Dynamics of the Chilean Web structure

Ricardo Baeza-Yates *, Barbara Poblete

Center for Web Research, Department of Computer Science, University of Chile, Blanco Encalada 2120, Santiago, Chile

Abstract

In this paper we present a large scale study on the evolution of the Web structure of the Chilean domain (.cl) from 2000 to 2004, focusing on the Web site transitions in the structure. This is the study of the largest time span and the most detailed of its kind. Our results show that there are many stable Web sites, but also a majority of chaotic changes. We also present the first known results on the death behavior of Web sites.

Keywords: Web structure dynamics; Web growth; Website lifecycle

1. Introduction

The Web is highly dynamic and not too much is known about its evolution. There has been some work on page evolution, obtaining models that predict when a page will change, but this differs a lot from site to site. There are also generative models for Web growth, but they usually do not include Web death (an exception is [5]).

In this study we focus on the Web site graph or host-graph. Web sites are better study subjects than Web pages for many reasons. First, a Web site most of the time is a logical information unit, this being less true for pages. Second, the main events on the evolution of the Web are related to sites. In fact, new Web sites appear and others disappear, but little is known about how this happens. Third, most external links in a site are to home pages, so the Web structure of sites is the glue of the Web connectivity. Fourth, most sites are strongly connected (it is enough to have a link to the home page in every page). Otherwise, a Web site would have pages in more than one component of the structure, which does not make any sense as a Web site should be atomic with respect to the overall structure (see similar and additional arguments in [6]).

The only paper that focuses in the dynamics of the host-graph is [6], but it does not study the structure of the host-graph. In [3] we presented the evolution of the structure composition of the Chilean Web at the site and domain level, based on data gathered from a search engine targeted to this country's Internet domain, TodoCL.cl, between the years 2000 and 2002. We extended our results and their analysis to 2003 in [4]. In this paper we include data of 2004, extending our previous results and visualizations. We focus not only on macro statistics, but also on the transitions of Web sites among different structural components. That is, we try to

* Corresponding author. Tel.: +56 2 689 5531; fax: +56 2 689 2736.

E-mail addresses: rbaeza@dcc.uchile.cl (R. Baeza-Yates), bpoblete@dcc.uchile.cl (B. Poblete).

answer the following question: are the size changes in the Web structural components due to a small number of sites going from one component to another in one direction or to a larger number of sites that go in both directions? Our results show that for some Web components the first is true, while for others the second is true.

We define the Chilean Web as all .cl sites, which in practice represent more than 98% of the sites (other non .cl sites hosted in Chile are estimated to number less than 1000). The first year the crawl started from an initial sample of sites, but subsequent years it started with all .cl domains thanks to NIC Chile (www.nic.cl). Hence, the number of unconnected sites was low the first year. Also, the last three crawls contain more dynamic pages, which in general do not change the Web structure. In addition, the last two crawls, although larger in pages compared to 2002, may not reflect an actual growth in the Chilean Web as the number of sites did not increase that much. Table 1 shows the data gathered for our study. Although our results depend on our crawling policies, we have used always the same crawler, changing only the seed URLs. Obviously, each year our seed set is larger.

Our results present how the structure evolves, how sites migrate from one component to another component, and where sites appear and disappear in the structure. The changes are dramatic, showing more chaos than order, and we elaborate on this in the conclusions. This is a first step to measure and follow the evolution of the structure of a part of the Web, as well as try to understand the process behind the changes. To the best of our knowledge there are no other studies on Web structure composition as detailed as ours, both in results and time span. Most statistical studies deal with global attributes such as language or size.

In Section 2 we review the results on the structure of the Web and the problems faced to obtain it. Section 3 shows the evolution of this structure, and Section 4 analyzes the migrations of Web sites in the structure in relation to the expected typical life cycle of a Web site. In Section 5 we analyze the dynamics of the size of Web sites. The last section contains our concluding remarks.

2. Web structure

The most complete (and unique) study of the Web structure [7] focuses on page connectivity. One problem with this is that a page is not a logical unit (for example, a page can describe several documents and one document can be stored in several pages). Hence, we started by studying the structure of how Web sites were connected, as Web sites are closer to being real logical units. Not surprisingly, we found in [1] that the structure at the Website level was similar to that of the global Web, and hence we were able to use the same notation of [7]. The components are

(a) MAIN, sites that are in the strong connected component of the connectivity graph of sites (that is, we can navigate from any site to any other site in the same component);
(b) IN, sites that can reach MAIN but cannot be reached from MAIN;
(c) OUT, sites that can be reached from MAIN, but there is no path to go back to MAIN; and
(d) other sites that can be reached from IN (T.IN, where T is an abbreviation for tentacles), sites in paths between IN and OUT (TUNNEL), sites that only reach OUT (T.OUT), and unconnected sites (ISLANDS).

