Chapter 8

Characterization of the Chilean Web

As an application of the crawler implementation presented in Chapter **??**, we downloaded and studied the pages under the `.CL` top-level domain.

The WIRE crawler includes a module for generating statistics and reports about its collection. In this chapter, we present several characteristics of the Chilean Web, and we compare some of them with the characteristics of the Greek Web.

8.1 Reports generated by WIRE

The procedure for generating reports has the following steps:

1. Analysis of the metadata that is kept in the data structures described in Section **??** (page **??**), and generation of statistics as plain text files.
2. Generation of `gnuplot` scripts to generate graphs, and invocation of `gnuplot`.
3. Generation of the report using LaTeX.

This procedure makes maintenance of the reports easier, as the data is separated from the representation. The exact commands for generating reports are detailed in the user manual that is available on-line at `http://www.cwr.cl/projects/WIRE/doc/`.

The generated reports include:

- A report about characteristics of the pages that were downloaded.
- A report about links found in those pages.
- A report about languages.

- A report about Web sites.
- A report about links in the Web site graph.

In this chapter, most of the data tables and graphics (except for pie charts, due to a limitation of the GNU plot program) were generated using the WIRE report generator.

8.2 Collection summary

Table 8.1 summarizes the main characteristics of the collection, which was obtained in May 2004.

Table 8.1: Summary of the characteristics of the studied collection from the Chilean Web.

Downloaded Web pages	3,313,060	
Static	2,191,522	66.15%
Dynamic	1,121,538	33.85%
Unique	3,110,205	93.88%
Duplicates	202,855	6.12%

Downloaded Web sites	49,535
Static pages per site	40.40
Dynamic pages per site	26.73
Pages per site	67.13

We downloaded up to five levels of dynamic pages and also up to 15 levels of static pages. We also limited the crawl to HTML pages, downloading at most 300 Kb of data per URL, with a maximum of 20,000 pages per Web site.

8.3 Web page characteristics

8.3.1 Status code

Figure 8.1 shows the distribution of the HTTP response code. In the figure, we have merged several HTTP response codes for clarity:

- **OK** includes requests that lead to a page transfer: `OK` (200) and `PARTIAL CONTENT` (206) responses.
- **MOVED** includes all the redirects to other pages: `MOVED` (301), `FOUND` (302) and `TEMPORARY REDIRECT` (307).
- **SERVER ERROR** includes all failures on the server side: `INTERNAL SERVER ERROR` (500), `BAD GATEWAY` (502), `UNAVAILABLE` (503), and `NO CONTENT` (204).

- **FORBIDDEN** includes all the requests that are denied by the server: `UNAUTHORIZED` (401), `FORBIDDEN` (403) and `NOT ACCEPTABLE` (406).

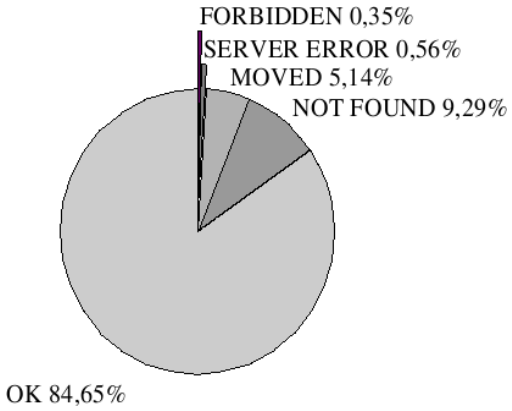


Figure 8.1: Distribution of HTTP response code.

In all our experiments, we usually have had between 75% and 85% of responses leading to an actual transfer of data. From the point of view of Web crawler, the fraction of failed requests is significant, and it should be considered in short-term scheduling strategies when "overbooking" the network connection.

The fraction of broken links, over 9%, is very significative. This means that the Web is dynamic, and quality control on existent Web sites is neither meticulous nor frequent enough.

8.3.2 Content length

To save bandwidth, we downloaded only the first 300 KB of the pages. The center of the distribution of page sizes follows a Zipf's law of parameter -3.54, as shown in Figure 8.2. Close to 300 KB the number of pages looks higher than expected because of the way in which we enforced the download limit.

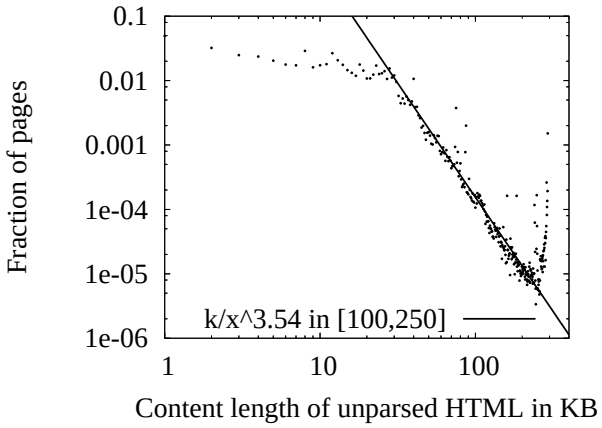


Figure 8.2: Distribution of content length of pages.

We observe that below 12 Kilobytes, there are fewer pages than predicted by the Zipf's law. This is because of a limit of HTML coding: the markup is not designed to be terse and even a short text requires a certain amount of markup. As HTML is used as a presentational language, controlling the formatting attributes of the pages, it generates a significant overhead over text size, especially for complex designs.

For a Web crawler, 300Kb seems to be a safe limit for HTML pages, as there are very few Web pages with more than this amount of data.

8.3.3 Document age

We observed the last-modification date returned by Web servers, and applied the heuristic described in Section **??** (page **??**) to discard wrong dates. We found that 83% of the Web sites studied returned valid last-modified dates for their pages. The distribution of page age in terms of months and years is shown in Figure 8.3.

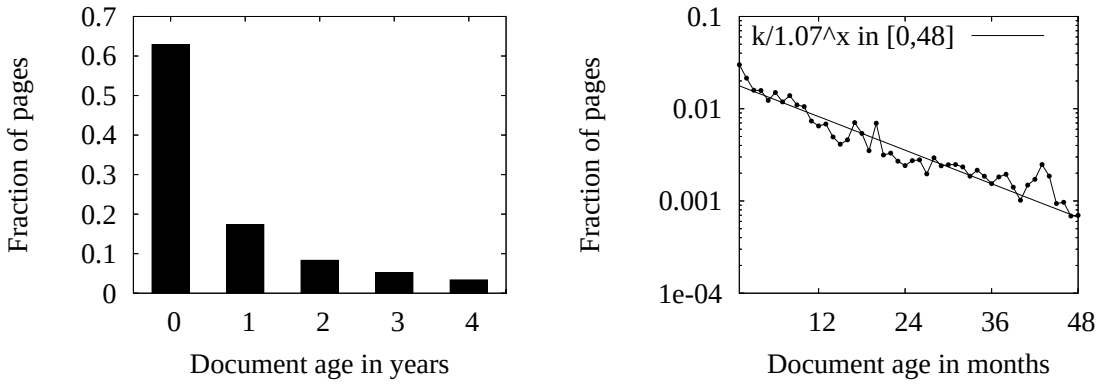


Figure 8.3: Distribution of page age. Note that for the graph of Page age in months the scale is semi-log.

Page changes exhibit an exponential distribution, as seen in the graphic of document age in months. Note that more than 60% of pages have been created or modified in the last year, so the Chilean Web is growing at a fast pace.

8.3.4 Page depth

We limited the browser to download only five levels of dynamic pages, and up to 15 levels of static pages. The distribution of pages by depth is shown in Figure 8.4.

The distribution of static pages follows a shape whose maximum is in the fifth level, but the distribution of dynamic pages tends to grow without bounds. This is because dynamic pages have links to other dynamic pages, as discussed in Chapter **??**.

