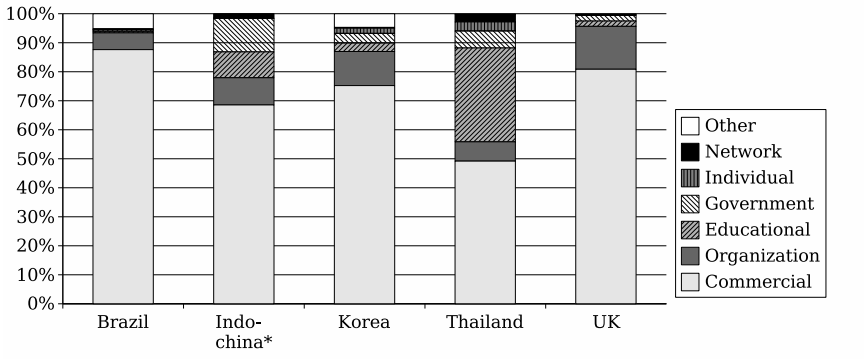

Fig. 9. Distribution of the number of domains per second-level domain, for the samples in which it is not possible to register a second-level domain directly. (*) includes only Cambodia, Myanmar and Vietnam.

Domains for commercial usage comprise on average 70% of the registrations, ranging from 50% in Thailand to 82% in the UK and 88% in Brazil. This is followed by educational and government institutions, with roughly 10% each on average.

4. LINKS

In this section, we study the Web as a directed graph, in which each page is a node, and each hyper-link is an edge.

4.1 Degree

The distribution of in-links is shown in Figure 10, which is consistent with a power-law distribution.

Indegree links range from 8.3 pages for Chile to 26.2 pages for Indochina and 27.9 for Italy, with a median of 14.9 in-links. The exponent ranges from $\theta = 1.6$ for Indochina to $\theta = 2.1$ for Spain, with a median of $\theta = 1.9$ for the eight studies reported.

In samples of the global Web, it has been observed an average of 7.2 out-links per page [Kleinberg et al. 1999], and the distribution of out-links is also very skewed, as shown in Figure 11. The distribution of average outdegree links range from 3.6

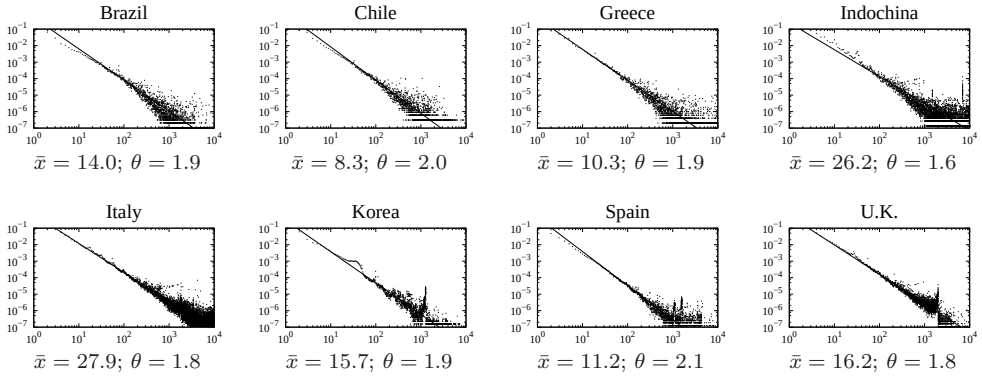

Fig. 10. Histograms of the indegree of Web pages. The number of different pages pointing to a page is in the x-axis, and the relative frequency in the y-axis. The average indegree $\bar{x}$ (counting only pages with in-links) and the power-law exponent θ are given.

pages for Spain to 31.8 pages for Indochina and 31.9 pages for Italy, with a median of 18.8 pages.

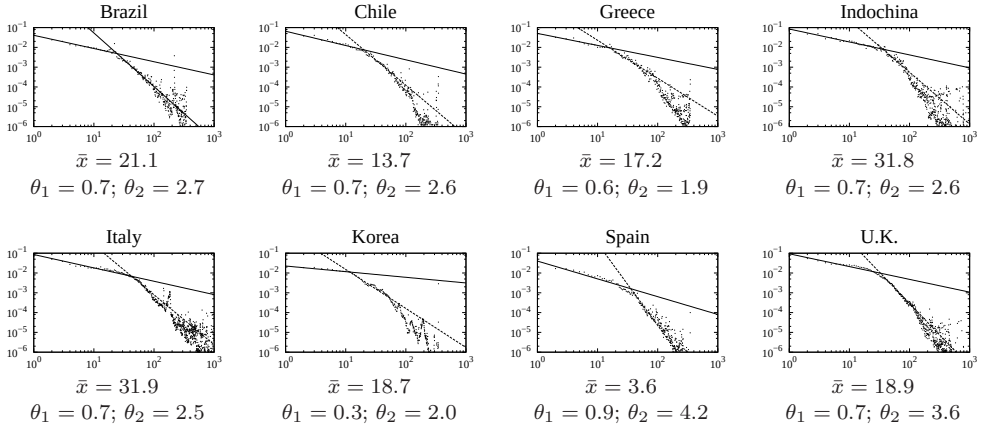

Fig. 11. Histograms of the outdegree of Web pages. The number of different pages pointed by a page is in the x-axis, and the relative frequency in the y-axis. The average outdegree $\bar{x}$ (for pages with at least one out-link) and two parameters for the power-law are given: θ_1 for pages with few out-links ($\leq$ 20–30) and θ_2 for pages with more out-links.

When examining the distribution of outdegree, we found two different curves: one for smaller outdegrees –less than 20 to 30 out-links– and another one for larger outdegrees. They both show a power-law distribution and we estimated the exponents for both parts separately. The corresponding exponent values for θ_1 and θ_2 range from $\theta_1 = 0.3$ for South Korea to $\theta_1 = 0.9$ for Spain with a median of $\theta_1 = 0.7$, and $\theta_2 = 1.9$ for Greece to $\theta_2 = 4.2$ for South Korea with a median of $\theta_2 = 2.6$.

The fact that for smaller outdegrees there is a power-law distribution can be explained by the same argument that Zipf used: because Web page authors make a minimal effort. However, pages with more out-links are typically generated by content management systems or Web page generators that are not bound by effort constraints, as making a program that generates 100 links is as easy as making a program that generates 1000 links. Consistently with this, we observe that there are more deviations from the power-law in the right part of the histograms.

Finally, when looking at the averages of both indegree and outdegree links we observe an increased number of links from those reported by earlier studies. An explanation for the increase is two-fold. Over the past seven years that cover the AltaVista study of Broder *et al.* [2000] people are authoring more elaborate Web sites that have more links. We assert that this is because Web sites authors have matured and they also use more links in hopes of increasing their PageRank score.

4.2 Ranking

One of the main algorithms for link-based ranking of Web pages is PageRank [Page et al. 1998]. We calculated the PageRank distribution for several collections and found a power-law in the distribution of the obtained scores, with exponents between 1.8 and 2.0. In theory, the PageRank exponent should be similar to the indegree exponent [Pandurangan et al. 2002], and this is indeed the case. The distribution of PageRank values can be seen in Figure 12.