In [1] we analyzed the data for year 2000 and we extended this notation by dividing the MAIN component into four parts:

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

Table 1
TodoCL collections

	Year				
	2000	2001	2002	2003	2004
Pages	695,546	794,046	1,987,804	3,135,020	3,252,779
Sites (crawled)	7468	21,204	38,307	38,208	53,527
Sites (known)	7468	22,882	45,606	56,018	78,477
Domains (crawled)	6261	19,386	34,869	33,912	47,468
Domains (known)	6261	20,644	41,184	49,258	69,073

(b) MAIN-IN, which are sites that can be reached directly from the IN component but are not in MAIN-MAIN;
(c) MAIN-OUT, which are sites that can reach directly the OUT component, but are not in MAIN-MAIN;
(d) MAIN-NORM, which are sites not belonging to the previously defined subcomponents.

Fig. 1 shows all these components. The average update time of pages and sites, and their relation to structure and link ranking techniques was studied in [2] for the first two collections (2000 and 2001). We could consider domains in our study, but domains may contain sites that are quite different. For example, Web hosting in an ISP provider using a common second-level domain such as co.cl.

Given this structure, with good seeds, it is possible to crawl MAIN and OUT without problems. The rest is more difficult if we do not have a complete list of seeds, and most studies do not find, for example, all of the ISLANDS. In our case, we have most of the Chilean domains, hence our study has a very large coverage. On the other hand, because any crawling is incomplete (for example, dynamic pages can be unbounded), any Web graph will be incomplete. That means that any analysis of the Web structure will be an approximation. Moreover in our case, as we are not considering paths through links outside the Chilean Web, we cannot know a path between two pages if the path goes outside the .cl domain. Nevertheless, our Web subset is a very coherent one and it is not just a Web sample. To know if a site exists, it is enough to crawl the home page. However, to know all the links for that site, a thorough crawling of the site is needed.

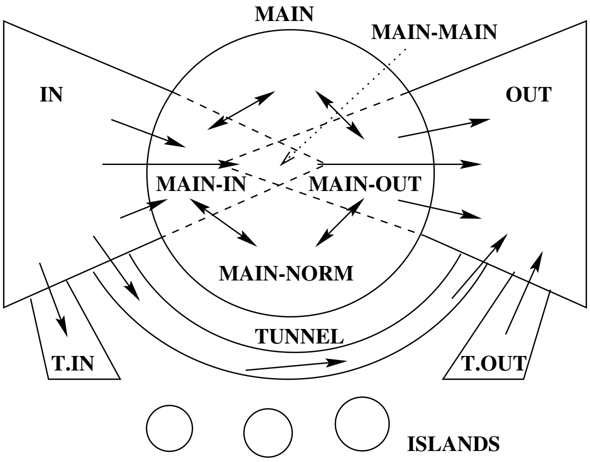

Fig. 1. Structure of the Web.

3. Evolution of the structure composition

Table 2 shows the number of sites that have appeared and disappeared from year to year, from a total of 78,477 different sites belonging to 69,073 domains, crawled at some point. As of April 6, 2005, there were 119,408 registered domains in .cl, with 94,348 having a DNS server. Hence, in the worst case our data covers 73% of all domains in .cl. However, we estimate that the coverage is over 80%. The last three rows represent the new sites (NEW), the sites that were not crawled but exist (UNKNOWN), and the sites that disappeared (DEAD), respectively. In both cases, we count on a year to year basis. That is, it is NEW from a year to the next, not to the overall period considered. UNKNOWN include non-crawled existing sites and sites with connectivity or access problems. NEW sites may not be really new, as the crawling coverage is not 100%. Death of a site means that there is no IP address associated with it (this might be incorrect if the site changes its name, but then it is considered as a new site and there are few such cases) and death of a domain means that there are no sites associated with it (in particular the domain name itself or prefixed by www).[1]

In Table 3 we give the relative size of each component. Notice the size of ISLANDS in 2004, which is over 45% of the Chilean Web sites (but only a small percentage of the total number of pages). These sites are usually recent, and the main growth of the Web is in that component. As our collection is not complete, the percentages for MAIN are lower bounds while for ISLANDS, upper bounds. As we checked for non-crawled sites to see if they exist, but we do not know the actual component they belong to, we can have upper and lower bounds for MAIN and ISLANDS, by adding and subtracting the number of sites with an unknown component, respectively. For example, the real number of sites in MAIN is between MAIN-UNKNOWN and MAIN+UNKNOWN.