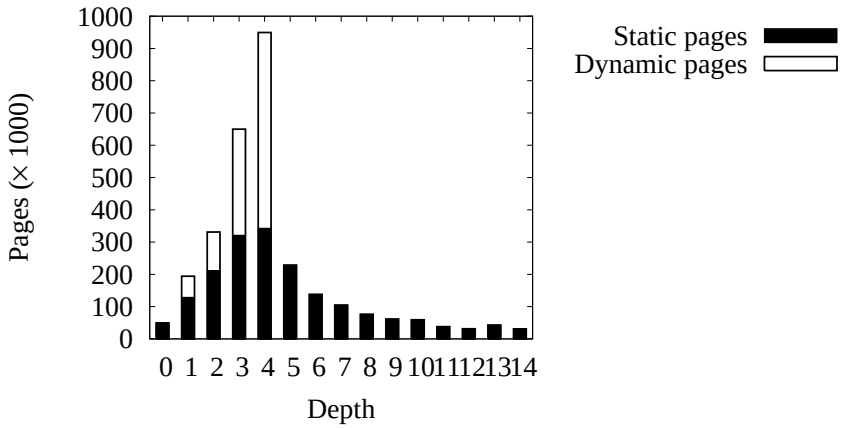


Figure 8.4: Distribution of pages at different depths.

8.3.5 Languages

We took a sample of 5,000 pages, and analyzed their contents to compare their word lists against a series of lists of stop words in several languages on the Chilean Web. We found about 71% of the pages in Spanish, and 27% in English. Other languages appeared with much less frequency.

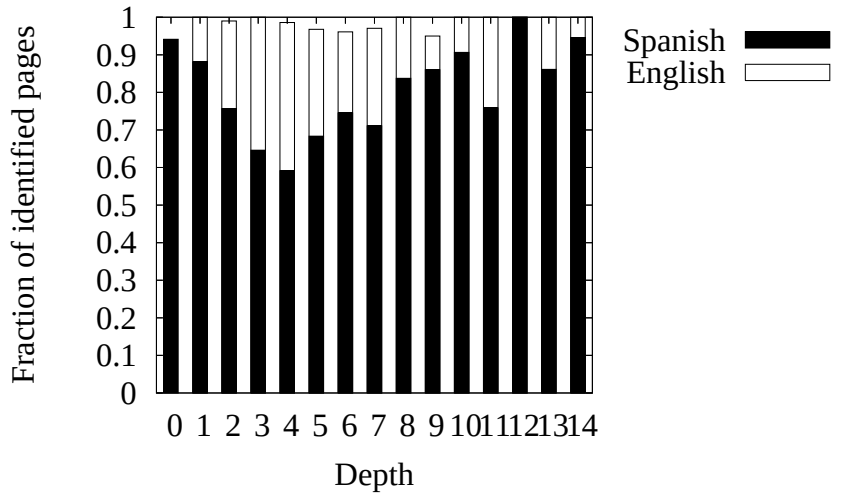


Figure 8.5: Distribution of languages by page depth.

There are large variations in this distribution if we check Web sites at different levels, as shown in Figure 8.5. For instance, over 90% of home pages are mostly in Spanish, but this figure goes as low as 60% if we take pages at depth 5, as shown in Figure 8.5. By inspecting a sample of the English pages at deeper levels, we found that their are mostly mirrors of Web sites such as the Linux Documentation Project, or the TuCows shareware repository.

8.3.6 Dynamic pages

About 34% of the pages we downloaded were dynamically generated. The most used application was PHP [php04], followed by ASP [asp04] and pages generated using Java (`.jhtml` and `.jsp`). The distribution is shown in Figure 8.6.

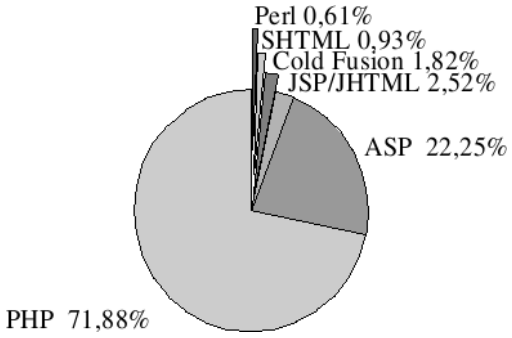


Figure 8.6: Distribution of links to dynamic pages.

PHP, an Open Source technology clearly dominates the market. Dynamic pages are mostly built using hypertext pre-processing (PHP, ASP, JHTML, ColdFusion), in which commands for generating dynamic content, such as accessing a database, are embedded in documents that are mostly HTML code. It must be considered also that some dynamic pages use HTML extension, and that some of the static pages in HTML are generated automatically using batch processing with content management systems, so there are other technologies for dynamic pages that could be missing from this analysis.

8.3.7 Documents not in HTML

We found 400,000 links to non-HTML files containing extensions used for documents. Portable Document Format (PDF) is the most widely used format and the de facto standard, followed by plain text and Microsoft Word. The distribution is shown in Figure 8.7.

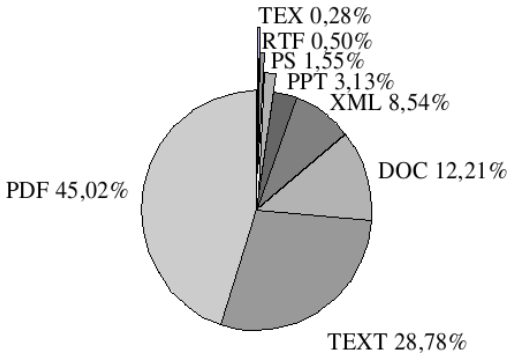


Figure 8.7: Distribution of links to documents found on Chilean Web pages, excluding links to HTML files.

Despite the fact that Microsoft Windows is the most used operating system, file types associated with Microsoft Office applications such as Word or Powerpoint are not used as much as we could expect, probably because of concerns of viruses or lost of formatting.

There are over 30,000 XML files in the Chilean Web, including files with the extensions DocBook, SGML, XML and RDF. In our opinion, this amount of links suggest that it is worth to download those XML files and analyze them, as we could start searching on them.

8.3.8 Multimedia

There are several links to multimedia files, including over 80 million links to images, 50,000 links to audio files, and 8,000 links to video files. The distribution of file formats is shown in Figure 8.8.

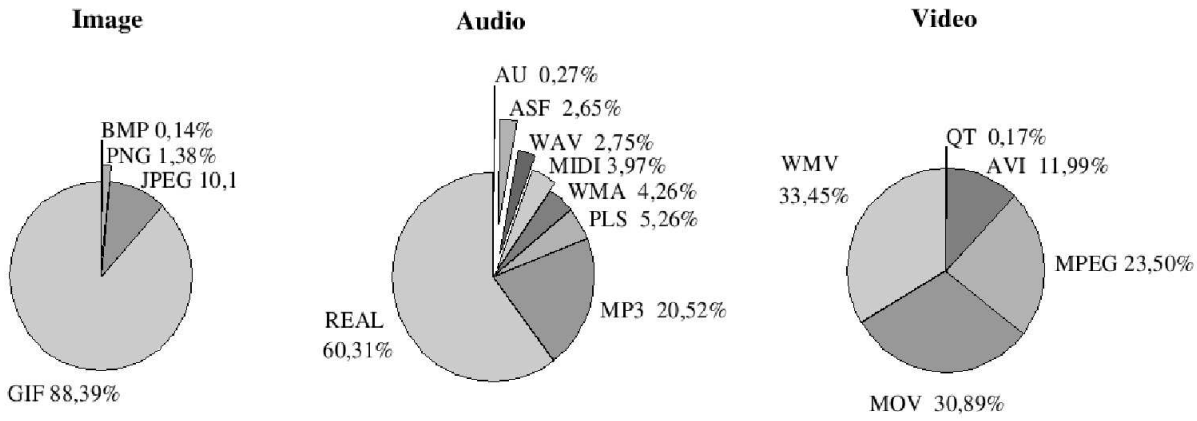


Figure 8.8: Distribution of links to multimedia files found on Chilean Web pages.

