Comparing the Characteristics of the Korean and the Chilean Web

Ricardo Baeza-Yates*
Felipe Lalanne

Carlos Castillo
Georges Dupret

Korean-Spanish Information Retrieval Project
IT Cooperation Center Chile-Korea
University of Chile

Center for Web Research
Dept. of Computer Science
University of Chile

December 2004

Executive Summary

This report summarizes the results of a comparison between the characteristics of two public Web spaces: the pages under the .KR (South Korea) domain, and the pages under the .CL (Chile) domain. We show several similarities that contribute to validate more general models for the characteristics of the Web, specially in terms of link structure.

The key findings are:

- We downloaded 4.7 million pages from the `.kr` (Korea) domain and 3.2 million pages from the `.cl` (Chile) domain.
- Domains in Korea host on average 6 Web sites each, while in Chile, a domain has on average only 1.1 Web sites.
- Korean Web sites have almost 4 times as many pages as Chilean Web sites.
- Chilean Web pages are about 50% larger than Korean pages.
- The most used technology in Korea for dynamic pages is ASP, and in Chile, PHP.
- There is a strong preference of Microsoft formats for audio and video in the Korean Web.
- The distribution of links in terms of in-degree and out-degree for both collections are nearly the same.

*Corresponding author: `rbaeza@dcc.uchile.cl`.

1 Introduction

This report compares the .kr and .cl domains on December 2004. The Web pages were obtained using the WIRE crawler [2] during December 2005. 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 5,000 pages per website; and considered only pages under the `.kr` and `.cl` domains.

Table 1 summarizes information about the page collection, as well as some demographic facts that provide the context for this study.

Table 1: Summary of characteristics.

	Korea	Chile
Population (2000) [13]	46.8 Million	15.2 Million
Gross Domestic Product [11]	476 US$ bn.	66 US$ bn.
Per-capita GDP, PPP [11]	17,111 US$	10,373 US$
Human development rank [12]	30^{th}	43^{th}
Number of internet hosts [10]	269,788	219,250
Number of internet domains [10]	4,984	23,820
Web servers contacted	54,895	53,528
HTTP requests	8.9 Million	2.7 Million
Pages downloaded with HTTP OK	4.7 Million (53%)	3.2 Million (85%)
Average pages per site	224.2	57.6

Both countries are comparable in terms of the number of pages, but have many differences in terms of language, history, economy, etc. The Korean Web is larger than the Chilean Web, so the Chilean collection represents a sample that is more complete given the restrictions (maximum page depth and maximum pages per site) than the Korean collection. Nevertheless, according to the Internet Society's Internet Domain Survey [10] the number of Internet hosts of Korea is only 1.23 times the case of Chile. This does not take in account all hosts behind firewalls, which surely are many more in Korea.

Also, we had several connection errors when dealing with Korean Websites. This can be due to limitations in our connection, or to the fact that there are thousands of Web sites hosted on a few IPs, so those IP could have become saturated when crawling.

2 Sites per domain

The policies for assigning domains in both domains are different. In Chile, all domains under the `.cl` domain are available for registration, while in Korea, a domain must be bought only as a sub-domain of a predefined set: `.co.kr` for business, `.ac.kr` for academia, and so on. For this reason, for Korea, we considered `example.co.kr` as a domain, and `www.example.co.kr` as a site. For Chile, we considered `example.cl` as a domain, and `www.example.cl` as a site under that domain.

On average, there were about 6 sites per domain for Korean Web sites, and only about 1.1 sites per domain in the Chilean case.

Table 2 presents the 15 domains with the largest number of sites in both collections. Note that in Chile `.co`, `.corp`, `.gov` and `.gob` are just regular domains offered by ISPs (the first two) or used by some governmental agencies (`.gov & .gob`).

Table 2: Domains with the largest number of sites.

Korea		Chile	
Domain	Number of sites	Domain	Number of sites
ynara.co.kr	3264	tripod.cl	765
koryshop.co.kr	493	uchile.cl	333
yddr.co.kr	391	co.cl	301
legendofzu.co.kr	351	corp.cl	230
tprentacar.co.kr	308	scd.cl	224
marketinsight.co.kr	294	tie.cl	122
welcomebook.co.kr	293	puc.cl	122
modeunsadari.co.kr	284	utfsm.cl	107
proposeu.co.kr	227	ucv.cl	107
netbankkorea.co.kr	226	usach.cl	91
srbcafe.co.kr	222	udec.cl	75
saygoodby.co.kr	222	gov.cl	75
ykspong.co.kr	213	canal13.cl	73
ykcrystal.co.kr	212	utalca.cl	61
netgift.co.kr	208	gob.cl	57

3 Pages per site

On average there were 224.2 pages per site in the Korean Web and only 57.6 pages per site in the Chilean Web. Korean Web sites hold on average nearly 4 times as many pages as a Chilean site.

Figure 1 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.

4 Site size

The average site size considering all the downloaded pages was 1.1 MB for the Korean Web, and 7.9 MB for the Chilean Web. Figure 2 shows the distribution of site sizes.

The result of the two-sided Kolmogorov-Smirnov test of equality gives a p-value of 0.904, indicating that the distribution of page sizes are statistically equal, except for a scale factor.

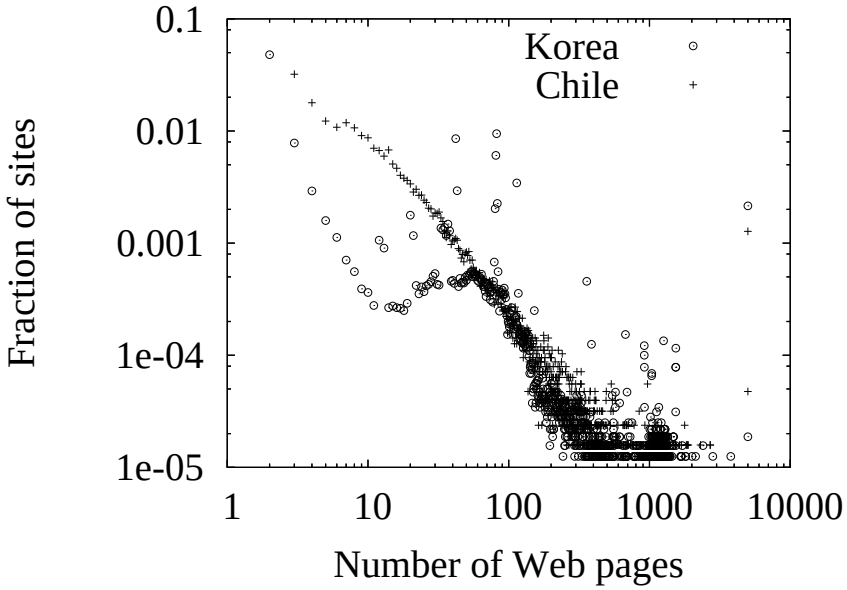


Figure 1: Number of pages per website.

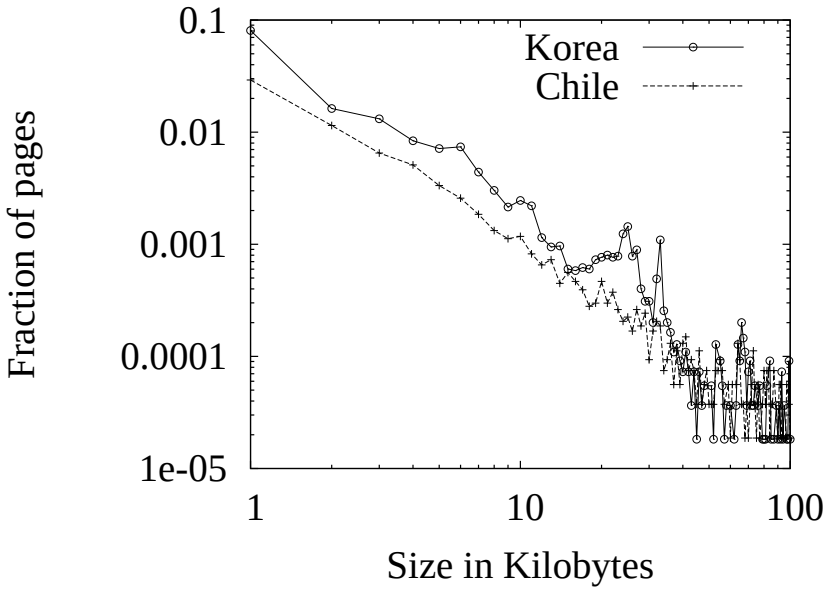


Figure 2: Histogram of site sizes, considering the total amount of data in HTML pages.

5 Page size

The average page sizes of the downloaded pages are 7.6 KB for the Korean Web, and 15.9 KB for the Chilean Web. After removing formatting, leaving only text and tags containing links (which we must preserve to

create the Web graph), the page sizes were reduced to 3.6KB (47% of the original size) and 5.7 KB (36% of the original size). This means that Chilean Web pages that we downloaded were larger on average than the Korean pages, and this was due more to HTML formatting than to larger textual content.

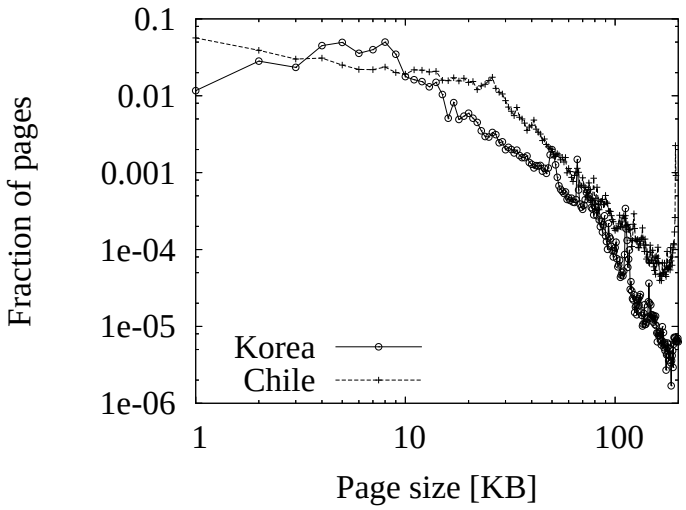


Figure 3: Histogram of page sizes.

The result of the two-sided Kolmogorov-Smirnov test of equality indicates that the two distributions are not equal, but as we saw in the previous section these differences tend to disappear when looking at entire Web sites.

Note: the increase in the frequency of pages with sizes close to 200Kb is an artifact of the way in which the maximum page length is enforced.

6 Dynamic pages

About 22% of the Korean pages and 35% of the Chilean pages that we downloaded were dynamic. We identified dynamic pages by recognizing common extensions such as `.CGI` or `.JSP` and by marking all pages with a "?" in the URL as dynamic. The technologies used for dynamic pages are shown in Figure 4.