To visualize the evolution, Fig. 2 shows the growth of each component including the number of sites dying (left) and the percentage for each component, including UNKNOWN sites (the dead sites are represented in a normalized fashion using the num-

[1] The domain name could be still registered and have a name server, though.

Table 2
Growth and death of sites (2000–2004)

	Year				
	2000	2001	2002	2003	2004
CRAWLED	7468	21,204	38,307	38,208	53,527
NEW	–	15,414	22,724	10,412	22,459
UNKNOWN	–	856	1766	3599	6195
DEAD	–	822	4343	8143	5474

Table 3
Relative size of the components of the Chilean Web (2000–2004)

	Component size (%)				
	2000	2001	2002	2003	2004
TIN	1.31	3.04	3.09	1.96	2.08
IN	10.81	5.84	10.07	8.22	6.65
MAIN	36.35	9.24	11.71	18.36	15.11
OUT	39.39	20.21	16.57	26.58	26.12
TOUT	4.03	1.68	3.1	3.74	3.65
TUNNEL	0.37	0.22	0.21	0.21	0.23
ISLANDS	7.71	59.73	55.21	40.9	46.16
MAIN-MAIN	3.88	3.43	4.10	4.65	3.64
MAIN-OUT	8.86	2.49	2.79	6.28	5.03
MAIN-IN	4.76	1.16	2.23	2.20	1.54
MAIN-NORM	18.95	2.15	2.90	5.24	4.90

ber of existing sites as the 100% level). The gray levels follow the order given by the boxed legend at the right.

4. Analysis of Website migration

In this section, we analyze how sites migrate in the structure. If a year a site S is in component A and the next year it is found in component B ($B \neq A$), we say that S migrated from A to B (a state transition in the structure). In Table 4 we show the sorted percentage of aggregated transitions for all the years.

In Appendix A we give the absolute numbers for the migration of sites per year among all the components. In most cases the UNKNOWN component sites will belong to ISLANDS or OUT, although in the later case, we just need one link back to MAIN to have that site in MAIN. Notice that OUT and

Table 4
Total sorted percentage of migrations between components of the Chilean Web (2000–2004)

Transition	Percent
NEW-ISLANDS	55.30
ISLANDS-DEAD	15.05
NEW-OUT	14.47
NEW-MAIN	8.53
NEW-IN	7.93
ISLANDS-OUT	7.11
MAIN-OUT	4.29
OUT-MAIN	3.95
OUT-ISLANDS	3.91
OUT-DEAD	3.16
ISLANDS-IN	2.37
IN-DEAD	2.18
IN-ISLANDS	2.17
IN-MAIN	1.72
MAIN-DEAD	1.53
ISLANDS-MAIN	1.48
IN-OUT	0.94
MAIN-IN	0.88
MAIN-ISLANDS	0.85
OUT-IN	0.57

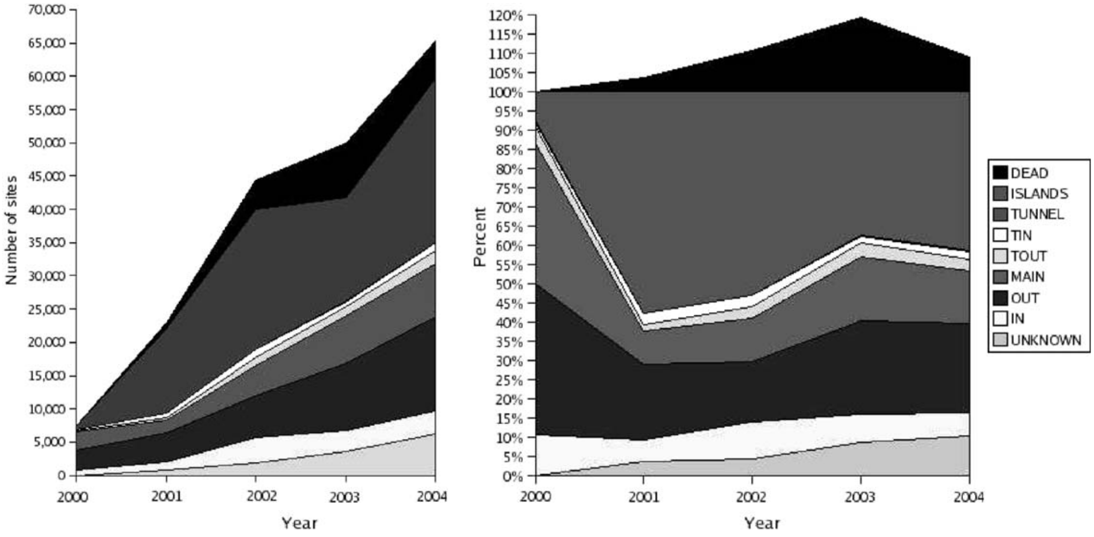

Fig. 2. Growth of the structural components, as well as site death: absolute value (left) and percentage (right).