Compuserve `GIF` is the most used file format for images, followed by `JPEG`. The Open Source PNG format, which was conceived as a replacement of GIF is still rarely used. The contents of these images was analyzed in the context of a face detection is analyzed in [**?**].

Realnetwork's `Realaudio` and `MP3` are the most used file formats for audio, and are mostly used for streaming in Internet radios. In the case of video, there is no clear dominant format and there are relative few video images on the Web (1/1000 of the quantity of image files). We also found over 700,000 links to Flash animations, mostly in the home pages of Web sites.

We found that about 1/3 of the links to multimedia files from home pages were not unique, and that this fraction falls to 1/10 of the links when internal pages are considered. This suggests that Web site designers usually have a small set of images that are used across their entire Web sites.

8.3.9 Software and source code

We found links to 30,000 files with extensions used for source code, and 600,000 files with extensions used for software. The later does not count software that is distributed in compressed files such as `.tar` or `.zip`. The distribution of the links found is shown in Figure 8.9.

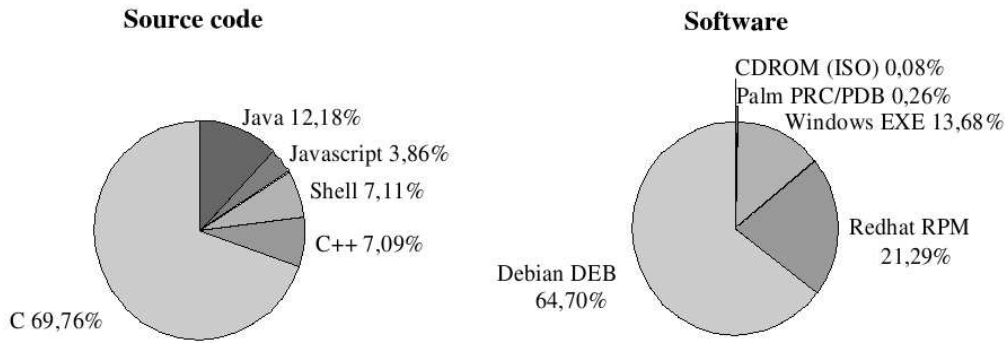


Figure 8.9: Distribution of links to source code and software.

Note that the number of files containing software packages for Linux distributions doubles the one for Windows software; the explanation is that in Linux an application is usually comprised of several packages. Nevertheless, this reflects a comparable level of availability of software packages for both platforms.

Software repositories are usually mirrored at several locations, and the prevalence of mirrors on the Web is in general very high, but the method for avoiding duplicates explained in Section **??** (page **??**) worked very well in removing these mirrors, as we only have 6% of duplicate pages.

8.3.10 Compressed files

We found links to 370,000 files with extensions used for packed or compressed files, and their distribution is shown in Figure 8.10.

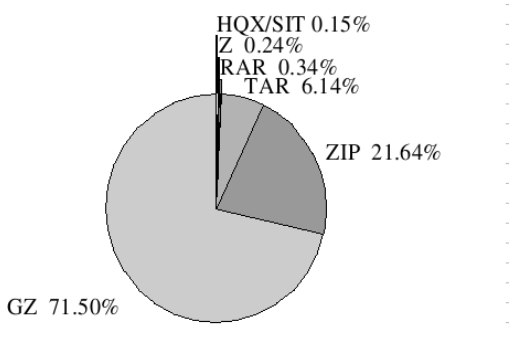


Figure 8.10: Distribution of links to compressed files.

The `GZ` extension, used in the GNU `gzip` program, is the most common extension. Note that in this

case these files probably include software packages that are not counted in Figure 8.9.

8.4 Web site characteristics

8.4.1 Number of pages

We observed an average of 67.1 pages per Web site, but the mode is much smaller than that. The distribution of the number of Web pages on Web sites is very skewed, as shown in Figure 8.11, following a Zipf's law of parameter -1.77.

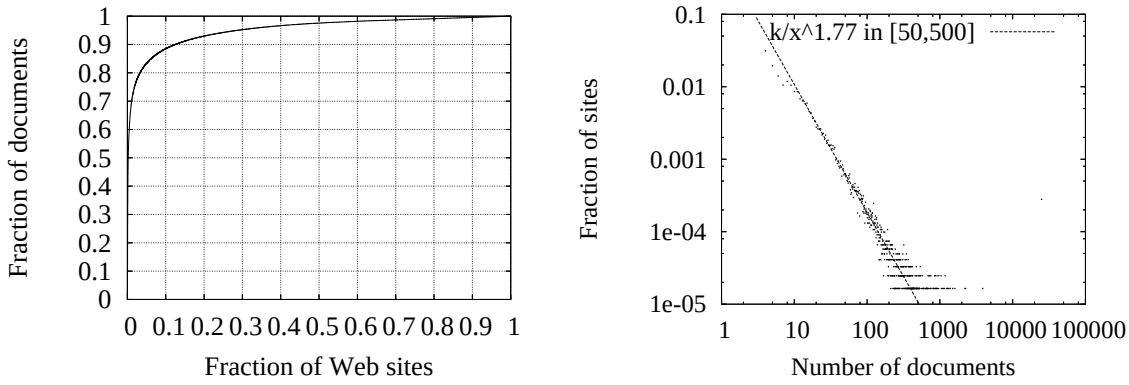


Figure 8.11: Pages per Web site.

.

There are many domain names that are registered with the sole purpose of reserving the name for later use. For instance, only half of the registered domain names under `.cl` have a Web site, and from those, about half have only one page, so only about a quarter of the Web sites are proper Web sites with at least two pages. Although the number of Web sites on the Chilean Web has doubled in the last three years, the fraction of Web sites with just one page remains constant.

On the other end, there are very large Web sites. The top 10% of the Web sites contain over 90% of the Web pages.

8.4.2 Page size

The average size of a complete Web site, considering only HTML pages, is about 1.1 Megabytes (we do not know the total amount of information on the Web site, as we did not download multimedia files). The distribution of the size of Web sites in terms of bytes is also very skewed, as can be seen on Figure 8.12. It is even more skewed than the distribution of the number of pages, as the top 10% of Web sites contain over 95% of the total page contents in bytes.

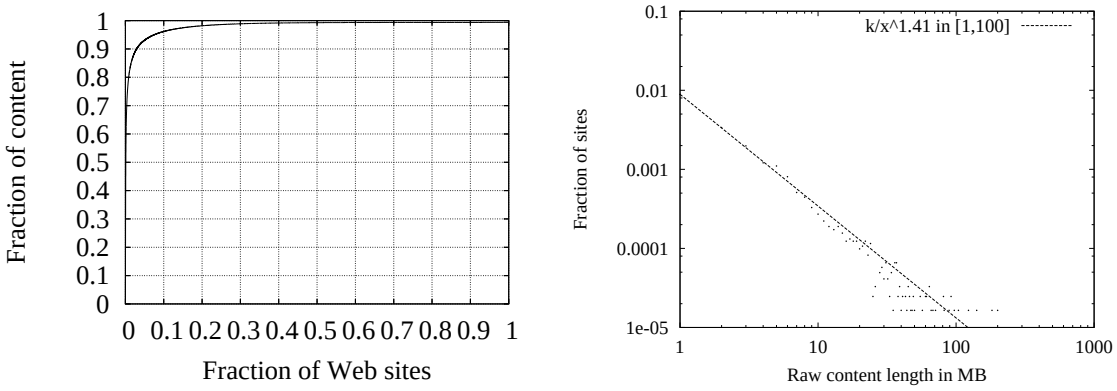


Figure 8.12: Page contents per Web site.

.