In the case of the Chilean Web, the most used language is `PHP` (Hypertext Preprocessor) and in the Korean Web, it is `ASP` (Active Server Pages).

7 Multimedia files

We found 140,000 links to non-HTML documents in the Korean Web, and 370,000 in the Chilean Web. The difference is mostly due to the high usage of PDF files in the Chilean Web. The distribution for the other types is shown in Figure 5. In the Korean Web, there is a high usage of structured documents in XML, and less use of Microsoft Office's `.doc` documents.

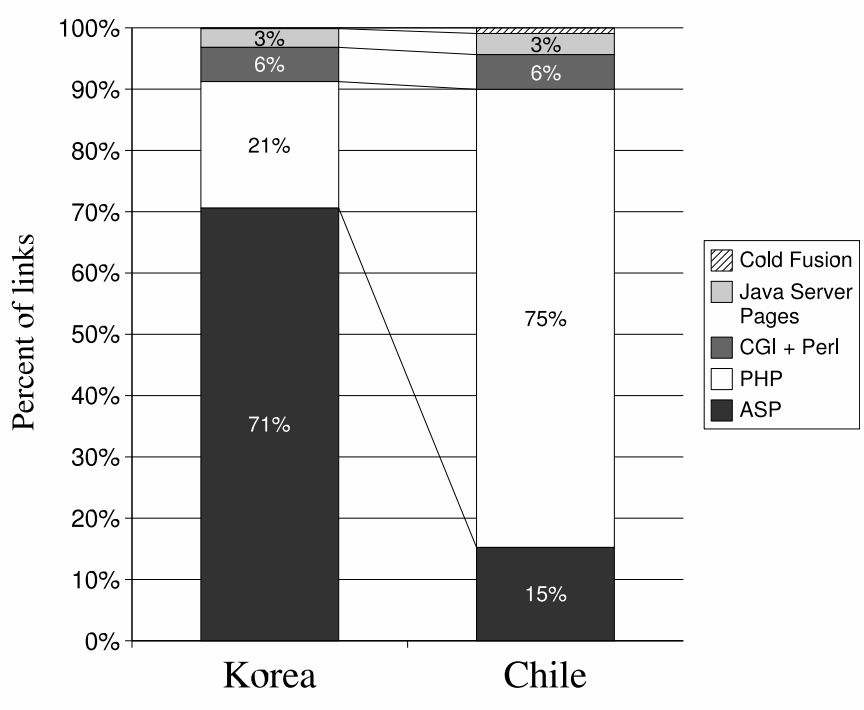


Figure 4: Distribution of extensions for dynamic pages.

Figure 6 shows the distribution of links to image files. Compuserve's GIF format is the most used format, followed by JPEG.

Figure 7 shows the distribution of links to audio and music files. In the Korean Web there is a higher usage of Windows Media formats such as `WMA` and `ASF`, while in the Chilean Web the `MP3` format is preferred. In both cases Realaudio files are also used, mostly because of their support for streaming.

Figure 8 shows the distribution of links to video files. In the case of the Korean Web, the Windows Media formats such as `WMV` is preferred, while in the Chilean case there is no clear winner, with several competing formats. We also found over 1.7 million links to Flash animations in the Korean Web and about 670,000 in the Chilean Web.

Figure 9 shows the distribution of links to compressed files. The two most used formats are `GZ` and `ZIP` in both data sets.

8 Links between pages

The web graph is usually characterized as a random graph created by a process of preferential attachment [8], that induces scale-free properties [5].

Figure 10 shows that these two sub-graphs of the Web have these characteristics, revealing the existence of self-similarities. The power law parameter varies with the range used, and it is roughly -2.0 to -2.1 for the center part of the distribution of in-degree, and about -3.0 to -3.3 for the distribution of out-degree.

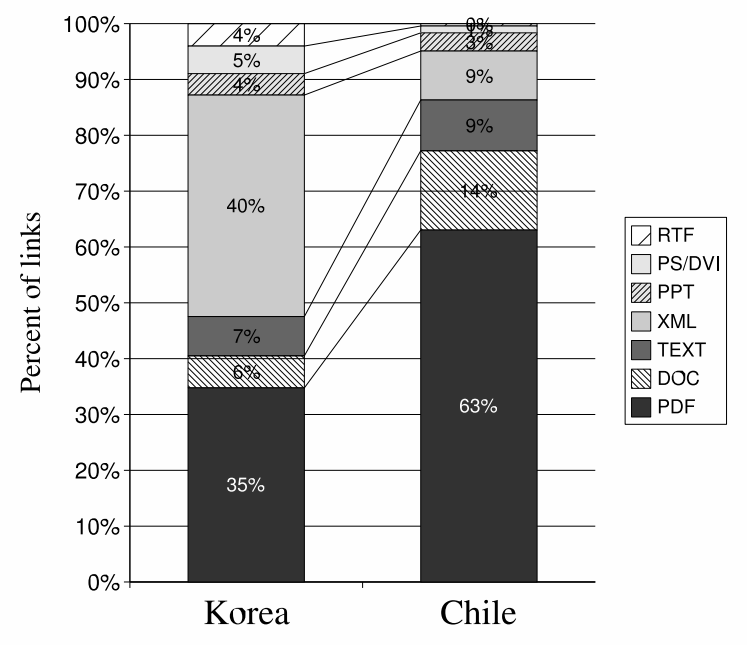


Figure 5: Distribution of file types used for documents, excluding HTML pages.

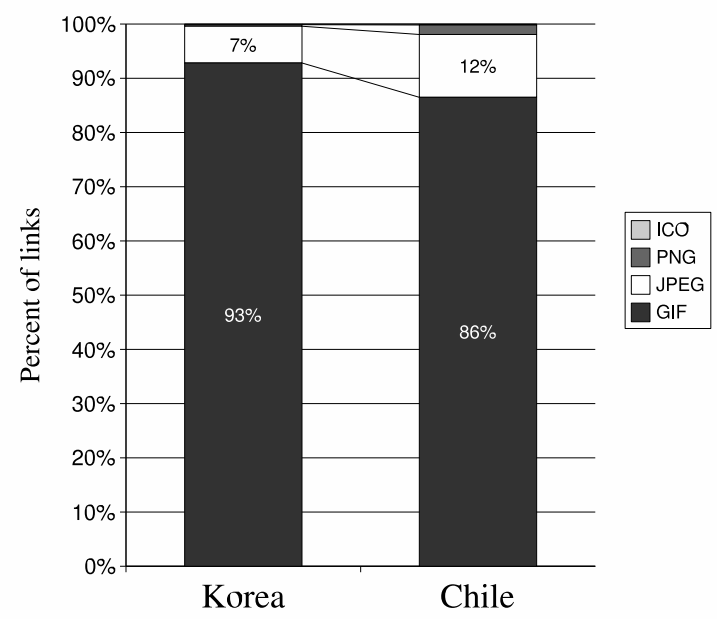


Figure 6: Distribution of file types used for images.

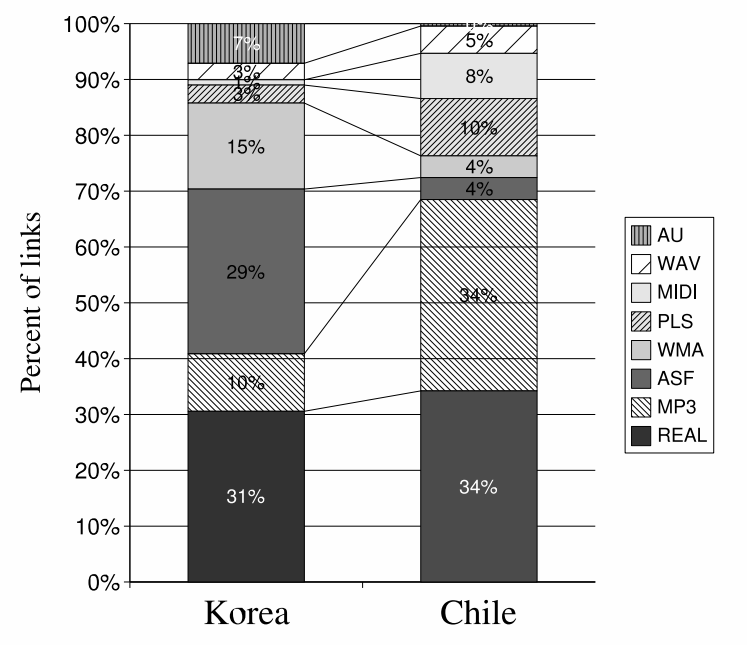


Figure 7: Distribution of fi le types used for audio and music.

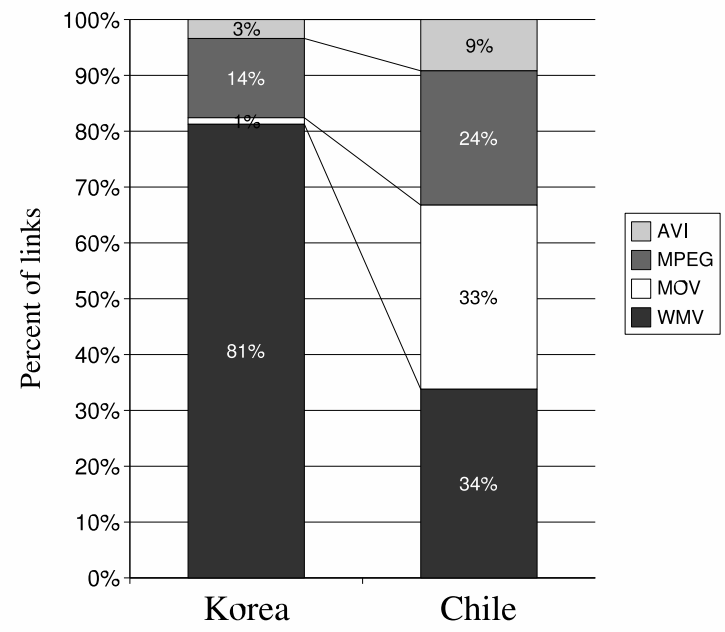


Figure 8: Distribution of fi le types used for video.