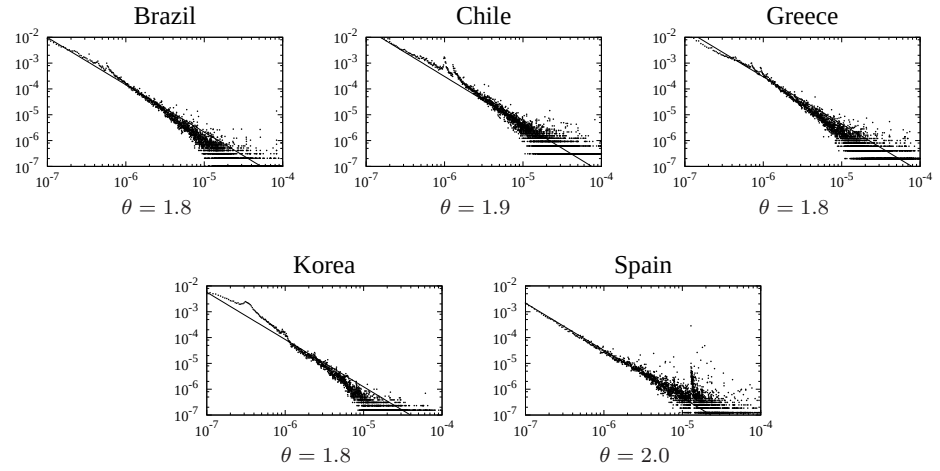

Fig. 12. Histograms of the PageRank of Web pages, with relative frequencies in the y-axis. The parameter θ is obtained by fitting a power-law to the data.

Finally, in some collections we also calculated a static version of the HITS scores [Kleinberg 1999], counting only external links and calculating the scores in the whole graph, instead of only on a set of pages. The tail of the distribution of authority-score also follows a power law. In the case of hub-score, it is difficult to assert that the data follows a power-law because the frequencies seems to be much

more disperse, as can be seen in the top row of Figure 13. The parameters for the authority score (in the bottom row of the figure) and for the PageRank are the same up to two decimal points, but both variables are not correlated.

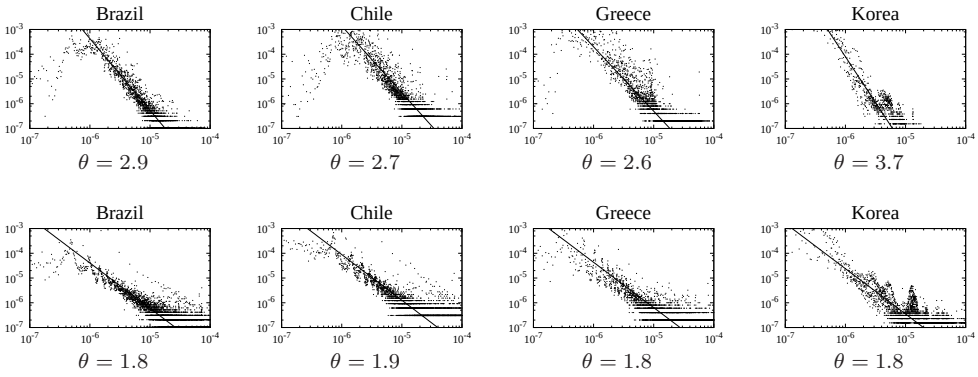

Fig. 13. Histograms of the static hub-scores and authority-scores of Web pages in several collections. The x-axis, on the top row, shows the hubs score, and on the bottom row the authority scores. The y-axis gives the relative frequency. All graphs are in the same logarithmic scale.

A summary of the power-law exponents found in this section is shown in Table III.

Table III. Summary of power-law exponents in the graph of links between pages. For the outdegree, there are two exponents: one for pages with roughly less than 20 out-links, and one for pages with more out-links.

Collection	In-degree	Outdegree		Page-Rank	HITS	
		Small	Large		Hubs	Auth.
Africa	1.92	n/a	n/a	n/a	n/a	n/a
Brazil	1.89	0.67	2.71	1.83	2.9	1.83
Chile	2.01	0.72	2.56	1.85	2.7	1.85
Greece	1.88	0.61	1.92	1.83	2.6	1.83
Indochina	1.63	0.66	2.62	n/a	n/a	n/a
Italy	1.76	0.68	2.52	n/a	n/a	n/a
South Korea	1.90	0.29	1.97	1.83	3.7	1.83
Spain	2.07	0.86	4.15	1.96	n/a	n/a
United Kingdom	1.77	0.65	3.61	n/a	n/a	n/a
[Broder et al. 2000]	2.1	n/a	2.7	n/a	n/a	n/a
[Dill et al. 2002]	2.1	n/a	2.2	n/a	n/a	n/a
[Pandurangan et al. 2002]	n/a	n/a	n/a	2.1	n/a	n/a
[Kleinberg et al. 1999]	≈ 2	n/a	n/a	n/a	n/a	n/a

4.3 Hostgraph

We studied the hostgraph [Bharat et al. 2001; Dill et al. 2002], that is, the graph created by changing all the nodes representing Web pages in the same Web site by a single node representing the Web site. The hostgraph is a graph in which there is a node for each Web site, and two nodes A and B are connected iff there is at least one link on site A pointing to a page in site B.

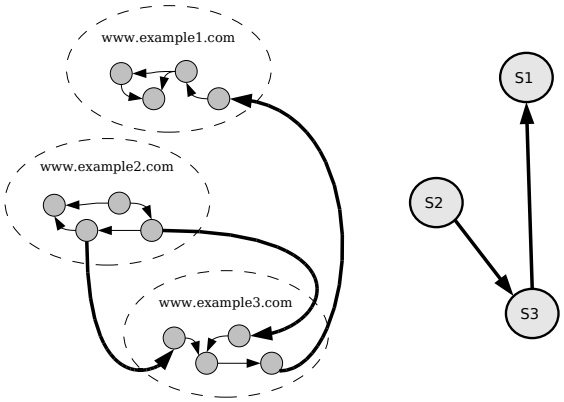

Fig. 14. The Web graph (left) can be transformed into a hostgraph (right). The hostgraph is a representation of the links between different Web sites, with multiple links merged.

The average indegree per Web site (average number of different Web sites inside the same country linking to a given Web site) was 3.5 for Brazil, 1.2 for Chile, 1.6 for Greece, 37.0 for South Korea and 1.5 for Spain. The distribution of indegree is shown in Figure 15.

By manual inspection we observed that in Brazil and specially in South Korea, there is a significant use –and abuse– of DNS wildcarding. DNS wildcarding [Barr 1996] is a way of configuring DNS servers so they reply with the same IP address no matter which host name is used in a DNS query. For instance, if `example.com` is using DNS wildcarding, then *`string`*`.example.com` always points to the same IP address no matter which *`string`* is used. This technique aims at increasing the ranking of a group of pages on search engine's results, by including several keywords in the host part of the URLs. We have observed that almost all the domains that use DNS wildcarding use it for spamming, with the exception of domains used for providing aliases for Web hosting.

The average outdegree per Web site (average number of different Web sites inside the same country linked by a given Web site) was 2.2 for Brazil, 2.4 for Chile, 4.8 for Greece, 16.5 for South Korea and 11.2 for Spain. The distribution of outdegree is shown in Figure 16.