The distribution of the page sizes suggests that using the schemes for server cooperation, presented in Chapter **??**, with just a few large Web sites could be very efficient.

8.4.3 Maximum depth

As defined in Chapter **??**, the home page of a Web site is at level 0, and the level of a page is the shortest path from that page to the home page of its Web site.

Most of the Web sites we studied are very shallow, as shown in Figure 8.13. The average maximum depth of a Web site is 1.5.

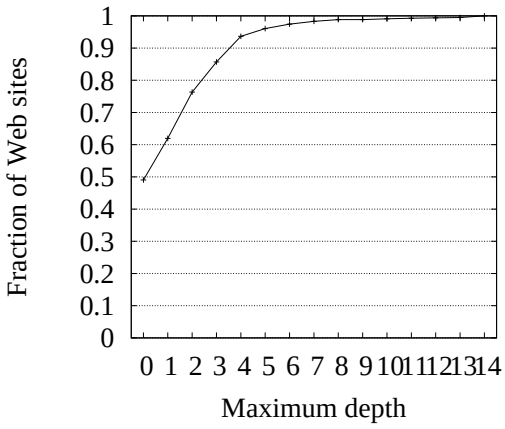


Figure 8.13: Cumulative Web site maximum depth.

.

The distribution of the maximum depth of Web sites is further evidence in favor of what is proposed in

Chapter **??**, namely, downloading just a few levels per Web site.

8.4.4 Age

We measured the age of Web sites, observing the age of the oldest and newest page, as well as the average age. The age of the oldest page is a upper bound on how old the Web site is, and the age of the newest page is a lower bound on how often the Web site is updated. The results are shown in Figure 8.14.

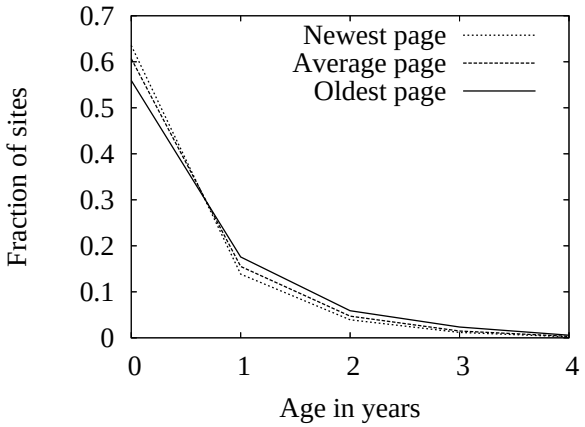


Figure 8.14: Web site age.

.

According to this figure, about 55% of the Web sites were created this year, and about 3/4 of the Web sites in the last 2 years. This implies that for obtaining a large coverage in a national top-level domain, it is necessary to obtain the most recently registered domain names frequently.

8.5 Links

8.5.1 Degree

The distribution of links is skewed, with very few pages having large amounts of links. The distribution of in-degree is much more skewed than the distribution of out-degree, as shown in Figure 8.15: having a Web page with a large in-degree is much more difficult than having a page with a large out-degree.

The distribution of out-degree is similar to the distribution of page-sizes, and there is indeed a correlation between both, as a page cannot have too many links if it is too small, as shown in Figure 8.16.

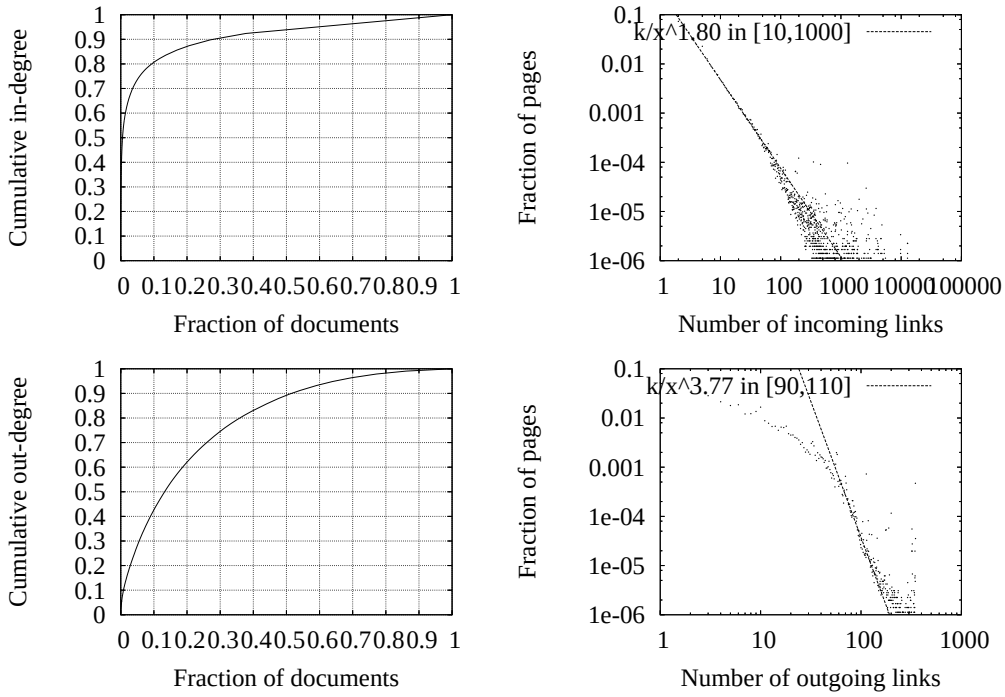


Figure 8.15: Distribution of in- and out-degree.

8.5.2 Link scores

We compared the distributions of Pagerank [PBMW98], and a variation of the HITS algorithm [Kle99], in which we use the entire Web as the expanded root set (this can be seen as a static version of HITS).

The distribution of link scores is shown in Figure 8.17.

As Pagerank is calculated using random jumps to other pages, even pages with very few links have a "parasitic" Pagerank value (no page has zero probability) that is lower than $1/N$, where N is the number of pages.

On the other hand, a page needs "good" links (out-links to authorities in the case of hubs, in-links from hubs in the case of authorities) to have a non-zero value. Only 12% of pages have a non-zero hub value and only 3% of pages have a non-zero authority value.

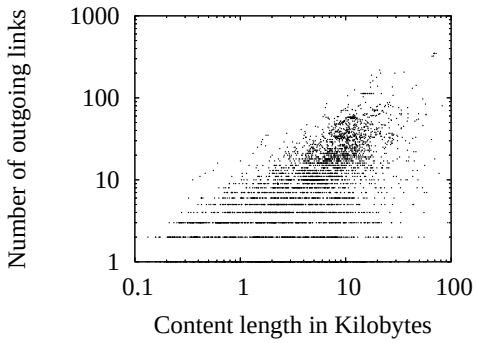


Figure 8.16: Content length vs number of outgoing links.