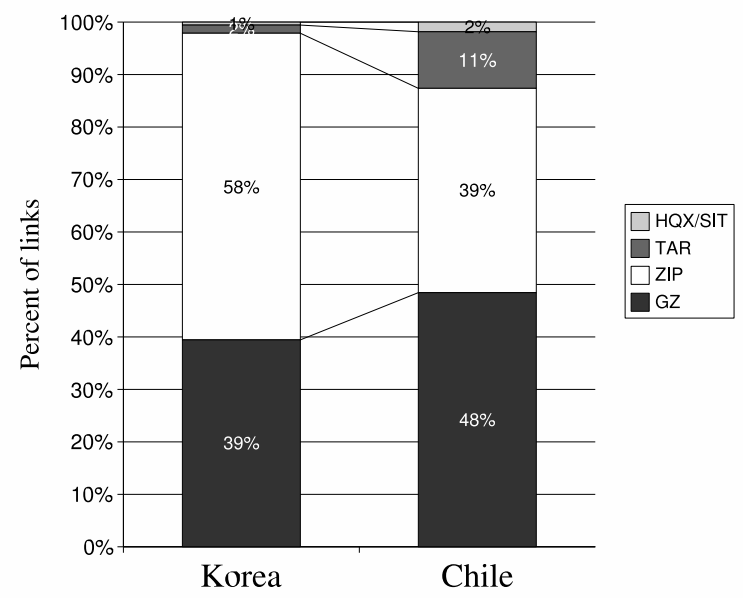


Figure 9: Distribution of file types used for compressed files.

For an estimation of the power law parameter across the entire range of observations, we used a weighted linear regression that takes the number of observations per degree into account. The slope for the in-degree was -1.75 and -1.79 for Korea and Chile, and for the out-degree it was -1.56 and -1.52 for Korea and Chile. These differences are not significant so we conclude that both distributions can be considered statistically the same.

The distribution of out-degree is different than the distribution of 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. Also note that in the case of out-degree, there are less pages with between 1 and 10 links than predicted by the model, so the model fit is better for the in-degree.

9 Link-based scores

We used a static version of Kleinberg's HITS algorithm [7] to calculate Hubs and Authorities in both Web collections. For this we considered only links across Web sites (internal links do not confer authority). We also calculated the Pagerank score [9] for each page in the collection.

The power-law parameters obtained are summarized in Table 3. In all three cases the statistical analysis concluded that the distributions follow the same law but with different parameters. The details about how these parameters were obtained are provided in the rest of this section.

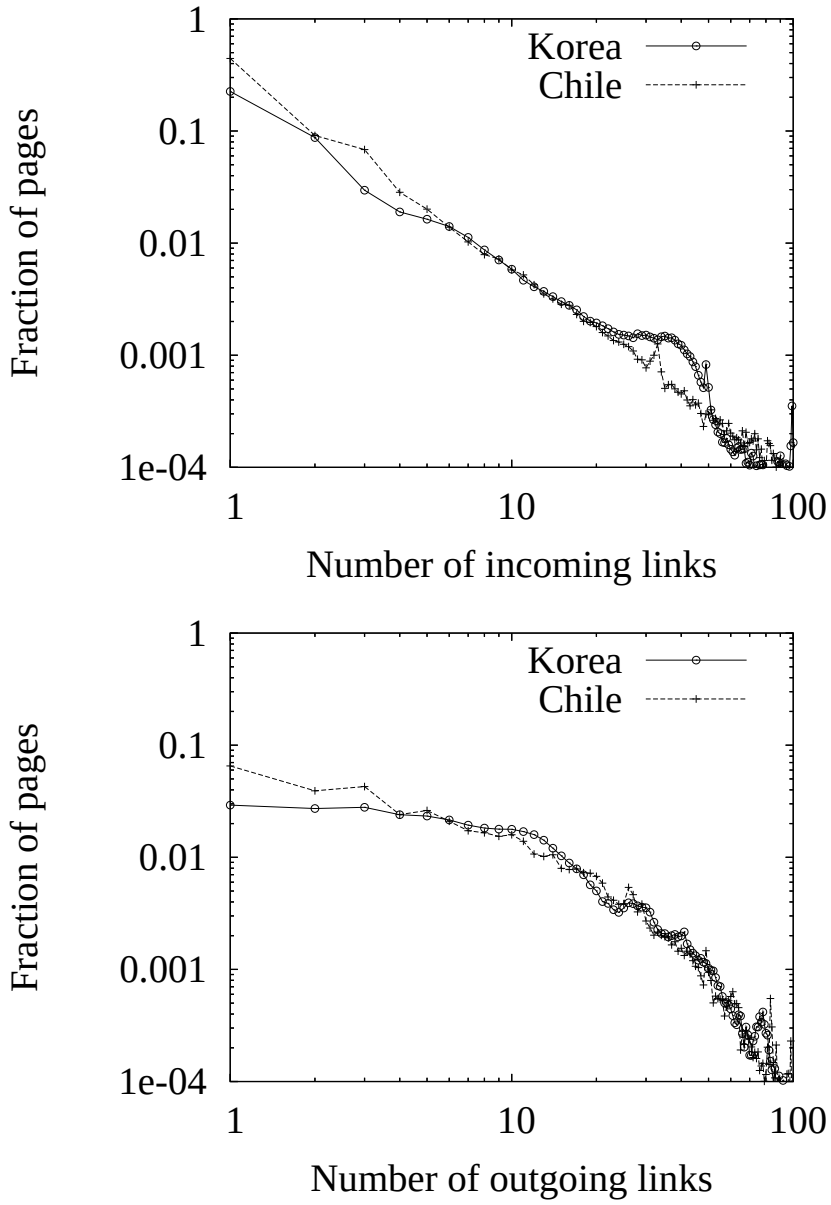


Figure 10: Distribution of in-degree and out-degree.

9.1 Hub score

We test the power law model on the part of the data that seems adecuate. For Chile, we keep the data larger than a threshold of $7.5\ 10^{-6}$ that corresponds to 2375 observations out of a total of 2447 (97%). For Korea, the threshold is $3.\ 10^{-7}$ and we keep 1194 out of 1269 observations (94%).

Table 3: Parameters for the power-law distribution of link scores in both samples.

Score	Korea	Chile
Hub	-2.36	-1.93
Authority	-1.00	-0.96
Pagerank	-2.02	-2.25

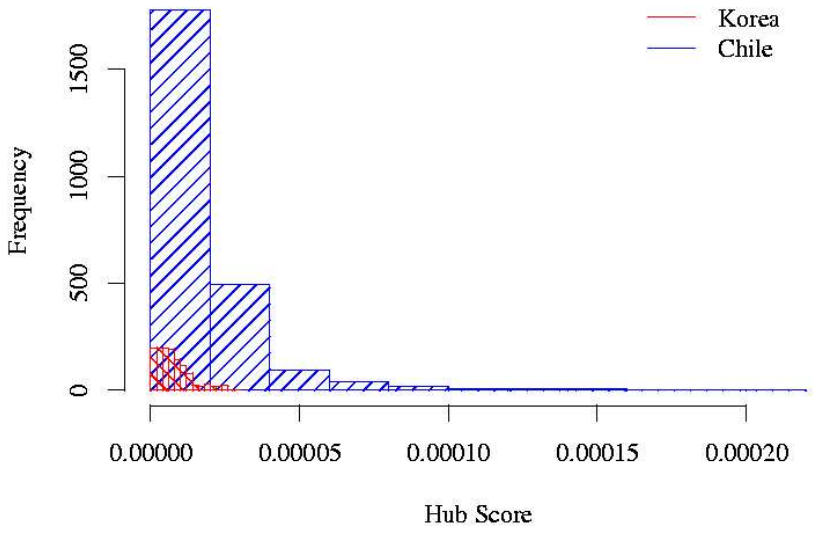


Figure 11: Histograms of hub scores in Chile and Korea. Hub scores are much smaller in Korea than in Chile.

The regression on Chile log-log data gives:

$$\text{Num. of documents} = \exp(-19.97707)(\text{Hub Score})^{-1.92720}$$

The regression on Korea log-log data gives:

$$\text{Num. of documents} = \exp(-26.30959)(\text{Hub Score})^{-2.36412}$$

The slopes and intercepts are significantly different between Chile and Korea.

Including all the points for the regression, the intercept and slope for Chile become -15.734231 and -1.548429 to be compared with -19.97707 and -1.92720 respectively. For Korea, keeping all observations leads to -24.129711 and -2.181706 instead of -26.30959 and -2.36412.

9.2 Authority score

The regression for the Korean data set gives:

$$\text{Num. of documents} = \exp(-8.802711)(\text{Authority Score})^{-0.998791}$$

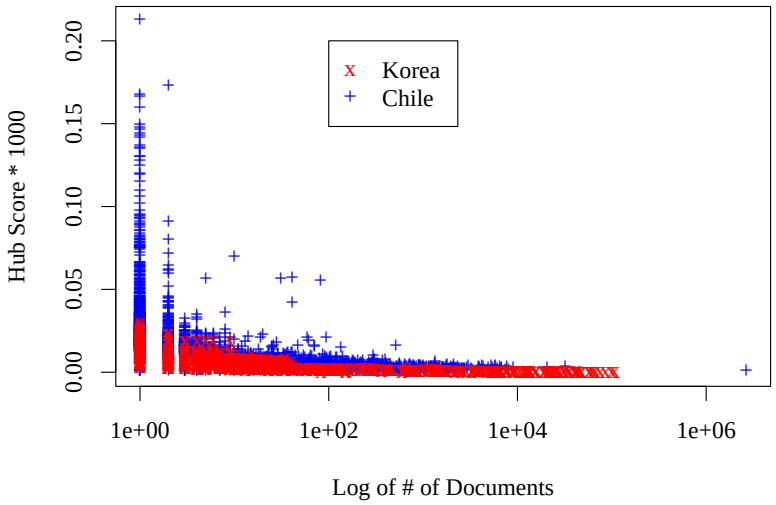


Figure 12: Number of documents per hub score in Chile and Korea

The regression for the Chilean data set gives:

$$\text{Num. of documents} = \exp(-8.558362)(\text{Authority Score})^{-0.956341}$$

We see that the slope is between 2 and 3 standard deviation, and we must reject the equality under a significance level of 5%.

We attempt to fit a power law on the part of the data that seems to follow a straight line on the log-log plot. Because the decision to include a particular observation in the transition region (authority $\in [10^{-7}, 10^{-6}]$) is partly arbitrary, we use different subset of observations to test the robustness of the coefficients. In conclusion, the slopes appear to be statistically significantly different, although their numerical values are similar.

9.3 Pagerank

We see on log-log plots that Pagerank for both countries follow quite well the power-law model. The obtained equations are very similar.