We also measured the number of internal links, that is, links going to pages inside the same Web site. We normalized this by the number of pages in each Web site, to be able to compare values. In the case of Brazil, Chile and Greece, we observed a combination of two power-law distributions: one for Web sites with up to 10 internal links per Web page on average, and one for Web sites with more

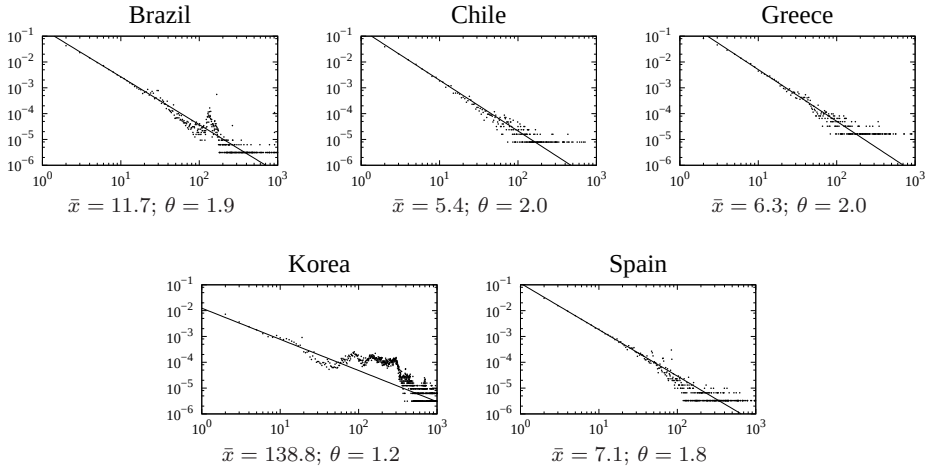

Fig. 15. Histograms of the indegree of Web sites. The x-axis is the number of different Web sites pointing to a site, and the y-axis the relative frequency. The average indegree $\bar{x}$ for Web sites with at least one in-link and the parameter θ of the power-law are given.

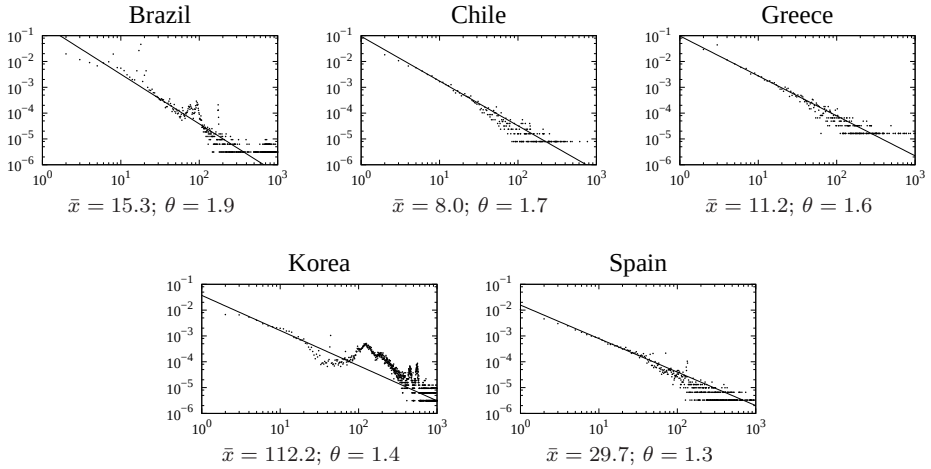

Fig. 16. Histograms of the outdegree of Web sites. The x-axis is the number of different Web sites pointed from a site, and the y-axis the relative frequency. The average outdegree $\bar{x}$ for Web sites with at least one out-link and the parameter θ of the power-law are given.

internal links per Web page. In the case of South Korea and Spain it resembles more a power-law with a single parameter, but we include an approximation with two different parameters for all the collections for consistency. The distribution is shown in Figure 17. This is consistent with Figure 11 that showed the outdegree of Web pages.

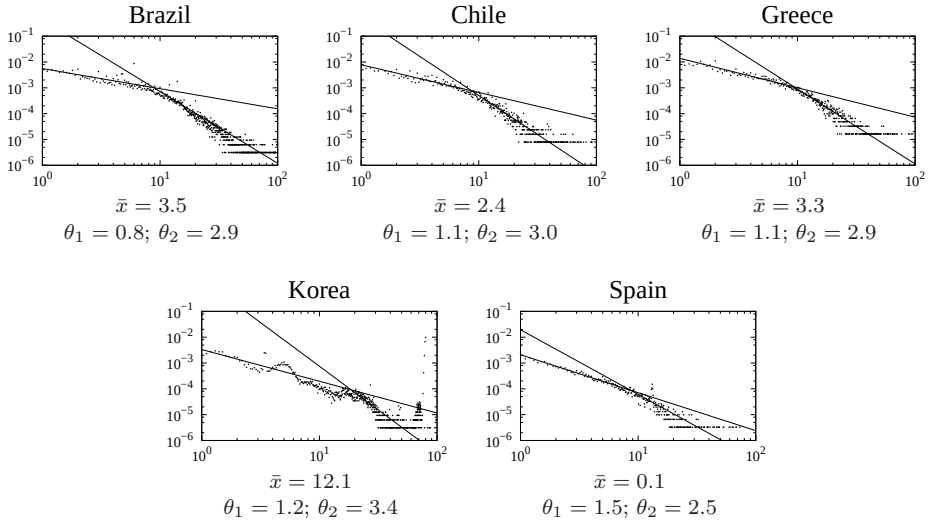

Fig. 17. Histograms of the average number of internal links. The x-axis is the number of internal links, normalized by the number of pages in each Web site, and the y-axis is the relative frequency. The average $\bar{x}$ and two parameters for the power-law fit θ_1 and θ_2 are given.

4.4 Web structure

Broder et al. [2000] proposed a characterization of the structure of the Web graph (known as the "bow-tie" model) based on the relationship of each page with the larger strongly connected component (SCC) on the graph. This induces a partition of the Web pages: the pages in the larger strongly connected component belong to the category MAIN. Starting in MAIN, if we follow links forward we find OUT, and if we follow links backwards we find IN. All of the Web pages which are part of the graph reachable from MAIN, disregarding the order of links, but that do not fit either in MAIN, IN, nor OUT are part of the components called TENDRILS and TUNNEL. A graphical depiction of these components is shown in Figure 18.

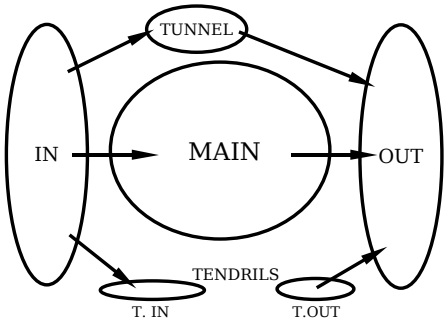

Fig. 18. Graphic depiction of the bow-tie structure of the Web. The arrows represent the flow of links.

Baeza-Yates and Castillo [2001] showed that this macroscopic structure is similar at the hostgraph level: the hostgraphs examined here are scale-free networks and have a giant strongly connected component. The distribution of the sizes of their strongly connected components is shown in Figure 19.

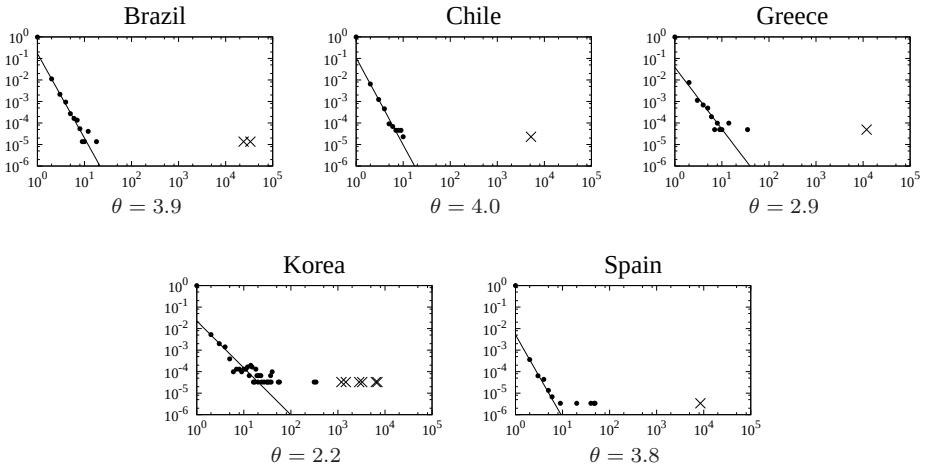

Fig. 19. Histograms of the sizes of strongly connected components (SCCs) in the hostgraph. For readability, SCCs with more than 1,000 sites are marked with a cross. The exponent θ was calculated by fitting a power-law to the smaller SCCs.