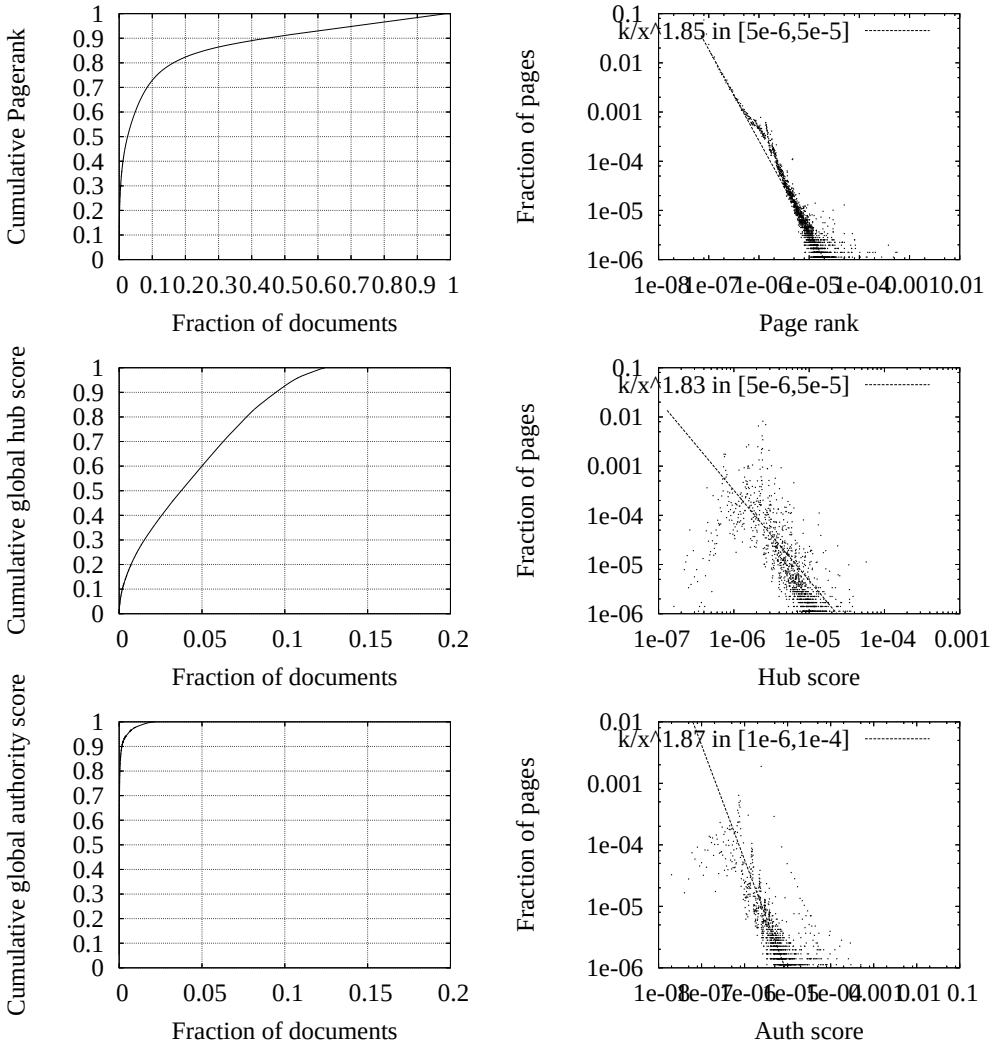


Figure 8.17: Distribution of Pagerank, global Hubs and Authority scores.

8.5.3 Links to external domains

We found 2.9 million outgoing external links, i.e.: links that include a host name part. After discarding links to other hosts in .CL we obtained 700,000 links. The distribution of links into domains for the top 20 domains is shown in Table 8.2.

Top level domain	Percent of links	Top level domain	Percent of links
COM	65.110%	PE - Peru	0.558%
ORG	11.806%	ES - Spain	0.494%
NET	8.406%	FR - France	0.464%
DE - Germany	1.621%	JP - Japan	0.462%
MX - Mexico	1.059%	NL - Netherlands	0.444%
BR - Brazil	0.977%	IT - Italy	0.431%
AR - Argentina	0.846%	VE - Venezuela	0.400%
CO - Colombia	0.809%	TW - Taiwan	0.382%
UK - United Kingdom	0.644%	SG - Singapur	0.371%
EDU	0.609%	KR - Korea	0.370%

Table 8.2: Fraction of links to external domains, top 20 domains

Most of the countries in the table are Latin American countries, but there are also links to large domains such as .COM or .DE. We took data from the exports promotion bureau of Chile, "ProChile" [Pro04], regarding the volume of exports from Chile to other countries, and we compared this with the number of links found. We took the top 50 countries that receive more exports from Chile. The USA –which is the largest destination of Chilean exports– was taken as the .COM domain. The results are shown in Figure 8.18.

There is a relationship between number of outgoing links and exports volume. The most important outliers are Asian countries, the outliers above the line have a high volume of exports but few links, probably because of a language barrier.

8.6 Links between Web sites

In the following, we consider links between Web sites. A link between two Web sites represents one or many links between their pages, preserving direction. Several links between pages are collapsed to a single link between Web sites, and self-links are not considered.

A summary of the characteristics of the links found is presented in Table 8.3.

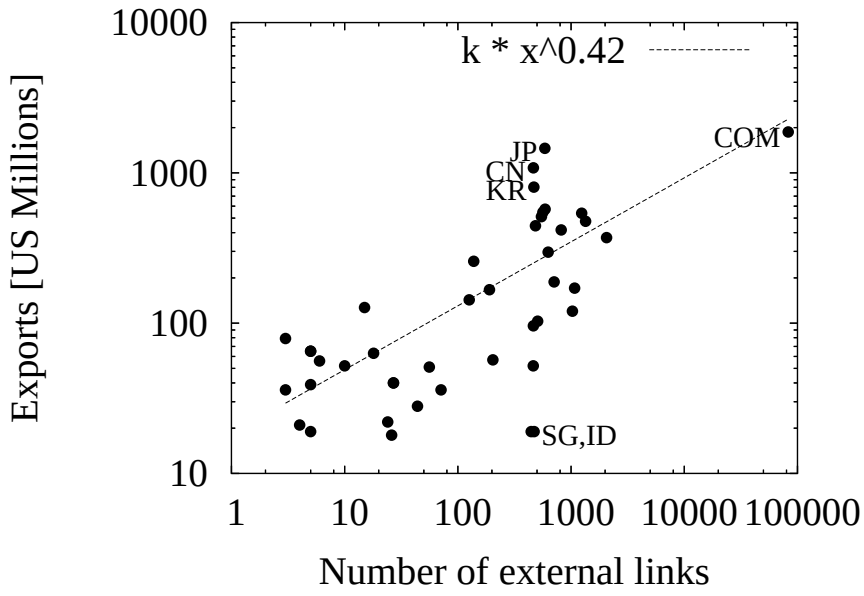


Figure 8.18: Relationship between number of external links from Chilean Web site and exports from Chilean companies to the top 50 destinations of Chilean exports.

Table 8.3: Summary of characteristics of links between Web sites.

Downloaded Web Sites	49,535	
At least one in-link	17,738	36%
At least one out-link	13,820	28%
At least one in-link or out-link	23,499	47%

8.6.1 Degree in the Web site graph

The distribution of in- and out-degree also reveals a scale-free network, as shown in Figure 8.19. The cumulative graphs consider only the Web sites with at least one in- or out-link respectively.