Results of the fit for Korea:

$$\text{Num. of documents} = \exp(-20.5576)(\text{Pagerank Score})^{-2.0203}$$

Results of the fit for Chile:

$$\text{Num. of documents} = \exp(-23.9387)(\text{Pagerank Score})^{-2.2491}$$

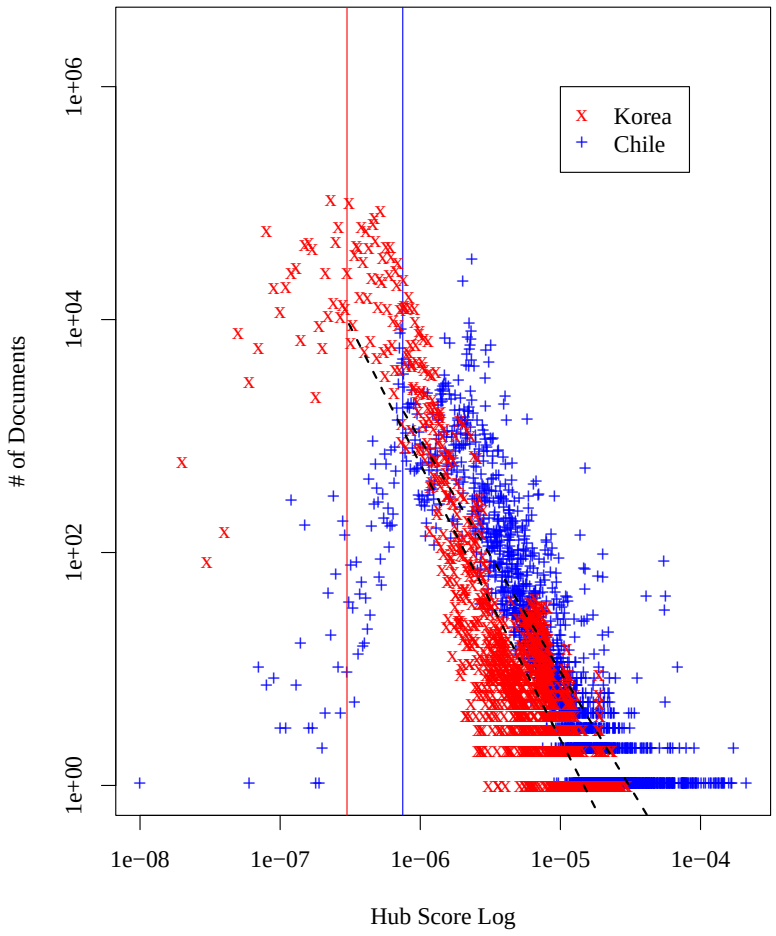


Figure 13: Log-log plot of number of documents versus hub score in Chile and Korea. A power law is fitted on the decreasing part of the data. The thresholds for Korea ($3.\ 10^{-7}$) and Chile ($7.5\ 10^{-6}$) are selected graphically. The model is weighted by the number of observations.

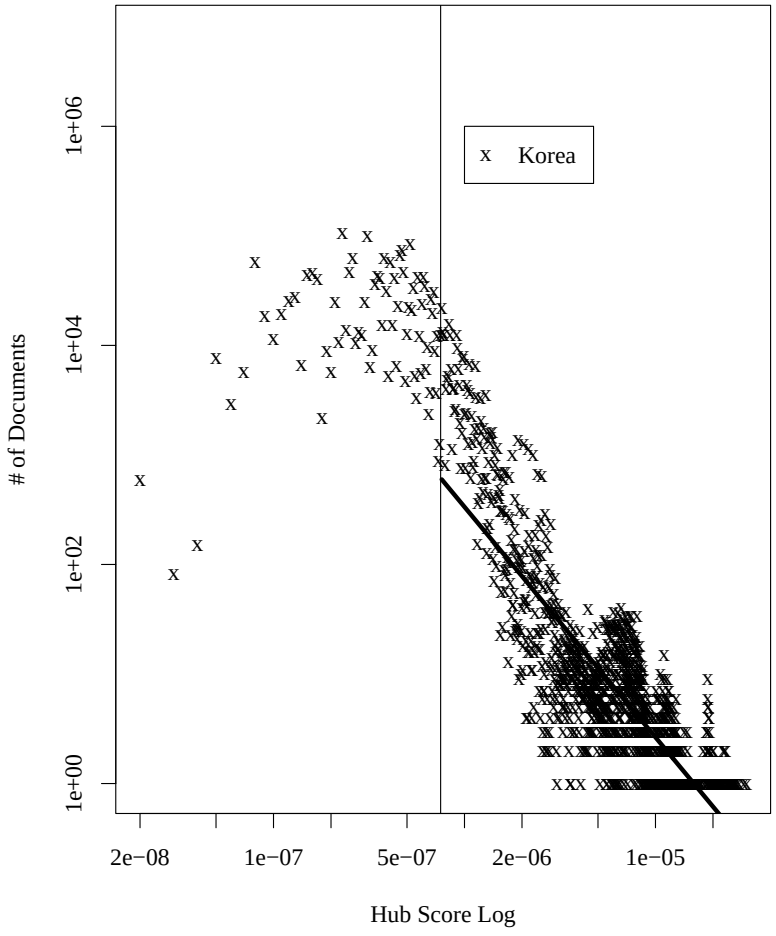


Figure 14: Log-log plot of number of documents versus hub score in Korea.

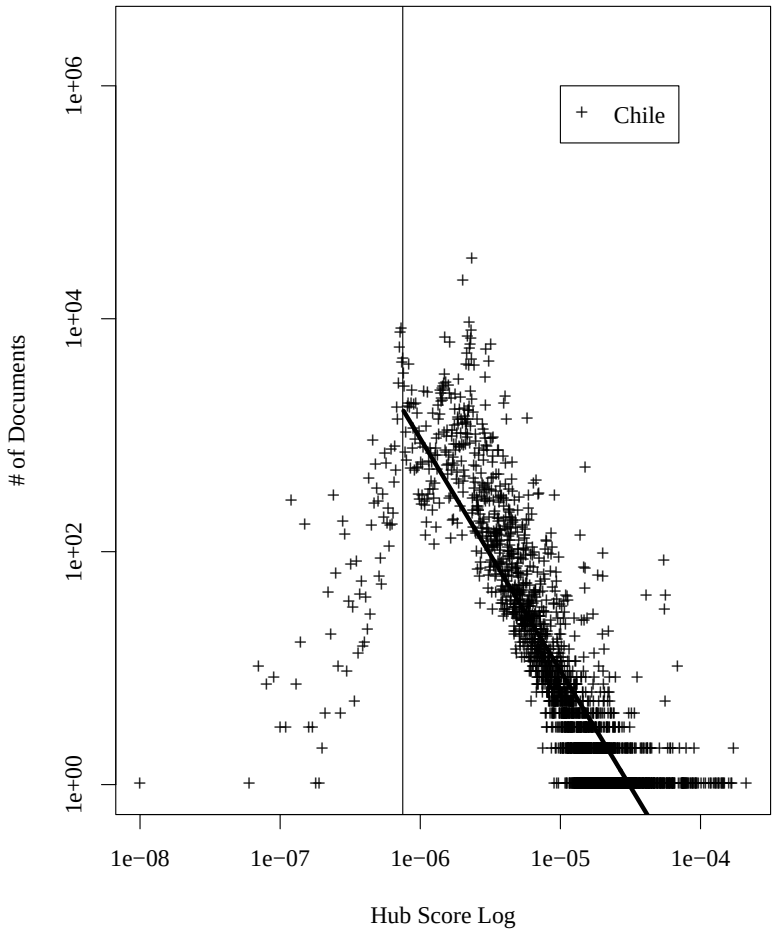


Figure 15: Log-log plot of number of documents versus hub score in Chile. A power law is fi tted on the decreasing part of the data.

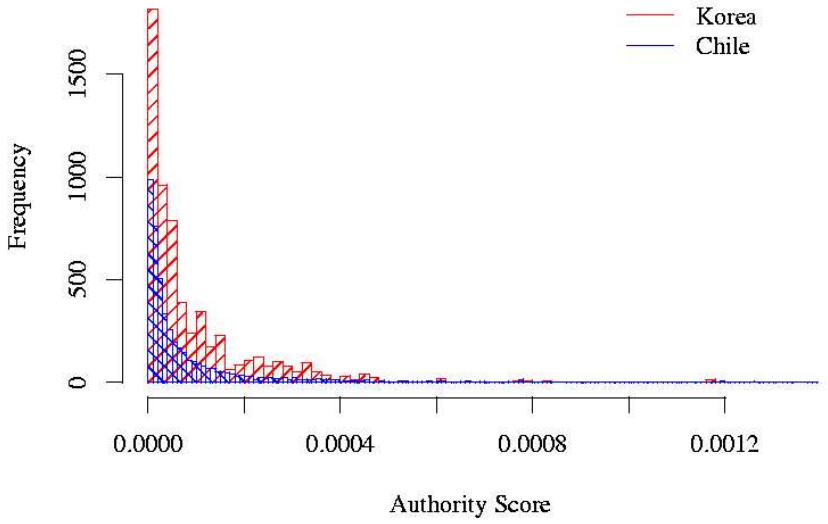


Figure 16: Authority histograms. Note that authority scores range from 0 to 0.04413966 in Chile and 0 to 0.0013334 in Korea. The Chilean values have been troncated to ease visualization. Korean authority scores are more concentrated in low values than Chile that is spread sparsely over a large range.

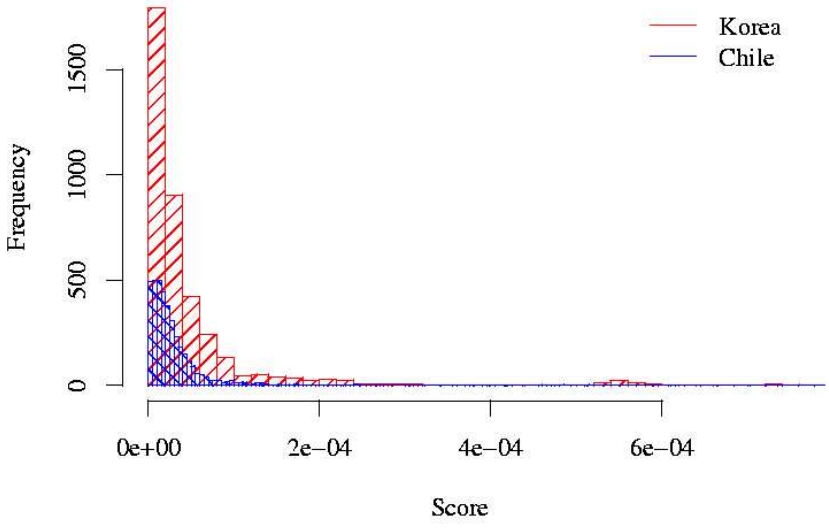


Figure 17: Histogram of Pageranks in Korea and Chile. Korean pages reach larger Pagerank scores than Chilean one.