In Chile, Greece and Spain, a sole giant strongly connected component appears, having at least 2 orders of magnitude more Web sites than the following component. In the case of Brazil, there are two giant SCCs. The larger one is a "natural" one, containing Web sites from different domains. The second larger is an "artificial" one, containing only Web sites under a domain that uses DNS wildcarding to create a "link farm" (a strongly connected community of mutual links). In the case of South Korea, we detected at least 5 large link farms. Table IV summarizes the power-law exponents found for the links in the hostgraph.

Regarding the Web structure, while at the level of pages the sizes of MAIN, IN, OUT and TENDRILS are very similar [Broder et al. 2000], the distribution between sites in general gives the component called OUT a larger share, as shown in Figure 20 (a). OUT is composed of Web sites that can be reached from the giant SCC, but that do not have many links to other Web sites. This is the typical case for the Web sites of small- and medium-sized companies or organizations, which have very few out-links.

When looking at the size of the sites in each component, it is clear that Web sites in component MAIN are larger than the others, as can be seen in Figure 20 (b). In the case of the South Korean Web, a possible explanation for the MAIN component being so small is that the largest strongly-connected component in this case is not a "natural" one, by one composed of spam Web sites.

Table IV. Power-law exponents in the hostgraph. For the number of internal links per page, there are two exponents: one for Web sites with roughly less than 10 internal links per page on average, and one for Web sites with more internal links per page. SCC is the exponent in the distribution of the sizes of strongly connected components, excluding the larger one.

Collection	Hostgraph degree		Internal links per page		SCC
	In	Out	Small	Large	
Brazil	1.85	1.92	0.78	2.88	3.93
Chile	1.97	1.73	1.07	3.02	4.05
Greece	2.00	1.55	1.14	2.90	4.20
South Korea	1.21	1.36	1.23	3.38	2.37
Spain	1.80	1.30	1.47	2.50	3.84
[Broder et al. 2000]	n/a	n/a	2.5	n/a	n/a
[Bharat et al. 2001]	1.62-1.73	1.67-1.80	n/a	n/a	n/a
[Dill et al. 2002]	2.34	n/a	2.1	n/a	n/a

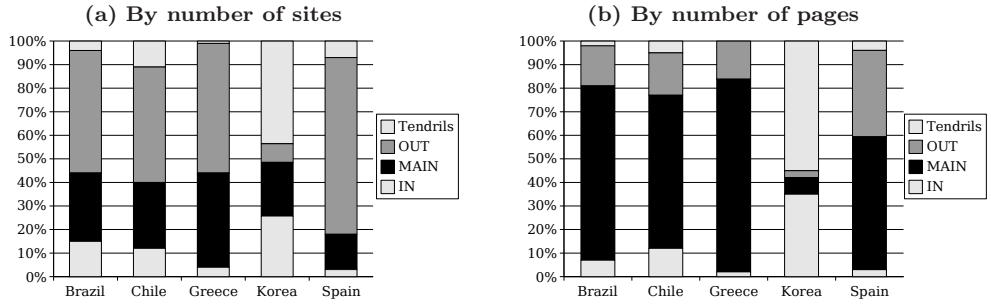

Fig. 20. Distribution of the sizes of components in the hostgraph, (a) by number of sites in each component, and (b) by the number of pages in those Web sites.

Finally, there are a large number of Web sites that are not reachable from MAIN, even if we disregard the direction of links. These isolated components, also called "islands", comprise 12% of the sites in Brazil, 46% in Chile, 9% in Greece, 56% in South Korea, and 82% in Spain. This fraction is very variable and depends on the strategy used for finding the starting URLs for the crawler, as the isolated sites can only be found if the exact site name is known in advance, so these percentages are most of the time just lower bounds. When the full list of domains registered in a country is known, or when the starting URLs are taken from the data from a large search engine, many isolated sites can be found, as was the case for Chile and Spain.

It can be argued that an isolated Web site is not so valuable. This is because the Web sites that are not connected to the rest of the Web contribute little in terms of content. In fact, their number of pages is much smaller: isolated Web sites contribute 4% of the pages in Brazil, 1% in Chile, 1% in Greece, 7% in South Korea and 28% in Spain. The percentage of isolated sites in Spain is high because:

(a) the initial set of sites is very complete, and (b) the initial sites include several `.com` sites that belong to Spanish companies, but do not link to other Spanish sites. Most of the islands in the Spanish Web are outside the `.es` domain. Notice that any national study does not take in account links coming from other countries, so a site that is an island in a country does not necessarily is an island in the whole Web.

5. TECHNOLOGIES

This section includes statistics about the technologies used for building Web sites, specially file formats and programming languages.

5.1 URL length

The distribution of the length of the URLs is important because it can help in the development of compression schemes. For instance, Suel and Yuan [2001] showed how to compress URLs of 50 bytes of length on average to around 13 bytes, by exploiting common prefixes.

Including the protocol part, the observed average length of URLs in the studied samples was 69 for Brazil, 64 for Chile, 81 for Indochina, 79 for Italy, 67 for Greece, 62 for Portugal, 67 for Spain and 76 for the United Kingdom. The distribution of the URL lengths is shown in Figure 21.

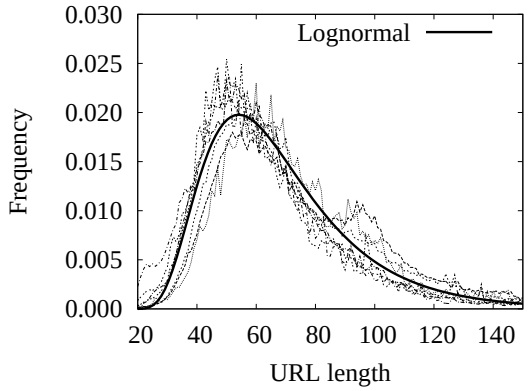

Fig. 21. Histogram of URL lengths and fit of a log-normal distribution.

We assumed a general log-normal distribution, with probability density function:

$$f(x) = \frac{e^{-((\log((x-\theta)/m)^2)/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \ ,$$

and fitted it to the data. The parameters obtained were: θ (location) $= 14.1 \pm 2.7$, m (scale) $= 49.2 \pm 3.6$ and σ (shape) $= 0.43 \pm 0.04$.

5.2 HTTP response code

As most crawlers work by recursively downloading pages and extracting links, there is no guarantee that a request for a given URL will succeed. In fact, several broken links (pages with a "`404 Not Found`" message) are found during the process. The HTTP response code from the Web servers indicates that about 80%-85% of the requests succeed, and that this fraction is similar across all domains, as shown in Figure 22. In the figure, the last column comes from data that was obtained in 1997 and published in [Pitkow 1999].