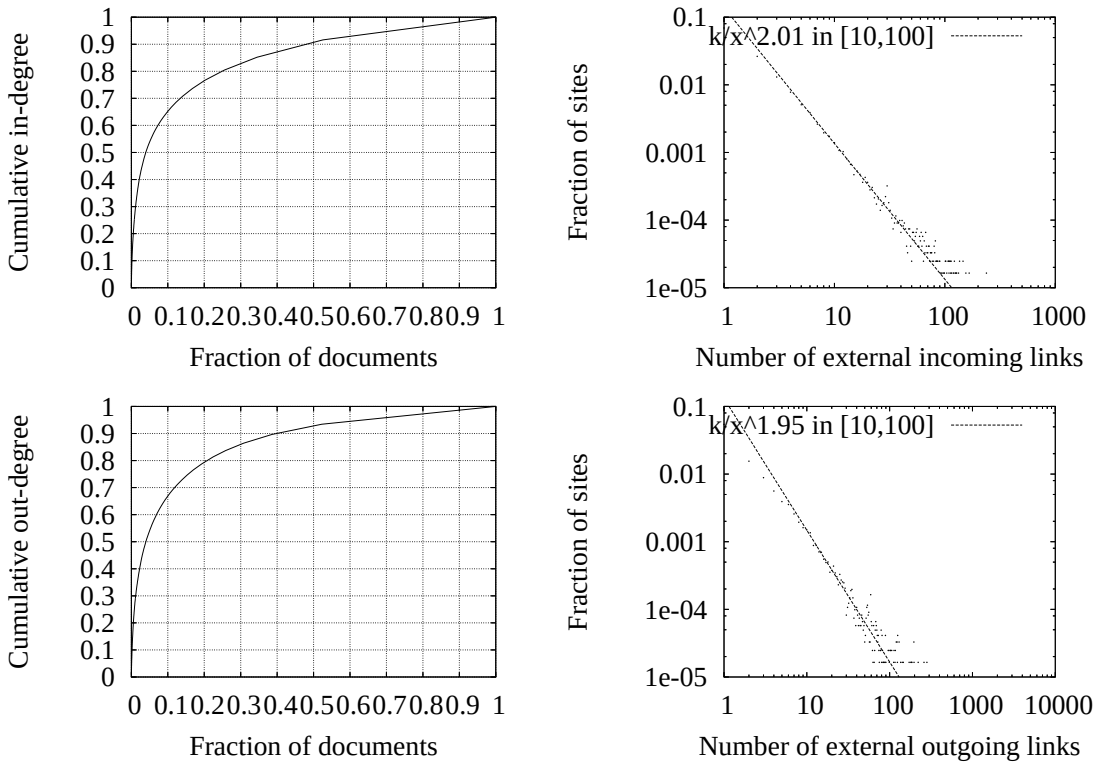


Figure 8.19: Distribution of in- and out-degree in Web sites.

8.6.2 Sum of link scores

We studied the link scores presented in Figure 8.17, and summed them by Web sites. The result is shown in Figure 8.20.

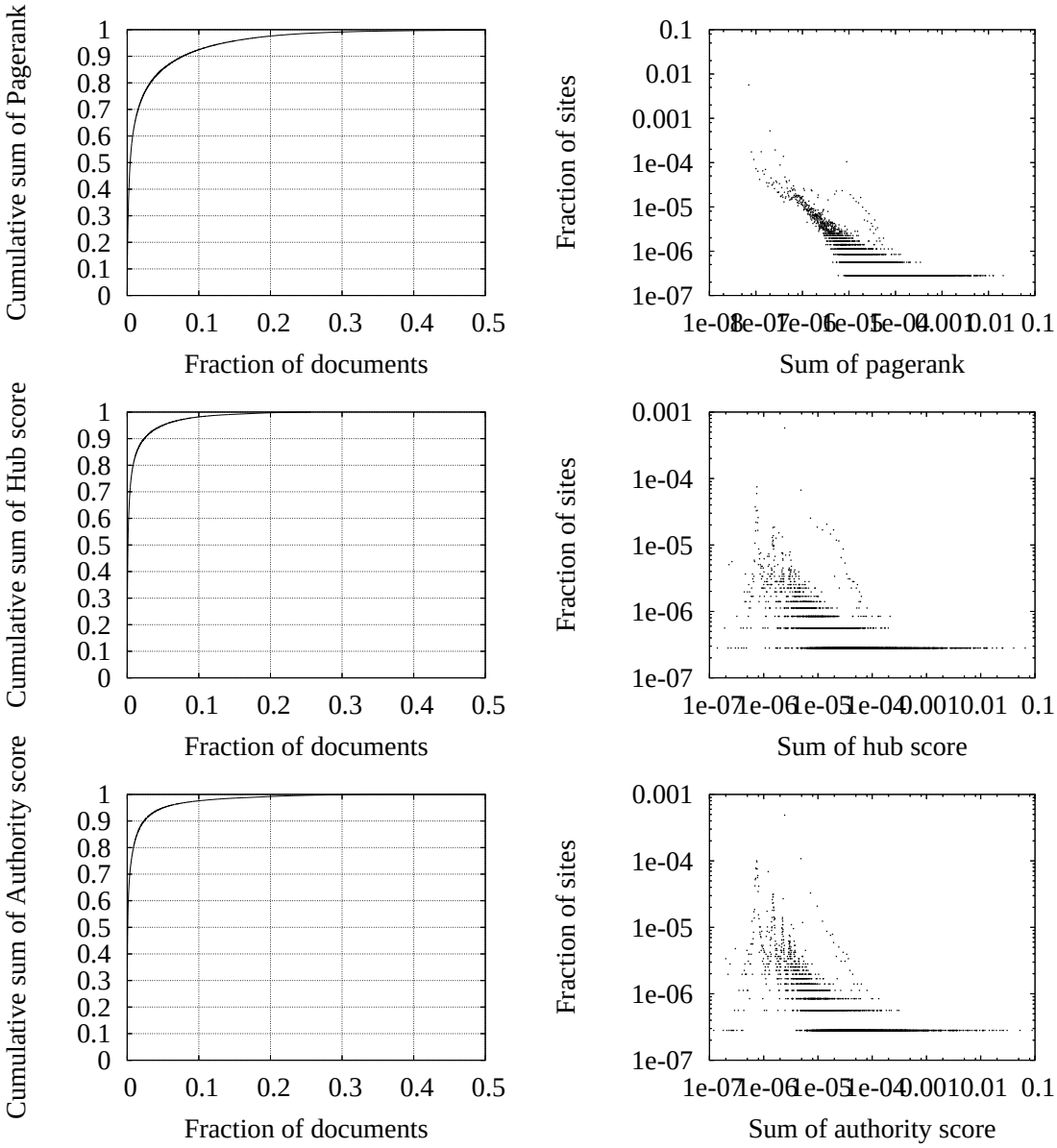


Figure 8.20: Distribution of Pagerank, global hubs and authority score in the graph of Web sites.

8.6.3 Most linked Web sites

The most linked Web sites on the Chilean Web are listed in table 8.4. There is a very strong presence of government-related Web sites in the top places, as well as universities.

Site name	Site type	Number of links
hits.e.cl	Access counter	675
www.sii.cl	Government (internal revenue service)	647
www.uchile.cl	University	595
www.mineduc.cl	Government (education)	513
www.meteochile.cl	Meteorology Service	490
www.emol.com	Newspaper	440
www.puc.cl	University	439
www.bcentral.cl	Government (bank)	404
www.udec.cl	University	366
www.corfo.cl	Government (industry)	354

Table 8.4: Most referenced Web sites, by number of in-links in the graph of Web site links.

8.6.4 Strongly connected components

We studied the distribution of the sizes of strongly connected components (SCC) on the graph of Web sites. A giant strongly connected component appears, as observed by Broder *et al.* [BKM$^+$00]. This is a typical signature of a scale-free network. The distribution of SCC sizes is presented in Table 8.5 and Figure 8.21, here we are considering only Web sites with links to other Web sites.

Component size	Number of components
1	17,393
2	283
3	54
4	20
5	4
6	3
7	2
8	2
9	2
10	1
5,202	**(Giant SCC) 1**

Table 8.5: Size of strongly connected components.

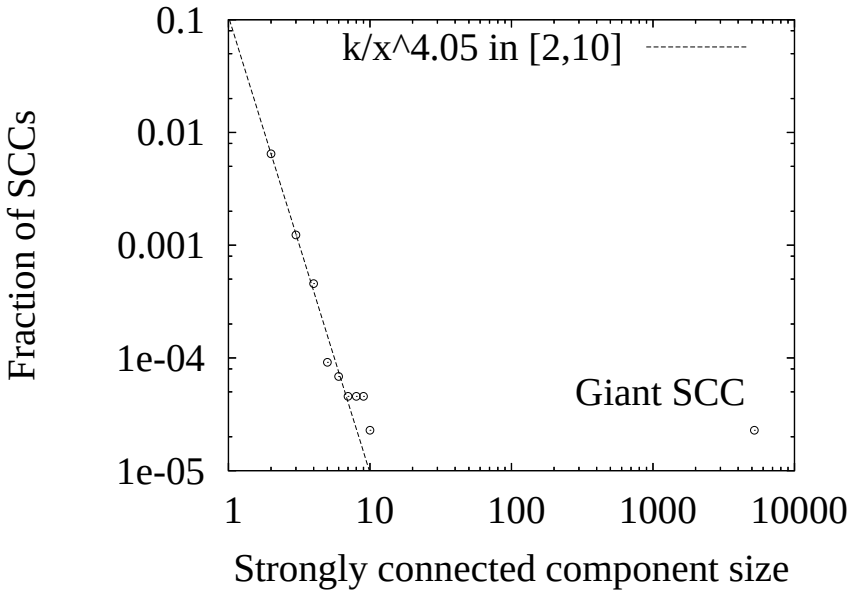


Figure 8.21: Distribution of strongly connected components.