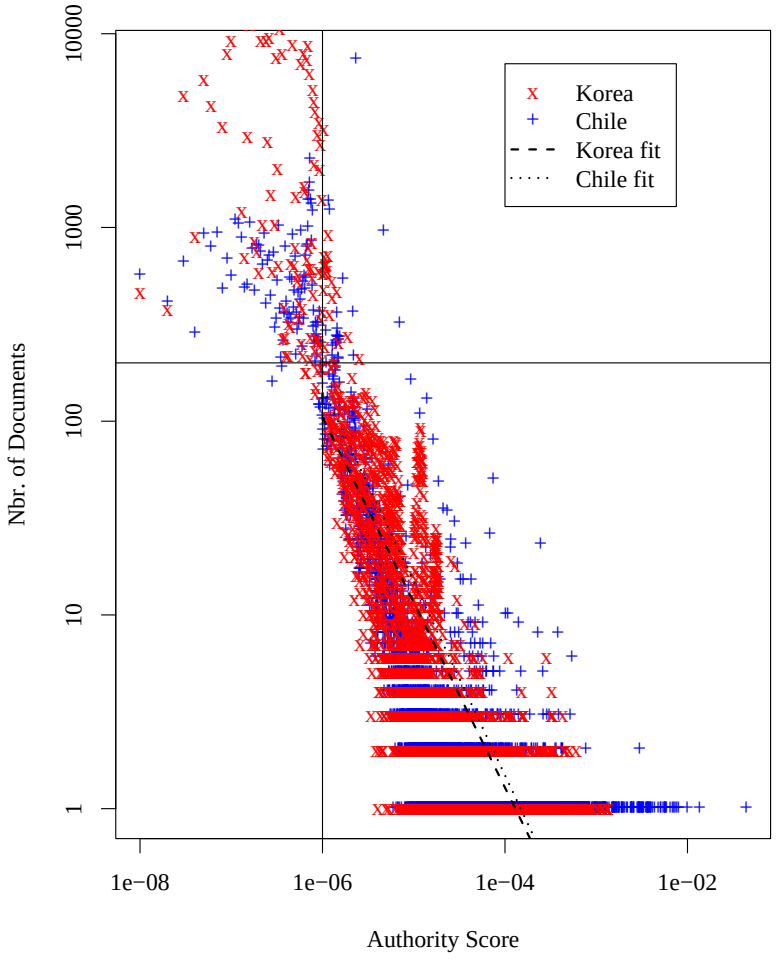


Figure 18: Power law fi tting of Authority. We restrict consideration to the observations where authority is larger than 10^{-6} and the nbr. of documents larger than 200. The linear model is weighted by the number of observations. Slopes are signifi cantly different.

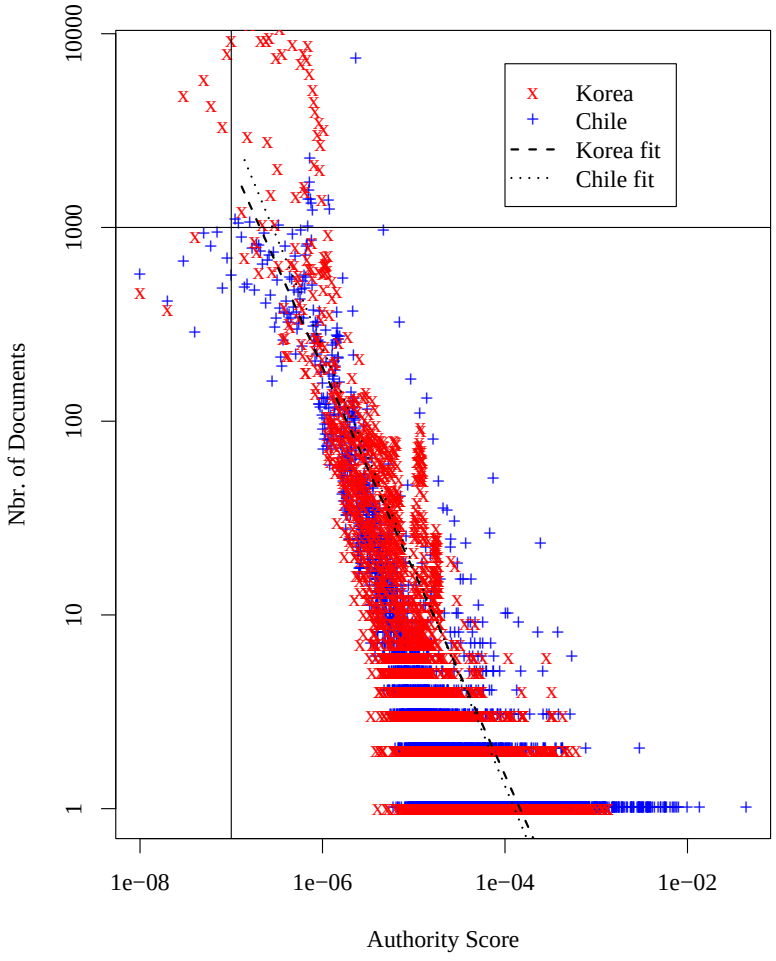


Figure 19: Power law fi tting of Authority. We restrict consideration to the observations where authority is larger than 2. 10^{-7} and the nbr. of documents larger than 1000. The linear model is weighted by the number of observations. Slopes are signifi cantly different.

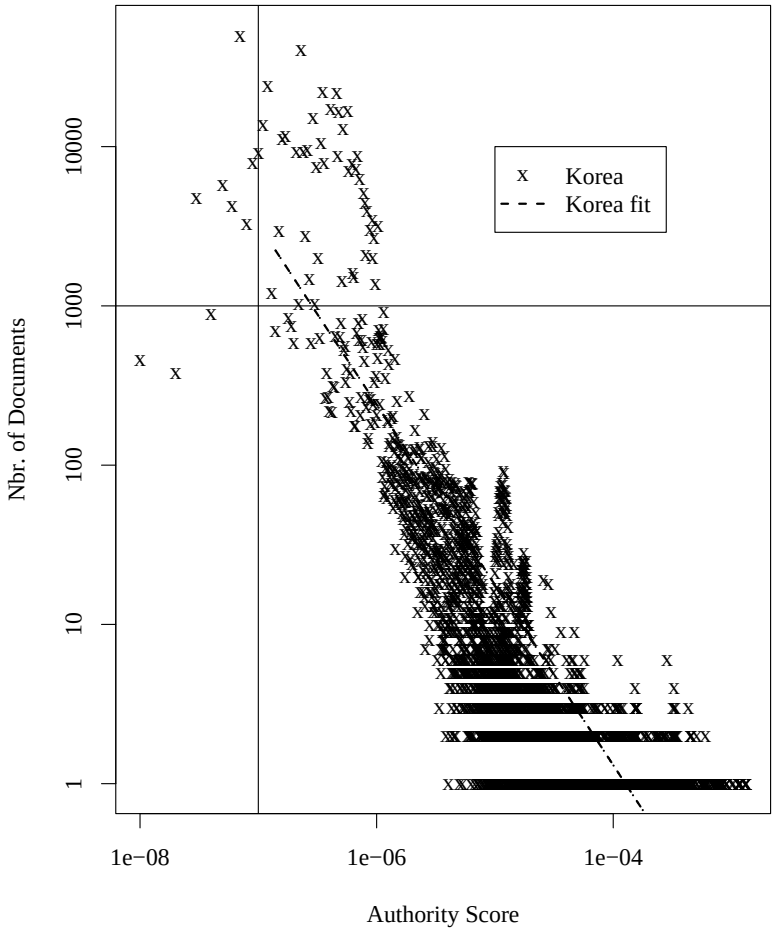


Figure 20: Detail of the power law fi tting of authority score for Korea. We restrict consideration to the observations where authority is larger than 10^{-7} and the nbr. of documents larger than 1000. The linear model is weighted by the number of observations.

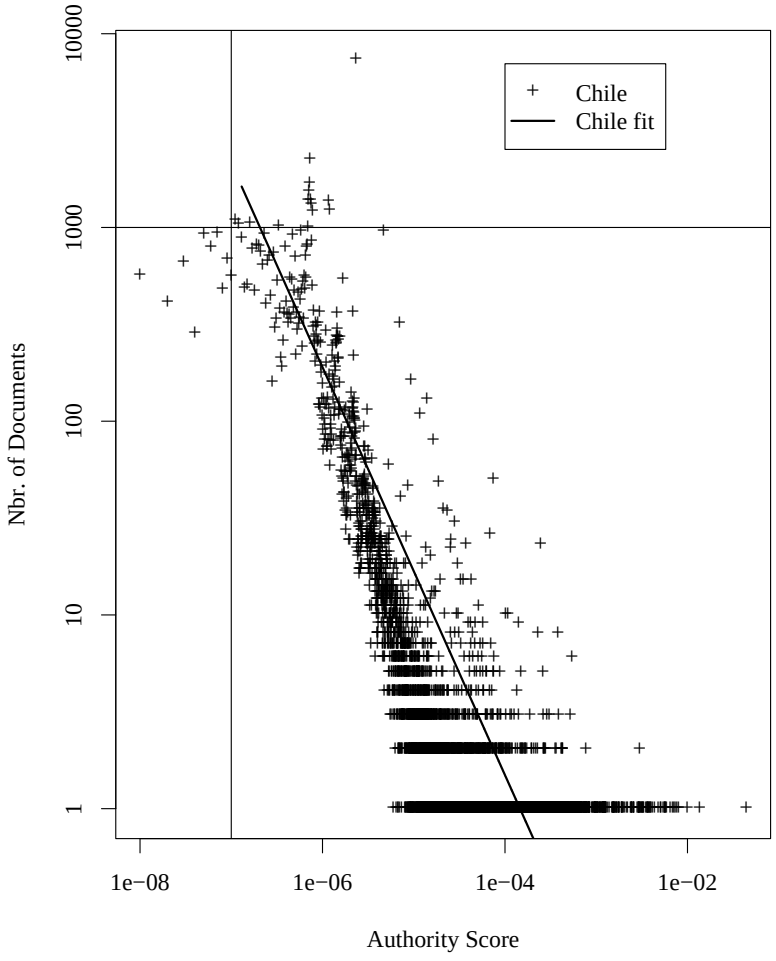


Figure 21: Detail of the power law fi tting of authority score for Chile. We restrict consideration to the observations where authority is larger than 10^{-7} and the nbr. of documents larger than 1000. The linear model is weighted by the number of observations.