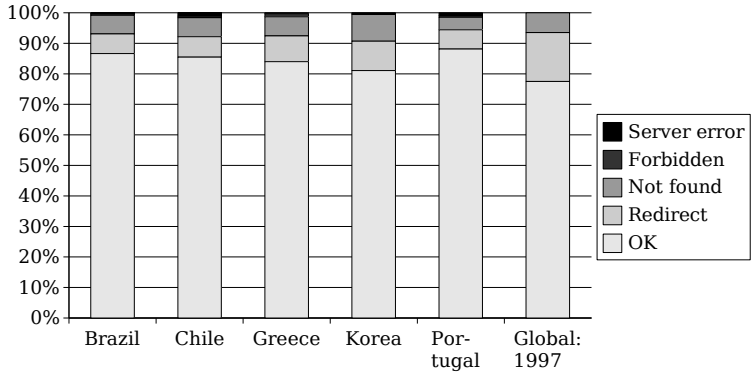

Fig. 22. Distribution of HTTP response codes.

Compared to the 1997 sample, the number of broken links ('Not found' in the figure) appears to be smaller. This may be due to the fact that only the links inside the country are checked, so we do not detect all the broken links. Also, nowadays there is a higher prevalence of "soft-404" messages [Yossef et al. 2004] and other types of redirects to hide broken links, probably because the reorganization of the contents of a Web site occurs several times during a Web site's lifetime. Furthermore, the general quality of Web sites may have changed, in part by the usage of tools for automating link creation and checking, and due to a stronger competition between Web site owners.

5.3 Document formats other than HTML

HTML is the preferred format for documents on the Web, and more than 95% of them are in this format. Other formats such as Adobe PDF and plain text are the most important ones after HTML. Together they account for 70%-85% of the non-HTML files, followed by Microsoft formats such as Word Document (`doc`) and Power Point slides (`ppt`). The distribution of non-HTML file types as determined by file type extensions (`.doc`, `.pdf`, `.ppt`, `.ps`, `.txt`, etc.) is shown in Figure 23, and is rather similar across collections.

5.4 Image formats

The GIF and JPEG formats comprise over 95% of the images, followed by the PNG format in a distant third place. There is evidence suggesting that most of the

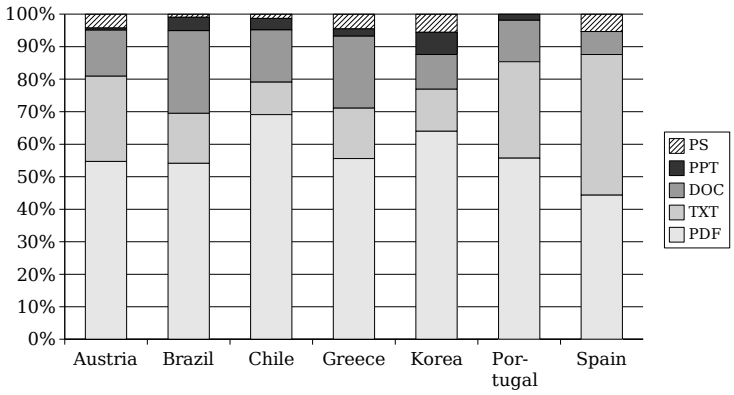

Fig. 23. File types of non-HTML documents.

images included in Web pages are not unique; for instance, in [Jaimes et al. 2004] it was found that 64% of images appearing in home pages were unique, and only 10% of the images in inner pages were unique.

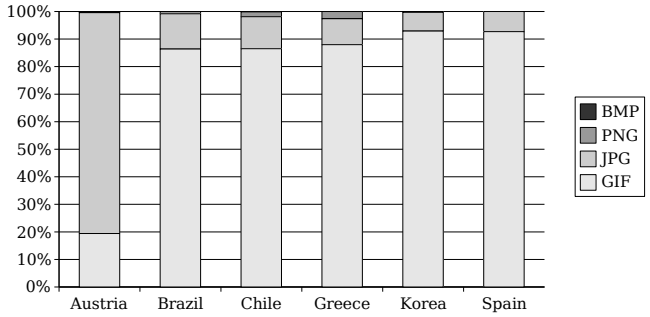

Fig. 24. Image formats. In the case of Austria, only unique images were counted.

The distribution of image types is shown in Figure 24. In the case of Austria, the methodology used for obtaining the distribution is different, as the number of unique images are counted. In the case of Spain, data is taken from a sample of university Web sites [Alonso et al. 2003].

5.5 Web sites that cannot be crawled correctly

Surprisingly, there is a large fraction of Web sites with only one page downloaded by the crawlers in all collections: 37% in Brazil, 40% in Chile, 31% in Greece, 29% in Indochina, 29% in Italy, 40% in South Korea, 38% in Portugal, 60% in Spain and 24% in the United Kingdom.

The most common causes for these Web sites are the following:

(1) The navigation relies completely upon Javascript, Flash or Java. This comprises about 60% of the one-page Websites on average, and is split evenly between

Flash- and Javascript- based navigations. Most Web crawlers cannot follow links embedded in these programs, so pages that have no regular links pointing to them are invisible for search engines.

(2) The home page contains a redirection to another Web site, or only links to external Web sites. This is sometimes done for aliasing, for instance, `www.bbcnews.com` may be easier to remember than `news.bbc.co.uk`. Sometimes this is also done for spamming, creating hundreds of Web sites with redirects. Web sites with redirects or only external links are about 30% of the cases on average.

(3) There is really only one page in the Web site, typically an "under construction" page. This is the remaining 10% of the cases on average.

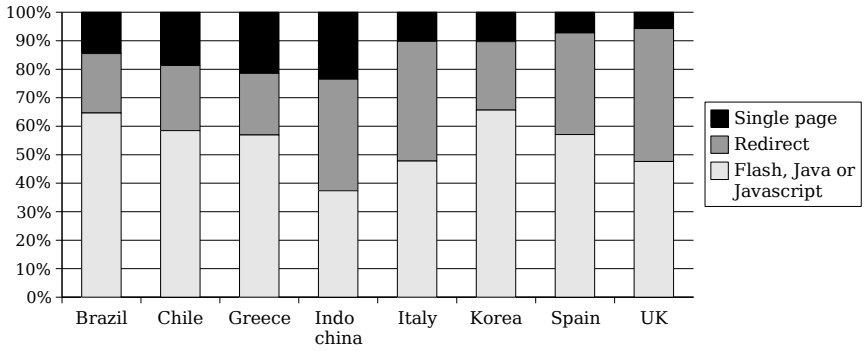

Fig. 25. Distribution of one-page sites.

Figure 25 depicts the distribution of the types of one-page sites. In the case of the Brazilian and Spanish Web, there are several large domains that include multiple redirects to the same page. In the case of the South Korean Web, the authors removed thousands of spam sites using a more elaborate redirection involving a Flash application.

5.6 Web server software

According to Netcraft[12], the most used Web server software is Apache with 63% of the sites, and the second most used is Microsoft Internet Information Server (IIS) with 25% of the sites. Figure 26 shows the distribution of Web server software in the studied domains.

In the African sample, the orders are reversed, this mean that in particular markets there could be important differences when measuring the share of these technologies.

[12]Netcraft Web server survey, accessed May 2006, <`http://news.netcraft.com/archives/web_server_survey.html`>

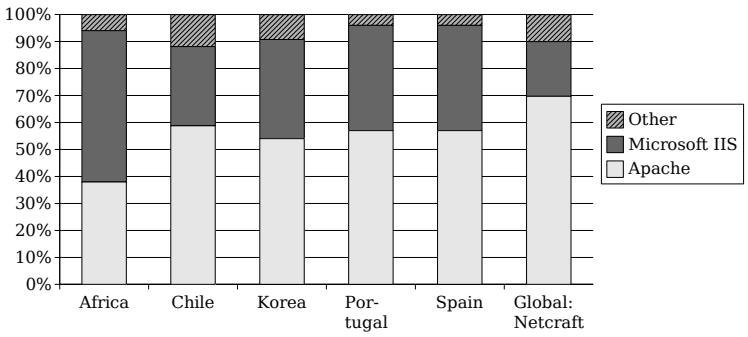

Fig. 26. Technologies used for Web servers.