8.6.5 Web links structure

In [BYC01] we extended the notation introduced by Broder *et al.* [BKM+00] for analyzing Web structure, by dividing the MAIN component into four parts:

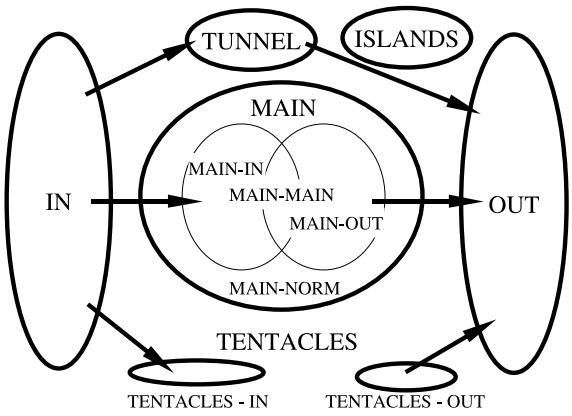
(e) MAIN-MAIN, which are sites that can be reached directly from the IN component and can reach directly the OUT component;

(f) MAIN-IN, which are sites that can be reached directly from the IN component but are not in MAIN-MAIN;

(g) MAIN-OUT, which are sites that can reach directly the OUT component, but are not in MAIN-MAIN;

(h) MAIN-NORM, which are sites not belonging to the previously defined subcomponents.

Note that the Web sites in the ISLANDS component are found only by directly accessing the home page of those Web sites. This is possible because we had a complete list of the registered domains under `.cl` at the time of our studies. The distribution of Web sites into components is shown in Figure 8.22. This structure evolves over time, as studied in [BYP03, BYP04].



Component name	Size
MAIN_NORM	2.89%
MAIN_MAIN	3.16%
MAIN_IN	1.20%
MAIN_OUT	3.26%
IN	7.23%
OUT	18.15%
TENTACLES-IN	2.75%
TENTACLES-OUT	4.23%
TUNNEL	0.33%
ISLAND	56.81%

Figure 8.22: Macroscopic structure of the Web.

.

8.7 Comparison with the Greek Web

We seek to understand to what extent the studies of the Chilean Web represent other subsets of the Web. Dill *et al.* [**?**] have shown that the Web graph is self-similar in a pervasive and robust sense. We compared some characteristics of the Chilean and the Greek Web, including the Web graphs but also other properties such as size or number of pages. The pages for this study were obtained simultaneously on the Greek and Chilean Web domains during January 2004.

We downloaded pages using a breadth-first scheduler for up to 5 levels for dynamically generated pages, and up to 15 levels for static, HTML pages. We limited the crawler to 20,000 pages per website; and considered only pages under the `.gr` and `.cl` domains.

Both countries are comparable in terms of the number of pages, but have many differences in terms of language, history, economy, etc. Table 8.7 summarizes information about the page collection, as well as some demographic facts that provide the context for this section.

	Greece	Chile
Population [Uni02]	10.9 Million	15.2 Million
Gross Domestic Product [The02]	133 US$ bn.	66 US$ bn.
Per-capita GDP, PPP [The02]	17,697 US$	10,373 US$
Human development rank [Uni03]	24^{th}	43^{th}
Web servers contacted	28,974	36,647
Pages downloaded	4.0 Million	2.7 Million
Pages with HTTP OK	77.8%	78.3%

Table 8.6: Summary of characteristics.

8.7.1 Web pages

Figure 8.23 shows the depth at which the pages of the collection were found; note that 5 is the limit we set for dynamic pages, as dynamic pages grows exponentially with depth. The distribution is almost identical.

Figure 8.24 shows the distribution of HTML page sizes, not considering images, showing a peak between 10 and 15 Kilobytes. The right-tail follows a power-law similar to previous results, and both distributions are very similar.

Figure 8.25 plots the number of pages per website. This has a very skewed distribution, as few websites account for a large portion of the total web; so we have plotted this in log-log scale.

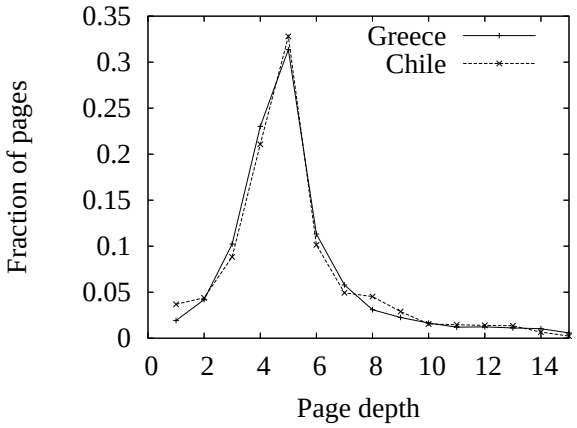


Figure 8.23: Comparison of page depth, 1 is the page at the root of the server.

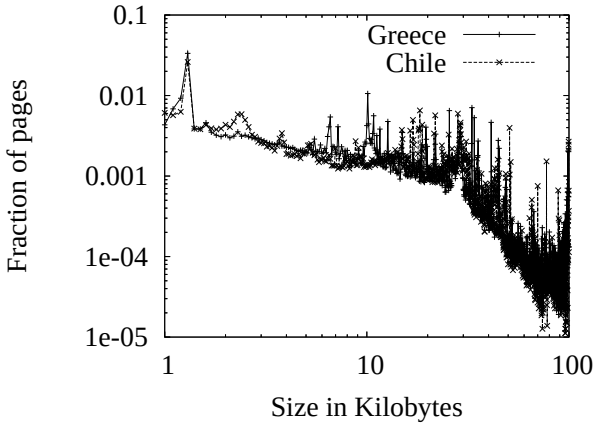


Figure 8.24: Comparison of the distribution of page size in Kilobytes.

8.7.2 Links

Figure 8.26 shows that these two sub-graphs of the web have these characteristics, revealing the existence of self-similarities. The power law parameter depends a lot on the range of data used. Taking degrees of at most 350, we obtain -2.02 and -2.11 for in-degree, and -2.17 and -2.40 for out-degree; for .GR and .CL, respectively. Discarding degrees smaller than 50, the parameters are closer to -2.3 and -2.8 for in-degree and out-degree. This should be compare with the results in [KRR$^+$00] that found -2.1 and -2.7, respectively, for 200 million pages in 1999.

The distribution of out-degree is different, as the in-degree in many cases reflects the popularity of a web page, while the out-degree reflects a design choice of the page maintainer. Also, it is much easier to have a page with many outgoing links than one with many incoming links.