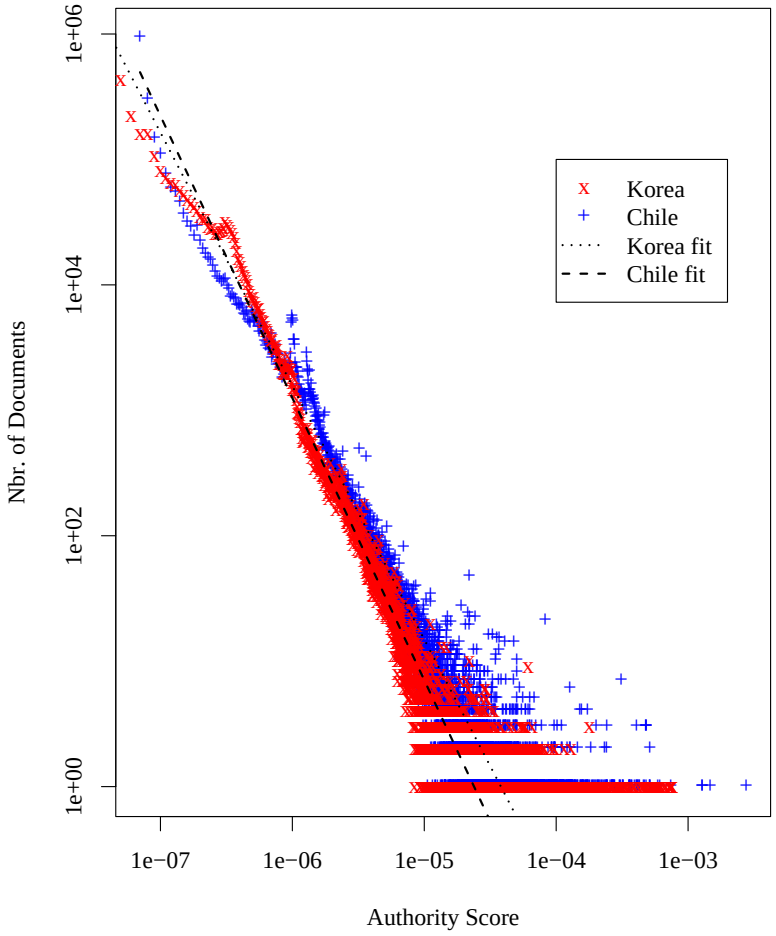


Figure 22: Power law fi tting of Pagerank: Chile. The linear model is weighted by the number of observations. Slopes are signifi cantly different.

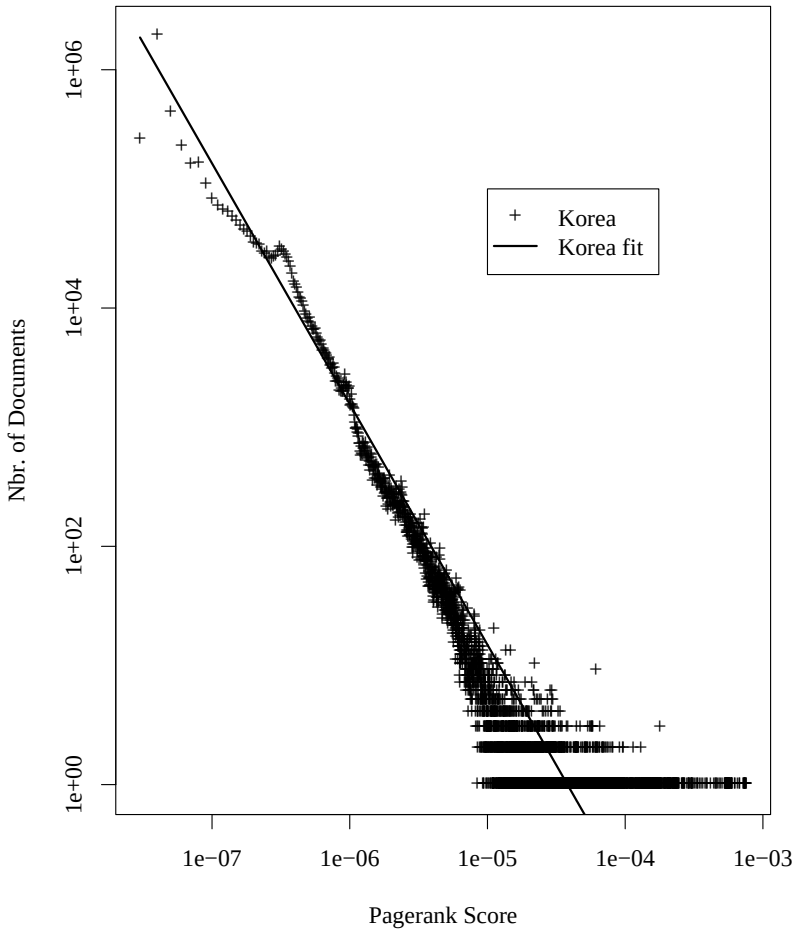


Figure 23: Detail of the power law fi tting of Pagerank in Korea. The linear model is weighted by the number of observations.

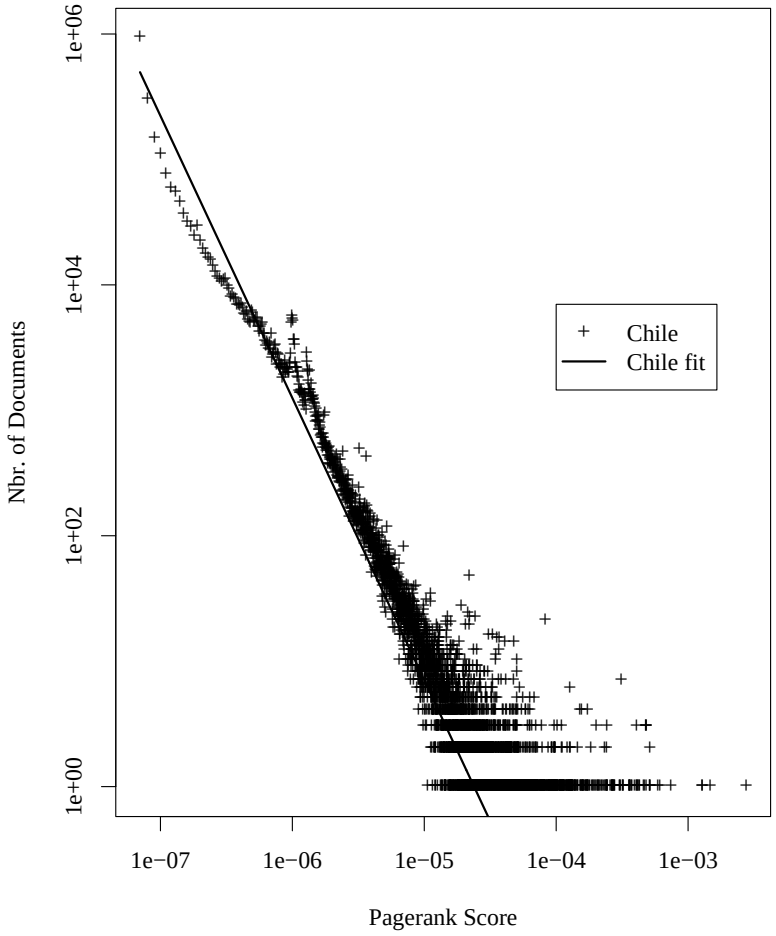


Figure 24: Detail of the power law fi tting of Pagerank in Chile. The linear model is weighted by the number of observations.

10 External links

We also studied the relationship of the collections with other top level domains, this is summarized in Table 4. The `.to` domain is a popular destination of links (used by sites like `go.to` or `welcome.to`).

Table 4: Most referenced external top-level domains.

Korea		Chile	
Domain	Percent	Domain	Percent
COM	79.8%	COM	67.9%
NET	15.1%	NET	8.8%
ORG	1.9%	ORG	10.9%
TO – Tonga	1.3%	ES – Spain	1.7%
INFO	0.2%	AR – Argentina	1.5%
JP – Japan	0.1%	DE – Germany	1.0%
BIZ	0.1%	UK – United Kingdom	0.8%
EDU	0.1%	BR – Brazil	0.6%
UK – United Kingdom	<0.1%	EDU	0.6%
CA – Canada	<0.1%	MX – Mexico	0.6%

Note that the table considers just the total number of links, without counting duplicates. We found one site with thousands of links to a single domain in `.be` and one site with thousands of links to a single domain in `.st` and we discarded those from the table.

11 Links between sites

We considered that site `A` is linked to site `B` if `A` and `B` are two different Web sites, and at least one page in site `A` links to one page in site `B`. This produces a graph of links between sites.

Table 5 shows the Web sites with the larger number of in-links in this graph, this is, the larger number of other sites linking to it. Note that multiple links from the same site are not counted.

The rest of this section studies the fitting of a power-law to the distribution of links between sites.

While the Chilean sites follow quite well the power-law model, the situation in Korea is quite different. The obtained equations are given below.

Result of the fit for Korea:

$$\text{Num. of documents} = \exp(8.446091)(\text{In-links})^{-1.003357}$$

Result of the fit for Chile:

$$\text{Num. of documents} = \exp(9.44318)(\text{In-links})^{-1.68765}$$

The increase in the number of sites with more than 50 other sites pointing to them in the graph of the Korean Web, seems to be a signal of collusion between spam Web sites, but this requires further checking.

Table 5: Sites with more references inside each domain. Only references from different Web sites in the same top-level domains are counted.

Korea		Chile	
Domain	In-links	Domain	In-links
www.jeandes.co.kr	3,618	www.sii.cl	715
www.100junyun.or.kr	3,579	www.uchile.cl	687
www.handlemat.co.kr	3,392	hits.e.cl	649
www.ibidder.co.kr	3,298	www.mineduc.cl	612
www.hieye21.co.kr	3,280	www.meteochile.cl	569
www.churchweb.or.kr	3,253	www.tripod.cl	502
www.ulsi.co.kr	3,221	www.puc.cl	486
www.chajungbi.co.kr	3,107	www.google.cl	448
www.btwo.co.kr	3,055	www.bcentral.cl	437
kim.ss3002.co.kr	3,011	www.udec.cl	433
www.sanjae114.co.kr	2,969	www.terra.cl	433
www.chamdiet.co.kr	2,964	www.corfo.cl	422
www.timebytime.co.kr	2,848	www.conicyt.cl	390
www.jejustory.co.kr	2,787	www.gobiernodechile.cl	383
www.liebewed.co.kr	2,751	www.latercera.cl	376

12 Site outgoing Links

Like for incident links, the Chilean outgoing links follow quite well the power-law model while the situation in Korea is quite different. Note the formal similarity of distributions between Korea in and out-going site links.