5.7 Programming languages for dynamic pages

Some pages on the Web are stored in Web servers and then retrieved by users; those pages are called "static pages". Other pages are created whenever they are requested, on demand, and they are called "dynamic pages". Dynamic pages are used to build Web applications, typically to access data sources that cannot be converted entirely into HTML pages due to space, privacy, or other constraints.

The first approach to measure the share of each programming language is to count the number of pages with the file extension that is associated with each programming language. In many systems, extensions can be disguised by configuring the server to hide them or replace them by another. While we cannot measure how frequent this is, we have no reasons to believe that this is done more frequently for some languages than for others, so we do not think that the fact that extensions can be hidden introduces a significant bias in this measurement. When counting in this way, three different groups appear: Africa, South Korea and the U.K. with predominance of ASP; Brazil, Chile and Greece with predominance of PHP; and Indochina, Italy and Spain in the middle. Other technologies are much less used, as shown in Figure 27.

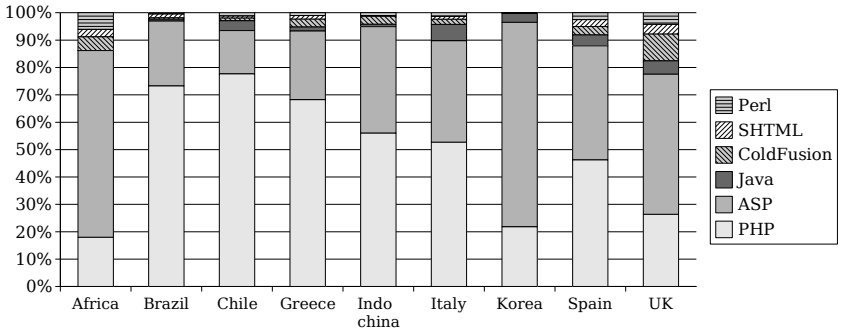

Fig. 27. Fraction of dynamic pages written in different programming languages.

Another approach is to measure which programming languages are used by each Web site. Most Web sites use only one programming language, but there are cases in which several languages are used for different parts of the Web site. On Figure 28, the part marked "MIXED" corresponds to Web sites that use two or more programming languages, comprising from 5% to 20% of the sites.

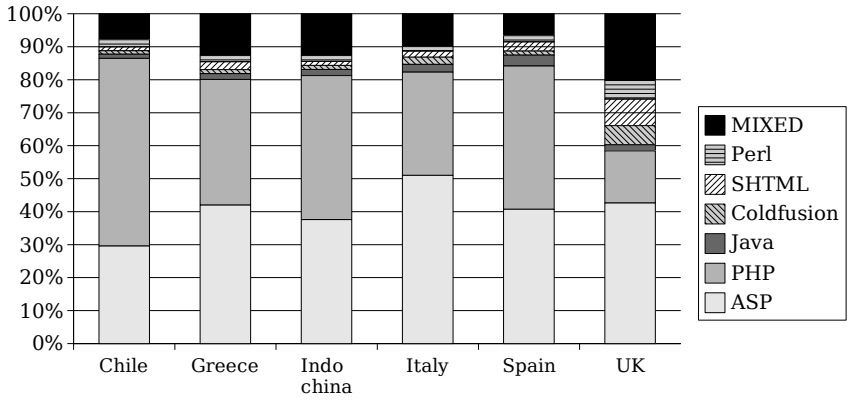

Fig. 28. Fraction of Web sites using different programming languages for dynamic pages.

Finally, in Figure 29, we calculate the average usage of different programming languages, and how frequent it is to use a language in conjunction with another one. For instance, about 3% of the Web sites use Java as a technology for dynamic pages, and when Java is used, in about 25% of the cases it is used with another programming language in the same Web site. ASP is used by over 45% of the dynamic Web sites, and it is used almost always exclusively.

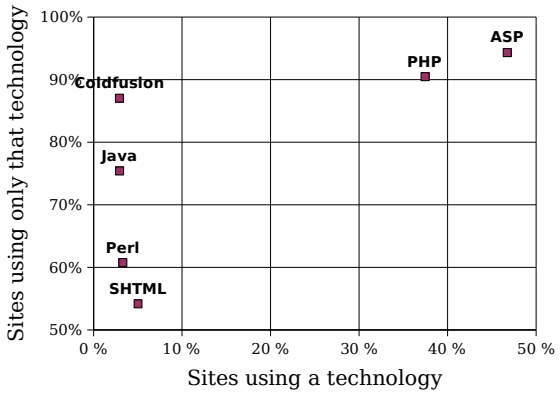

Fig. 29. Share of programming languages across Web sites, averaged across all the collections shown in the previous figure and weighted by the number of Web sites on each collection.

6. WEB CHARACTERISTICS AND SOCIO-ECONOMIC INDICATORS

In this section we compare the characteristics of the national web domains of the countries studied to a number of socio-economic indicators. More specifically, we examine the size of their economies and population and analyze the growth of Internet usage, and the penetration of said usage as a percentage of the population in that country and the region. We then studied the percentage of GDP invested in information and communication technologies. Finally, we look at the growth of Internet advertising by country.

Table V presents the growth of Internet users in the period 2000-2005. Countries are grouped by region, and we have also included data for the U.S.A. and the E.U. in order to facilitate comparisons. User growth in all countries but South Korea is in 3-digit figures, ranging from 118.7% for Italy to 346% for Brazil. The penetration as a percentage of the country's population ranges from 12.3% for Brazil and 12.8% for Thailand to 68.6% for the United States. Greece, Spain, and Chile, and especially, Thailand and Brazil are markets where user growth will be increasing at a fast pace in the very near future.

Table V. Growth of Internet users and usage in the countries studied. Source: World Internet Users and Population Statistics. Online: <http://www.internetworldstats.com/stats.htm>. Retrieved October 2005.

Region/ Country	Population	Internet users	Growth (users) '00-'05	Penetration users/pop.	% Users in region
European Union	460,270,935	225,006,820	141.5%	48.9%	100.0%
Austria	8,163,782	4,650,000	121.4%	57.0%	2.1%
Greece	11,212,468	3,800,000	280.0%	33.9%	1.7%
Italy	58,608,565	28,870,000	118.7%	49.3%	12.8%
Portugal	10,463,170	6,090,000	143.6%	58.2%	2.7%
Spain	43,435,136	16,129,731	199.4%	37.1%	7.2 %
United Kingdom	59,889,407	36,059,100	134.2%	60.2%	16.0%
North America					
United States	296,208,476	203,274,683	113.2%	68.6%	90.8%
South America					
Brazil	181,823,645	22,320,000	346.4%	12.3%	45.9%
Chile	15,514,014	5,600,000	218.7%	36.1%	11.5%
Asia					
Thailand	65,699,545	8,420,000	266.1%	12.8%	2.6%
South Korea	49,929,293	32,570,000	71.1%	65.2%	10.0%

Table VI shows the expenditure for information and communication technologies (ICT) by country for the period 2000-2003. It is noticeable that the U.S. and European countries, like Austria, Italy, and U.K., with advanced economies show a downward trend in ICT investing, nevertheless at relatively healthy levels. In

less developed economies within the EU, like Portugal, Spain, and Greece, ICT expenditure is at about the same level throughout the period. Chile, but especially Brazil, have been increasing their investment in ICTs. For example, in 2002 Brazil invested US$4.4 billion dollars in telecommunication infrastructure. This is yet another signal for the expected growth of the Internet in South America.