For the graph components, we use the bow-tie structure proposed by Broder et al. [BKM$^+$00]; but we

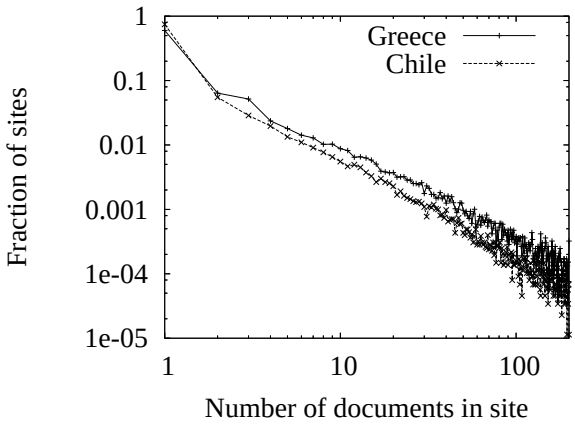


Figure 8.25: Comparison of the distribution of the number of pages per website.

considered only links between different websites, collapsing all the pages of a website to a single node of the graph. We show the relative size of components in Figure 8.27.

Note that that the `MAIN` (the giant strongly connected component) seems to be larger in the Greek web in expense of the `ISLAND` component - this can be an indicator of a better connected Web, although the seeds for the Chilean crawling had more islands.

We also studied the relationship of the collections with other top level domains reflecting cultural and economic relationships. The most linked domains (COM, ORG, NET, etc.) are equally important in both collections, but there are differences which are presented in Table 8.7. While the top linked Web sites for Greece are in Europe and Asia, for Chile they are mostly in America.

8.8 Conclusions

In this chapter, we have analyzed several characteristics of a large sample of the Web, and most of those characteristics suggest a distribution of quality that is very skewed. This is good for Web search, because only a few of the Web pages have some relevance, but this is also bad for Web crawling, because it is necessary to download large amounts of pages that are probably irrelevant.

World Wide Web users have a certain perception of how the World Wide Web is. This perception is based on what they see while interacting with the Web with the usual tool: a Web browser. The behavior of different users involves different parts of the Web, but in most cases it is limited to a few highly important Web sites with topics such as news, shopping or Web-based e-mail.

Most users do not go too deep inside Web sites. This means that there are thousands or millions of pages that are visited very rarely, or that are not visited at all. When characterizing the Web, we must forget what we have seen while browsing Web pages, because what we see through a Web browser is just the surface of

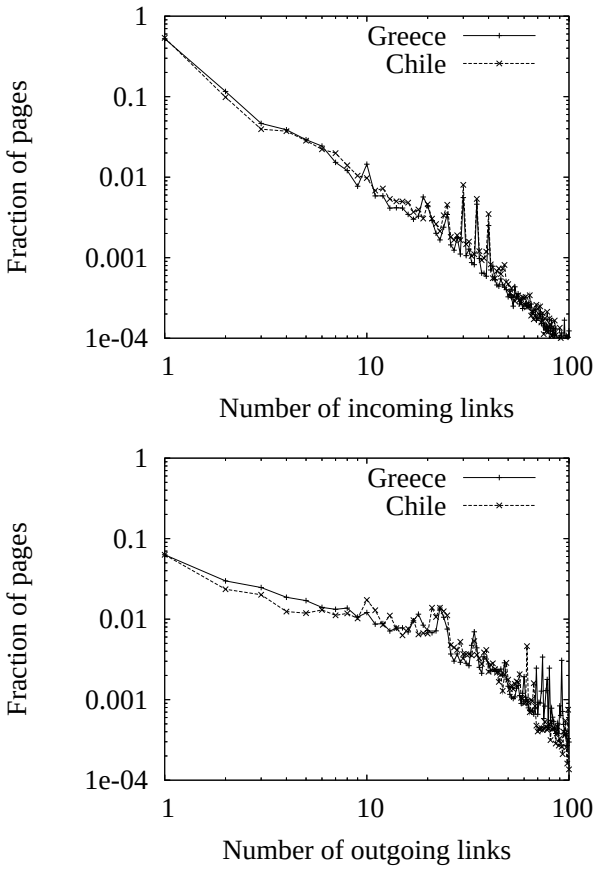


Figure 8.26: Comparison of the distributions of in-degree and out-degree.

something much deeper. For instance, there are very large and very small pages, pages with thousands of in-links and pages with only one, and so on.

Our results also show a dominance of standard formats such as PDF or plain text, and open source tools such as PHP and GZIP, which is quite natural given the open nature of the Web.

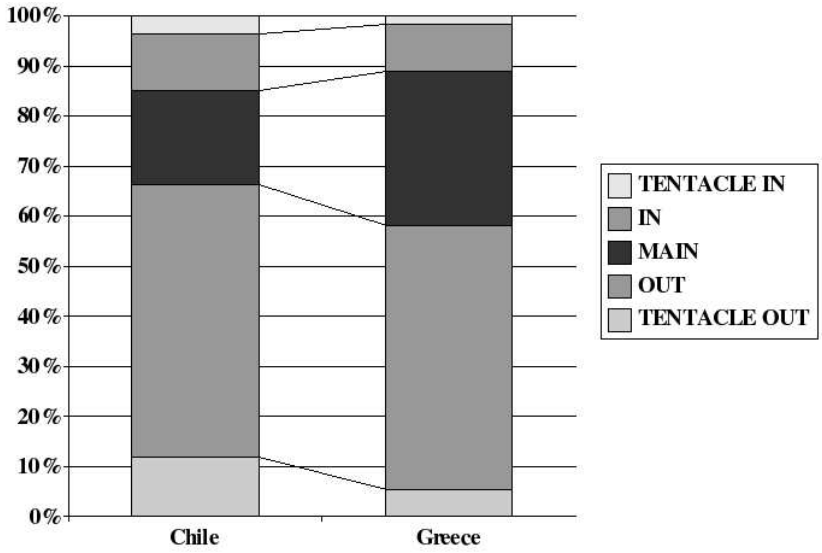


Figure 8.27: Comparison of the relative size of graph components.

Greece		Chile	
COM	49.2%	COM	58.6%
ORG	17.9%	ORG	15.4%
NET	8.5%	NET	6.4%
Germany	3.7%	Germany	2.6%
United Kingdom	2.6%	United Kingdom	1.4%
EDU	2.6%	EDU	1.3%
TV	1.3%	Mexico	1.2%
Russian Federation	1.3%	Brazil	1.1%
Taiwan	1.1%	Argentina	0.9%
Netherlands	0.9%	Spain	0.9%
Italy	0.8%	Japan	0.6%
GOV	0.6%	France	0.6%
Norway	0.6%	Netherlands	0.6%
France	0.5%	Italy	0.6%
Canada	0.5%	Australia	0.6%

Table 8.7: Comparison of the most referenced external top-level domains.

Bibliography

[asp04] Microsoft developer network - asp resources. http://msdn.microsoft.com/asp, 2004.

[Bar02] Albert-László Barabási. *Linked: the new science of networks*. Perseus Publishing, 2002.

[BKM+00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands, May 2000.

[BYC01] Ricardo Baeza-Yates and Carlos Castillo. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval*, pages 21–32, Laguna San Rafael, Chile, November 2001. IEEE CS Press.

[BYP03] Ricardo Baeza-Yates and Bárbara Poblete. Evolution of the Chilean Web structure composition. In *Proceedings of Latin American Web Conference*, pages 11–13, Santiago, Chile, 2003. IEEE CS Press.

[BYP04] Ricardo Baeza-Yates and Bárbara Poblete. Dynamics of the Chilean Web structure. In *Proceedings of the 3rd International Workshop on Web Dynamics*, pages 96 – 105, New York, USA, May 2004.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 57–65. IEEE CS Press, 2000.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The Pagerank citation algorithm: bringing order to the web. In *Proceedings of the seventh conference on World Wide Web*, Brisbane, Australia, April 1998.

[php04] Php - the hypertext preprocessor. http://www.php.net/, 2004.

[Pro04] ProChile. Estadísticas de exportaciones (statistics of exports). http://www.prochile.cl/-servicios/estadisticas/exportacion.php, 2004.

[The02] The Economist. Country Profiles, 2002.

[Uni02] United Nations. Population Division, 2002.

[Uni03] United Nations. Human Development Reports, 2003.