Result of the fit for Korea:

$$\text{Num. of documents} = \exp(8.356374)(\text{Out-links})^{-0.893614}$$

Result of the fit for Chile:

$$\text{Num. of documents} = \exp(8.959765)(\text{Out-links})^{-1.595429}$$

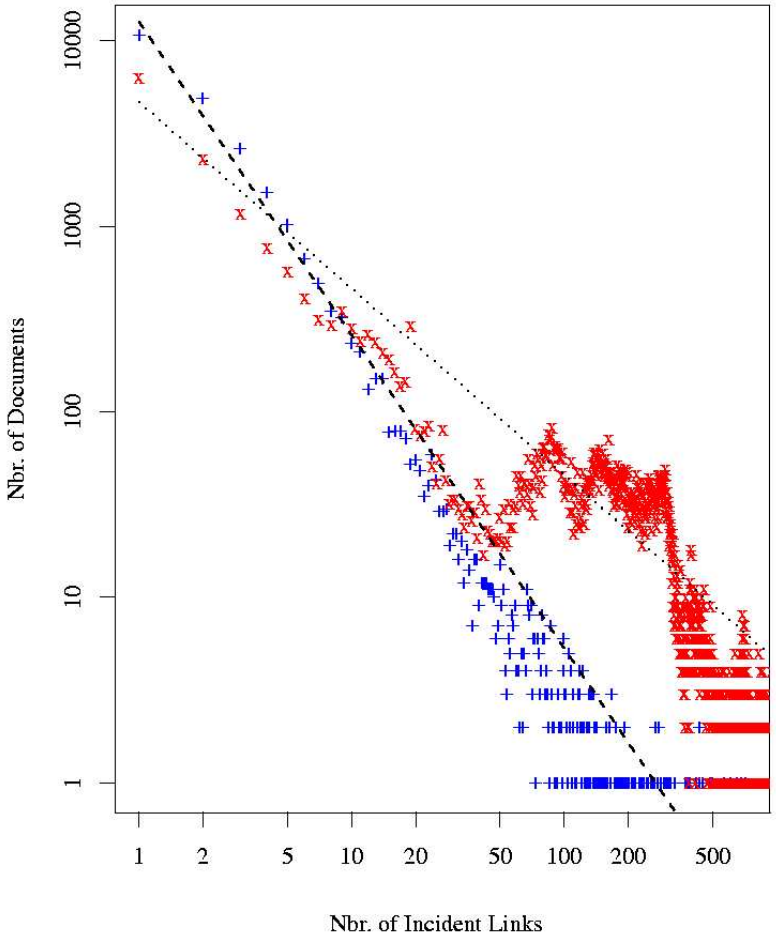


Figure 25: Power law fi tting of number of incident links to sites. The linear model is weighted by the number of observations.

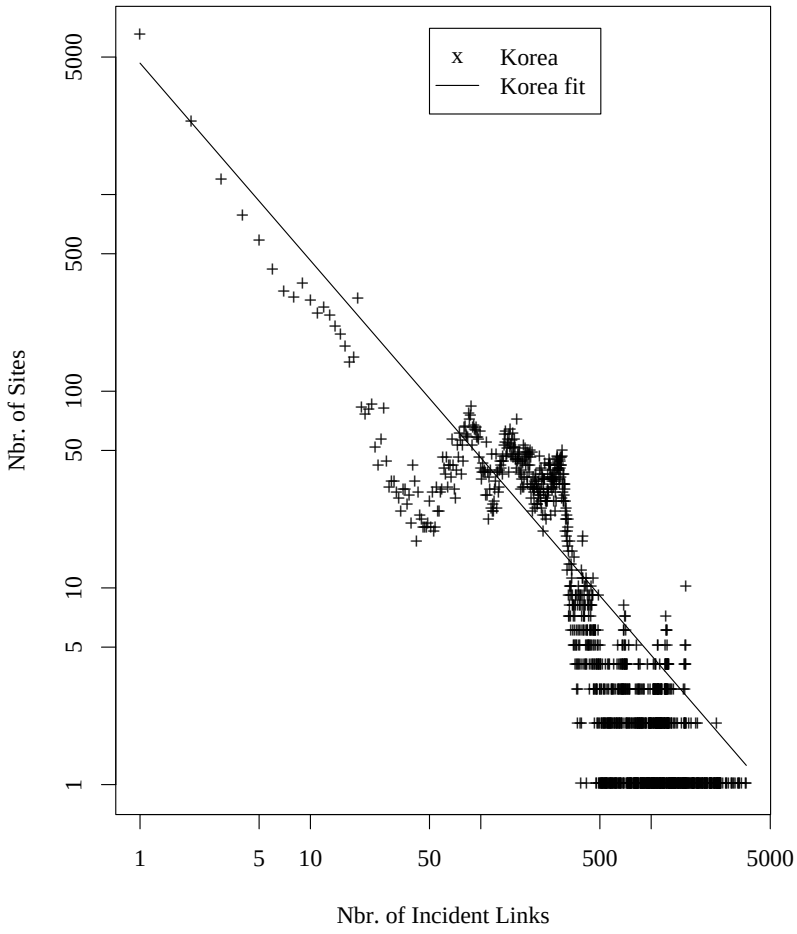


Figure 26: Detail of the power law fi tting of number of incident links to sites: Korea. The linear model is weighted by the number of observations.

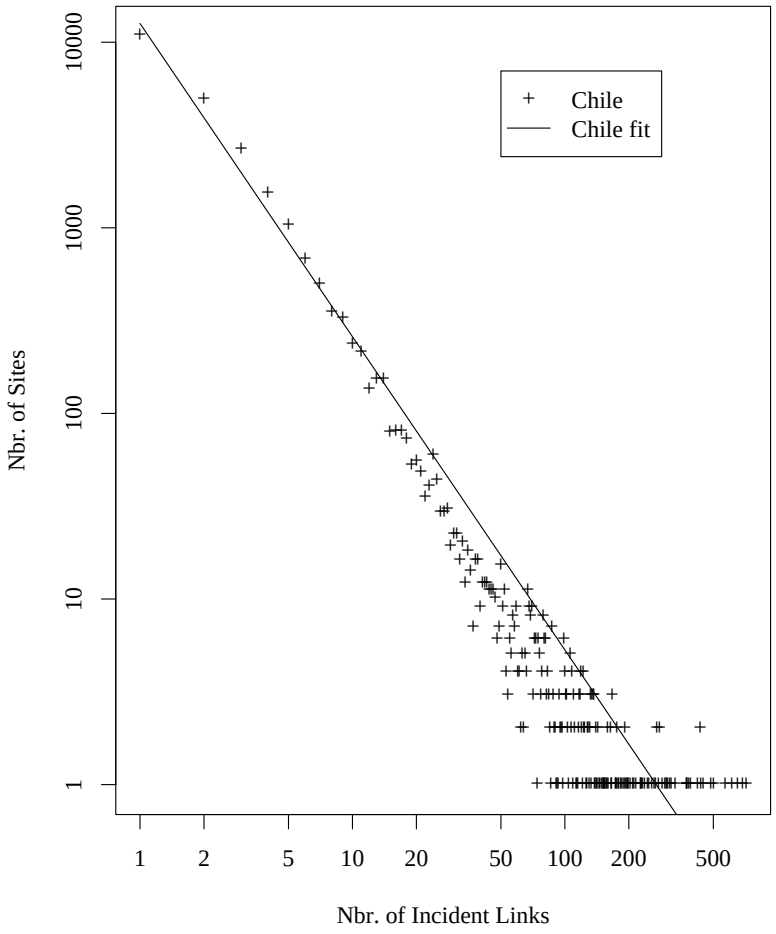


Figure 27: Detail of the power law fitting of number of incident links to sites: Chile. The linear model is weighted by the number of observations.

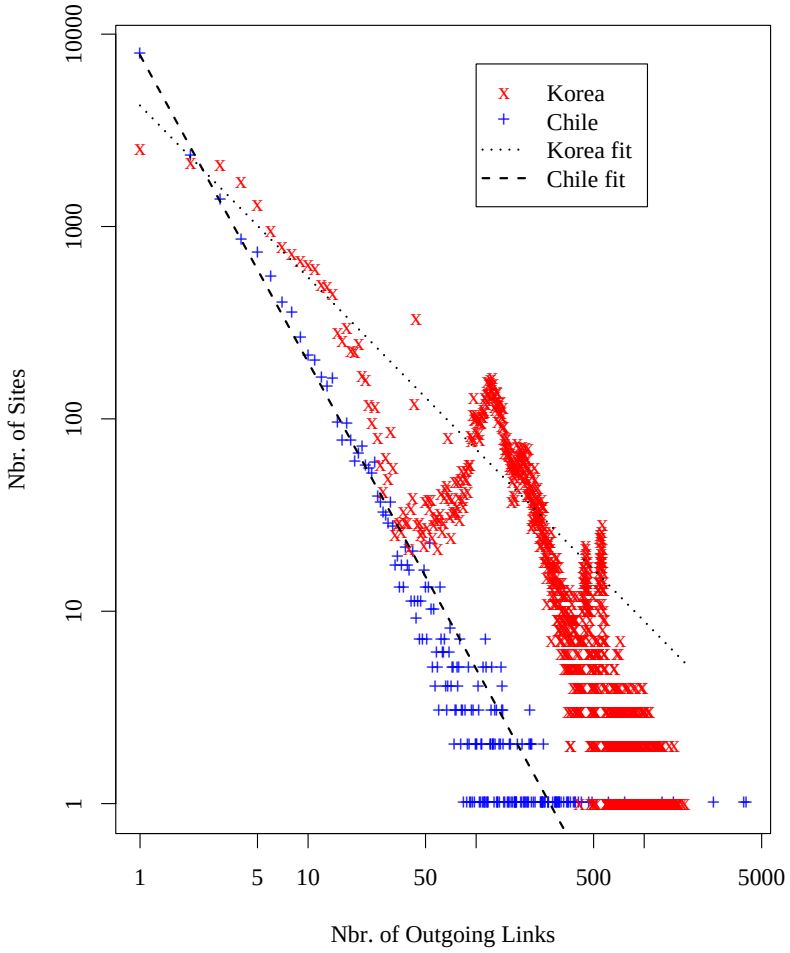


Figure 28: Power law fi tting of number of outgoing links to sites. The linear model is weighted by the number of observations.