Table VI. Information and communication technology expenditure (% of GDP). Source: Worldbank WDI Online. Retrieved in October 2005 from Worldbank WDI Online: <http://devdata.worldbank.org/wdi2005/>

Country	2000	2001	2002	2003
Austria	6.0	5.9	5.8	5.3
Brazil	5.6	6.0	6.9	6.7
Chile	6.0	6.2	6.7	6.7
Greece	4.5	4.4	4.4	4.3
Italy	4.8	4.6	4.5	4.1
South Korea	6.8	6.4	6.6	6.6
Portugal	4.4	4.3	4.3	4.2
Spain	4.1	4.1	4.1	3.8
Thailand	3.5	3.6	3.6	3.5
United Kingdom	8.1	7.8	7.6	7.3
United States	9.5	8.7	8.6	8.8

The growth of the Web is also reflected through the online advertising expenditures. Such data in general is difficult to get as the Internet Advertising Bureau (IAB) is still refining data collection and reporting methodologies. Data for the United States is relatively more readily available, whereas data for the other countries in the study are sparse and difficult to get. Nevertheless, very informative. Table VII gives advertising revenues for the United States for the period 1995-2004, and Table VIII provides forecasts for U.S. advertising revenues for the period 2005-2008. It is evident that the revenues have been exponential starting at $55 million in 1995 to $9.6 billion in 2004 and with a forecast to double to $18.5 billion by 2008. Notice that the 2004 forecast for 2005 was more than 10% less than the actual value.

Table VII. Internet advertising revenue report for the U.S. 1995-2005 (millions of dollars). Source: PriceWaterhouseCoopers LLP. Internet Advertising Bureau (IAB) Internet advertising revenue report. Retrieved in May 2006 from <http://www.iab.net/resources/adrevenue/pdf/IAB_PwC_2005.pdf>.

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
55	267	907	1,920	4,621	8,087	7,134	6,010	7,267	9,626	12,542

Table VIII. Internet advertising revenue forecasts for the US, 2005-2008 (in billions). Source: US Online Advertising Spending, eMarketer Report, July 2004 (Retrieved in October 2005 from <http://www.emarketer.com/Report.aspx?ad_spend_aug04>.)

	2005	2006	2007	2008
US	11.2	13.5	16	18.5

The efforts of the Internet Advertising Bureau outside the United States has been collected primarily in Europe. As seen in Table IX three-fold growth of advertising revenues have been observed in the U.K. and four-fold in Greece between 2000 and 2004. Greece's revenues are at 0.008% of the GDP and British revenues are at the 0.042% of the GDP. These figures will continue to grow because as seen in Table V the number of Internet users is at 33.9% for Greece and 60% for the U.K.

Table IX. European online advertising expenditures by country (millions of Euros). Source: Information taken from www.iabeurope.ws: Multimarket European Online AdSpend Figures: Spring 2005 Interactive Advertising Bureau Europe. IAB Europe: multimarket European online adspend figures. Retrieved in October 2005 from <http://www.iab.it/fmknet/View.aspx?da_id=1730>.

Country	1998	1999	2000	2001	2002	2003	2004
Greece	..	..	3.14	2.97	4.6	9.5	11
Italy	..	..	125.2	122.4	115.4	113.6	117.1
Spain	..	..	..	..	71.5	72.5	94.5
UK	29	76	231	248	294	562	653

7. CONCLUSIONS

We observed that the results across different collections are always consistent when the observed characteristic exhibits a power-law in one collection. In this class we include the distribution of page sizes, degrees, link-based scores, etc. On the other hand, for the distribution of Web site into the components of the Web graph, our results are mixed and include countries with very similar and very dissimilar distributions.

Some technological characteristics that are shared across countries are the distribution of URL lengths, which follow a log-normal distribution, and the HTTP response codes, which always show roughly the same ratio of success. The market shares of Web server software and image formats are also very stable across national domains, but other technologies vary more, such as non-HTML file types and programming languages.

Not surprisingly, natural language is the most varying characteristic across the national domains studied. We also found significant differences in the distribution of registrations under second-level domains across countries. In several aspects, the

collection of pages from South Korea was significantly different than the others; mostly because of a massive presence of spam.

Another important remark is that statistics based in national domains are incomplete in many cases. This is due to three main reasons:

(1) not all sites use the national domain; this is clear for USA where the `.us` domain is seldom used, but is also true for Spain and other countries,
(2) there are many unknown domains that are islands; and
(3) there are many sites that are not crawlable and hence their size and contents are also unknown.

Using our data we can approximate the real value of a measure M using the value of M for a country domain using:

$$\mathrm{M}(Total) = f_d \times f_{is} \times f_{nc} \times \mathrm{M}(Known)$$

where f_d, f_{is} and f_{nc} are estimated factors larger than 1, that depend on the "hidden size" due to other domains, islands and not-crawlable sites, respectively. For example, for Spain we can estimate the total number of pages using $f_d = f_{nc} \approx 5/3$ and $f_{is} = 1$ as a lower bound. These estimations come from the number of sites outside `.es`, the different number of pages per site on them, and the percentage of non-crawlable sites that we found. We know that `.es` has 9 million pages, so the overall Web of Spain has more than 25 million pages, which means that our study of Spain has at most 64% coverage (we crawled 16 million).

Web characterization studies of the Web using a Web crawler generate a view of the Web that is not what users are accustomed to see. For instance, while a Web crawler has no problem in downloading a page with 500 KB of HTML data (there are some examples), very few users will have the patience to wait for it. Most users have a routine of visiting a few selected, high-quality Web sites on a daily basis, and do not browse through obscure and mostly unknown Web pages. Web characterization studies focusing on what users actually *see*, instead of what is available, would be complementary with crawler-based studies. In that case, trace logs should be used for obtaining the Web pages.

Furthermore, the use of socio-economic indicators in Web characterization studies of national Webs supplement the information gathered from the crawlers and provide additional methods for explaining the behavior and growth of the Web.

Finally, any Web comparison is approximate, as the Web is not a static object, not only the content is constantly growing and changing, also connectivity and server performance changes. In addition, different crawlers will gather different samples. Nevertheless, in spite of all these factors, we believe that comparative studies give insight about Web characteristics and trends.

ACKNOWLEDGMENTS

We worked with Vicente López in the study of the Spanish Web, with Felipe Ortiz, Bárbara Poblete and Felipe Saint-Jean in the studies of the Chilean Web and with Felipe Lalanne in the study of the South Korean Web. We also thank the Laboratory of Web Algorithmics for making their Web collections available for research.

June 2006. Submitted for publication.

REFERENCES

Alonso, J. L., Figuerola, C. G., and Zazo, Á. F. 2003. *Cibermetría: nuevas técnicas de estudio aplicables al Web*. Ediciones TREA, Spain.

Arlitt, M., Friedrich, R., and Jin, T. 1999. Workload characterization of a Web proxy in a cable modem environment. *SIGMETRICS Performance Evaluation Review 27,* 2, 25–36.

Baeza-Yates, R. and Castillo, C. 2000. Caracterizando la Web Chilena. In *Encuentro chileno de ciencias de la computación*. Sociedad Chilena de Ciencias de la Computación, Punta Arenas, Chile.

Baeza-Yates, R. and Castillo, C. 2001. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval SPIRE*. IEEE CS Press, Laguna San Rafael, Chile, 21–32.

Baeza-Yates, R. and Castillo, C. 2002. Balancing volume, quality and freshness in Web crawling. In *Soft Computing Systems - Design, Management and Applications*. IOS Press Amsterdam, Santiago, Chile, 565–572.

Baeza-Yates, R. and Castillo, C. 2004. Crawling the infinite Web: five levels are enough. In *Proceedings of the third Workshop on Web Graphs (WAW)*. Lecture Notes in Computer Science, vol. 3243. Springer, Rome, Italy, 156–167.

Baeza-Yates, R. and Castillo, C. 2005. Características de la Web Chilena 2004. Tech. rep., Center for Web Research, University of Chile.

Baeza-Yates, R., Castillo, C., and Lpez, V. 2006. Caractersticas de la Web de Espaa. *El Profesional de la Informacin 15,* 1 (January).

Baeza-Yates, R. and Lalanne, F. 2004. Characteristics of the Korean Web. Tech. rep., Korea–Chile IT Cooperation Center ITCC.

Baeza-Yates, R. and Navarro, G. 2004. Modeling text collections and its application to the Web. *Applied Probability: Recent Advances*.

Baeza-Yates, R. and Poblete, B. 2003. Evolution of the Chilean Web structure composition. In *Proceedings of Latin American Web Conference*. IEEE CS Press, Santiago, Chile, 11–13.

Baeza-Yates, R., Poblete, B., and Saint-Jean, F. 2003. Evolución de la Web Chilena 2001–2002. Tech. rep., Center for Web Research, University of Chile.

Barr, D. 1996. RFC 1912: Common DNS operational and configuration errors. http://www.ietf.org/rfc/rfc1912.txt.

Bharat, K., Chang, B. W., Henzinger, M., and Ruhl, M. 2001. Who links to whom: Mining linkage between Web sites. In *International Conference on Data Mining (ICDM)*. IEEE CS, San Jose, California, USA, 51–58.

Björneborn, L. and Ingwersen, P. 2004. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology 55,* 14 (August), 1216–1227.

Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2002. Structural properties of the African Web. In *Proceedings of the eleventh international conference on World Wide Web*. ACM Press, Honolulu, Hawaii, USA.

Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2004. UbiCrawler: a scalable fully distributed Web crawler. *Software, Practice and Experience 34,* 8, 711–726.

Brewington, B., Cybenko, G., Stata, R., Bharat, K., and Maghoul, F. 2000. How dynamic is the web? In *Proceedings of the Ninth Conference on World Wide Web*. ACM Press, Amsterdam, Netherlands.

Brin, S., Motwani, R., Page, L., and Winograd, T. 1998. What can you do with a Web in your Pocket? *IEEE Data Engineering Bulletin 21,* 2, 37–47.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. 2000. Graph structure in the Web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*. ACM Press, Amsterdam, Netherlands, 309–320.

Cavnar, W. B. and Trenkle, J. M. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, US, 161–175.

da Silva, A. S., Veloso, E. A., Golgher, P. B., , Laender, A. H. F., and Ziviani, N. 1999. CoBWeb - A crawler for the Brazilian Web. In *Proceedings of String Processing and Information Retrieval (SPIRE)*. IEEE CS Press, Cancun, Mxico, 184–191.

Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., and Tomkins, A. 2002. Self-similarity in the web. *ACM Trans. Inter. Tech. 2,* 3, 205–223.

Dorogovtsev, S. N. and Mendes, J. F. F. 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press.

Downey, A. B. 2001. The structural cause of file size distributions. In *Proceedings of the 9th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS)*. IEEE CS Press, Cincinnati, Ohio, USA.

Efthimiadis, E. and Castillo, C. 2004. Charting the Greek Web. In *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*. American Society for Information Science and Technology, Providence, Rhode Island, USA.

Eiron, N., Curley, K. S., and Tomlin, J. A. 2004. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*. ACM Press, New York, NY, USA, 309–318.

Fetterly, D., Manasse, M., and Najork, M. 2004. Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages. In *Proceedings of the seventh workshop on the Web and databases (WebDB)*. Paris, France, 1–6.

Gomes, D. and Silva, M. J. 2005. Characterizing a national community Web. *ACM Transactions on Internet Technology 5,* 3.

Grefenstette, G. and Nioche, J. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of Content-Based Multimedia Information Access (RIAO)*. Paris, France, 237–246.

Gyöngyi, Z. and Garcia-Molina, H. 2005. Web Spam Taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*.

Heydon, A. and Najork, M. 1999. Mercator: A Scalable, Extensible Web Crawler. *World Wide Web Conference 2,* 4 (April), 219–229.

Huberman, B. A. and Adamic, L. A. 1999. Growth dynamics of the World-Wide Web. *Nature 399*.

Jaimes, A., , Verschae, R., Baeza-Yates, R., Castillo, C., Yaksic, D., and Davis, E. 2004. On the image Content of a Web segment: Chile as a case study. *Journal of Web Engineering 3,* 2, 153–168.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46,* 5, 604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. 1999. The Web as a graph: measurements, models and methods. In *Proceedings of the 5th Annual International Computing and Combinatorics Conference (COCOON)*. Lecture Notes in Computer Science, vol. 1627. Springer, Tokyo, Japan, 1–18.

Mitzenmacher, M. 2003. Dynamic Models for File Sizes and Double Pareto Distributions. *Internet Mathematics 1,* 3, 305–333.

Modesto, M., Pereira, á., Ziviani, N., Castillo, C., and Baeza-Yates, R. 2005. Um novo retrato da Web Brasileira. In *Proceedings of XXXII SEMISH*. So Leopoldo, Brazil, 2005–2017.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The PageRank citation ranking: bringing order to the Web. Tech. rep., Stanford Digital Library Technologies Project.

Pandurangan, G., Raghavan, P., and Upfal, E. 2002. Using Pagerank to characterize Web structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*. Lecture Notes in Computer Science, vol. 2387. Springer, Singapore, 330–390.

Pitkow, J. E. 1999. Summary of WWW characterizations. *World Wide Web 2,* 1-2, 3–13.

Rauber, A., Aschenbrenner, A., Witvoet, O., Bruckner, R. M., and Kaiser, M. 2002. Uncovering information hidden in Web archives. *D-Lib Magazine 8,* 12.

Sanguanpong, S., Nga, P. P., Keretho, S., Poovarawan, Y., and Warangrit, S. 2000. Measuring and analysis of the Thai World Wide Web. In *Proceeding of the Asia Pacific Advance Network conference*. Beijing, China, 225–230.

Sanguanpong, S. and Warangrit, S. 1998. NontriSearch: search engine for campus network. In *National Computer Science and Engineering Conference*. Bangkok, Thailand.

Suel, T. and Yuan, J. 2001. Compressing the graph structure of the Web. In *Proceedings of the Data Compression Conference DCC*. IEEE CS Press, Snowbird, Utah, USA.

Veloso, E. A., de Moura, E., Golgher, P., da Silva, A., Almeida, R., Laender, A., Neto, R. B., and Ziviani, N. 2000. Um retrato da Web Brasileira. In *Proceedings of Simposio Brasileiro de Computacao*. Curitiba, Brasil.

Yossef, Z. B., Broder, A. Z., Kumar, R., and Tomkins, A. 2004. Sic transit gloria telae: towards an understanding of the web's decay. In *Proceedings of the 13th conference on World Wide Web*. ACM Press, New York, NY, USA.

Zipf, G. K. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA, USA.

Received Month YYYY; revised Month YYYY; accepted Month YYYY.