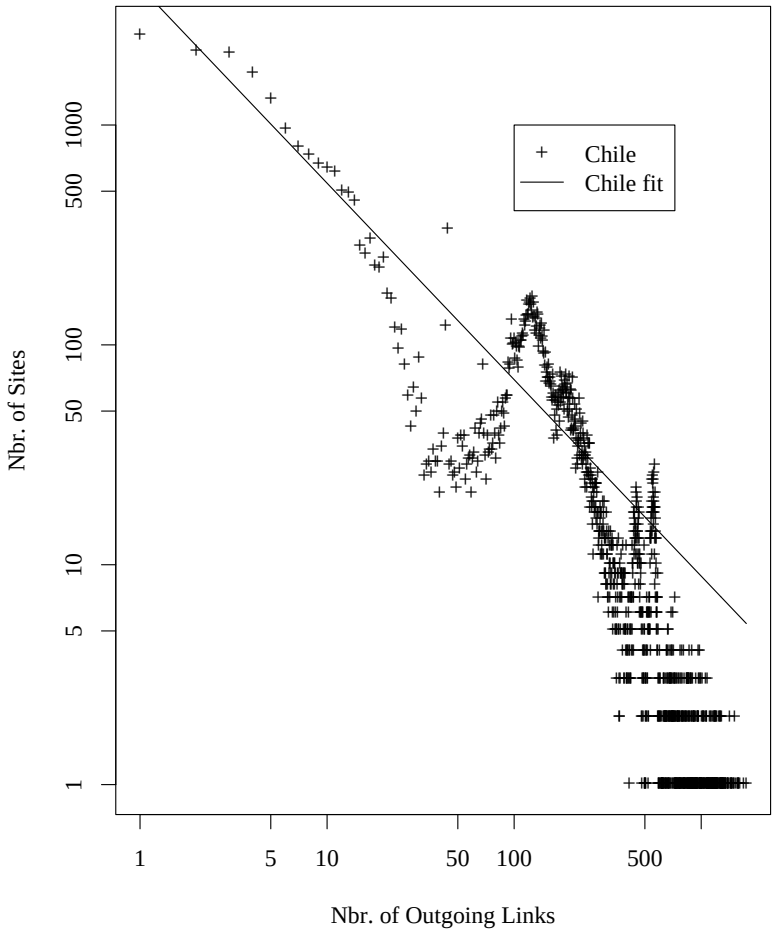


Figure 29: Detail of the power law fi tting of number of incident links to sites: Korea. The linear model is weighted by the number of observations.

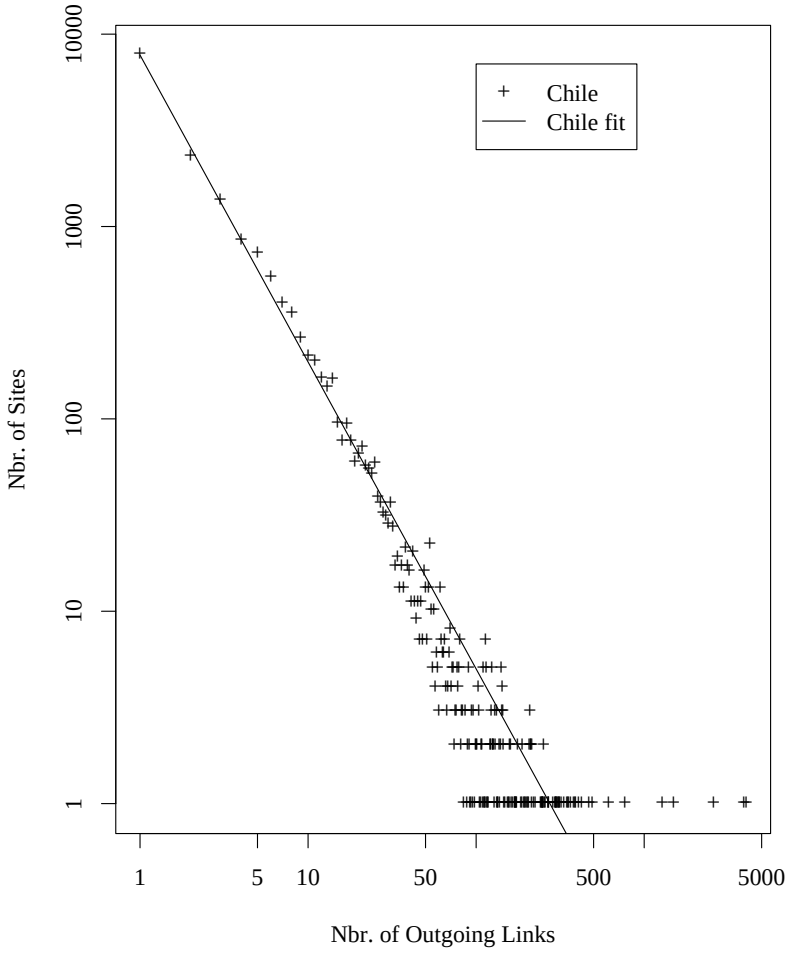


Figure 30: Detail of the power law fitting of number of incident links to sites: Chile. The linear model is weighted by the number of observations.

13 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.* [6]. This is a typical signature of a scale-free network.

Figure 31 shows the relative size of the strongly connected components. In both cases, over 99% of the strongly-connected components are singletons, composed of a single Web site. There is a single giant SCC in the Chilean Web, but there are several ones in the Korean Web. This means that either our sample is not still large enough for this specific type of analysis, or that the Korean Web is does not have this property when looked in isolation, and other Korean or foreign Web sites (e.g. in the `.NET` or `.COM` domains) are required to generate a single SCC.

Further inspection of the largest SCCs in the Korean Web showed that the first two come from unrelated domains, but the third one is of porn sites related to each other, which may explain the isolation of the later.

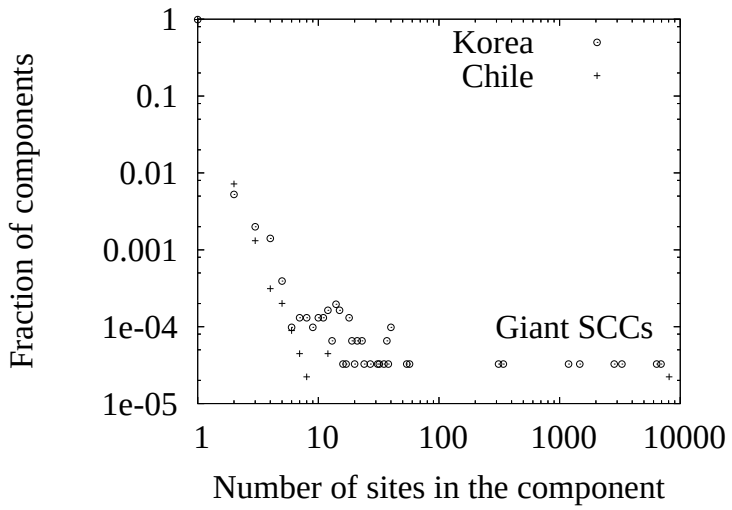


Figure 31: Distribution of sizes of the strongly connected components.

Strongly connected components can be used to characterize macroscopic characteristics of the Web graph. In [1] we extended the notation introduced by Broder *et al.* [6] for analyzing Web structure, by dividing the MAIN component into four parts:

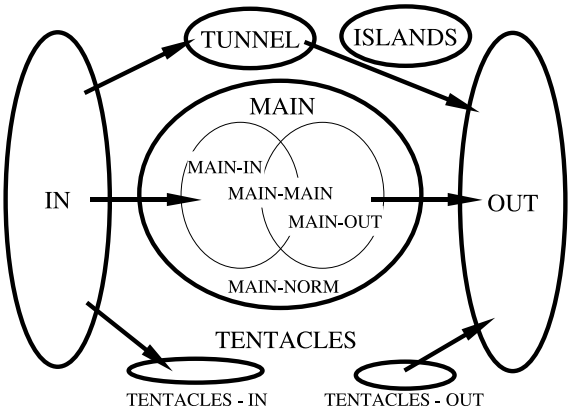
(e) MAIN-MAIN, which are sites that can be reached directly from the IN component and can reach directly the OUT component;

(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 only when having a complete list of the registered domains under the desired domain name, and this was possible only for the Chilean Web. The distribution of Web sites into components is shown in Figure 32. This structure evolves over time, as studied in [3, 4].



Component name	Korea	Chile
MAIN_NORM	6.10%	9.14%
MAIN_MAIN	2.45%	6.79%
MAIN_IN	2.14%	2.87%
MAIN_OUT	12.11%	9.37%
IN	25.93%	12.41%
OUT	7.54%	48.72%
TENTACLES-IN	43.69%	3.88%
TENTACLES-OUT	0.04%	6.82%

Figure 32: Relative size of components in the macroscopic structure of the Web.

.

14 Conclusions and Future Work

The closer similarities between the Korean and the Chilean Web are related to the degree distributions of links, showing self-similarities in the Web graph. On the other hand, different technologies and file formats are used in both Webs, and the external links relating to other countries are also very different.

The Korean Web is, in a sense, more mature than the Chilean Web in terms of having a larger number of actors and several businesses running on the Web. Pages are also larger, sites have more pages and they are more connected. Unfortunately, this means that there is also a larger presence of spam and porn Web sites than in the Chilean Web. For an analysis to be fair, some special cases such as a Web domain generating too many Web sites should be detected and removed from the study. This is a matter of further research for the next study.

However, it is interesting to see how two different countries that differ by a factor of 3 in population and by a factor of 6 in gross domestic product present so many similarities. This is further evidence for the existence of deeper rules that govern Web linking behavior.

References

[1] BAEZA-YATES, R., AND CASTILLO, C. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval* (Laguna San Rafael, Chile, November 2001), IEEE CS Press, pp. 21–32.

[2] BAEZA-YATES, R., AND CASTILLO, C. Balancing volume, quality and freshness in web crawling. In *Soft Computing Systems - Design, Management and Applications* (Santiago, Chile, 2002), IOS Press Amsterdam, pp. 565–572.

[3] BAEZA-YATES, R., AND POBLETE, B. Evolution of the Chilean Web structure composition. In *Proceedings of Latin American Web Conference* (Santiago, Chile, 2003), IEEE CS Press, pp. 11–13.

[4] BAEZA-YATES, R., AND POBLETE, B. Dynamics of the Chilean Web structure. In *Proceedings of the 3rd International Workshop on Web Dynamics* (New York, USA, May 2004), pp. 96 – 105.

[5] BARABÁSI, A.-L. *Linked: the new science of networks*. Perseus Publishing, 2002.

[6] BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. Graph structure in the web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web* (Amsterdam, Netherlands, May 2000), pp. 309–320.

[7] KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*, 5 (1999), 604–632.

[8] KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., SIVAKUMAR, D., TOMKINS, A., AND UPFAL, E. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)* (2000), IEEE CS Press, pp. 57–65.

[9] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The Pagerank citation algorithm: bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project, 1998.

[10] SOCIETY, I. Internet domain survey. http://www.isc.org/ds/, July 2004.

[11] THE ECONOMIST. Country Profiles, 2002.

[12] UNITED NATIONS. Human Development Reports, 2003.

[13] UNITED NATIONS. Population Division, 2004.