MAIN are quite stable components, because a large fraction of their sites stay there. It is also interesting to see that MAIN grows mainly from OUT or NEW sites, and that ISLANDS is the component with largest growth and also death, followed by OUT (and not IN as would be expected).

Web sites evolve and hence migrate inside the structure. First, a typical Web site should start as part of ISLANDS or IN (depending if they link or not to a good Web site). If the site becomes popular and they also link to known sites, the site migrates to MAIN. If links are not well chosen or updated, they start in or migrate to OUT. Fig. 3 shows the expected life path of a Website to migrate to MAIN. We also include migrations from MAIN to OUT if the site is not well maintained. On the other hand, the left side of Fig. 4, shows what really happened, aggregating all the transitions in our data (dark arrows are sites that disappear). The main differences from our intuition are that there are very few IN to MAIN and IN to ISLANDS transitions. However, some of the transitions involve changes in two links, for example, from IN to OUT or MAIN to or from ISLANDS. Assuming that the two links do not appear exactly at the same time, the transition from IN to OUT went through MAIN or ISLANDS, ISLANDS to MAIN went through IN or OUT, and MAIN to ISLANDS went through OUT or IN. This means that a finer time granularity on the Web snapshots is needed to understand 3.4% of the transitions.

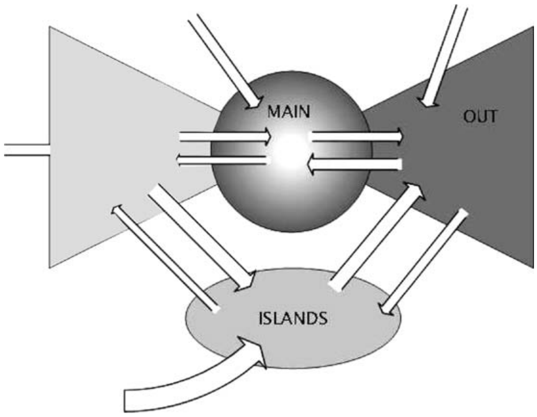

Fig. 3. Expected migrations of Web sites in the structure.

Using the transitions of Fig. 4 as a static Markov chain, assuming that the rest of the cases in each part of the structure are internal transitions to itself (except the NEW+DEAD case), we obtain a 31% upper bound on the size of MAIN or OUT, and a 19% upper bound in the size of IN. Similarly, we get a 19% lower bound for the size of the ISLANDS.

Fig. 5 shows the real migration of each site in the structure using one grey level per component. The order of the grey levels, from white to black is (NEW+UNKNOWN+DEAD, TIN, IN, MAIN, OUT, TOUT, TUNNEL, ISLANDS). Each column

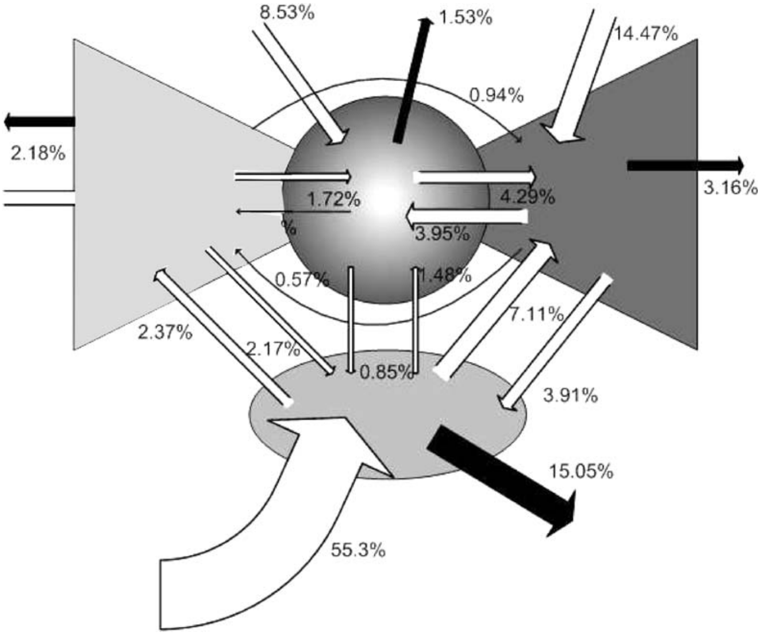

Fig. 4. Aggregated real migrations of Web sites in the structure.

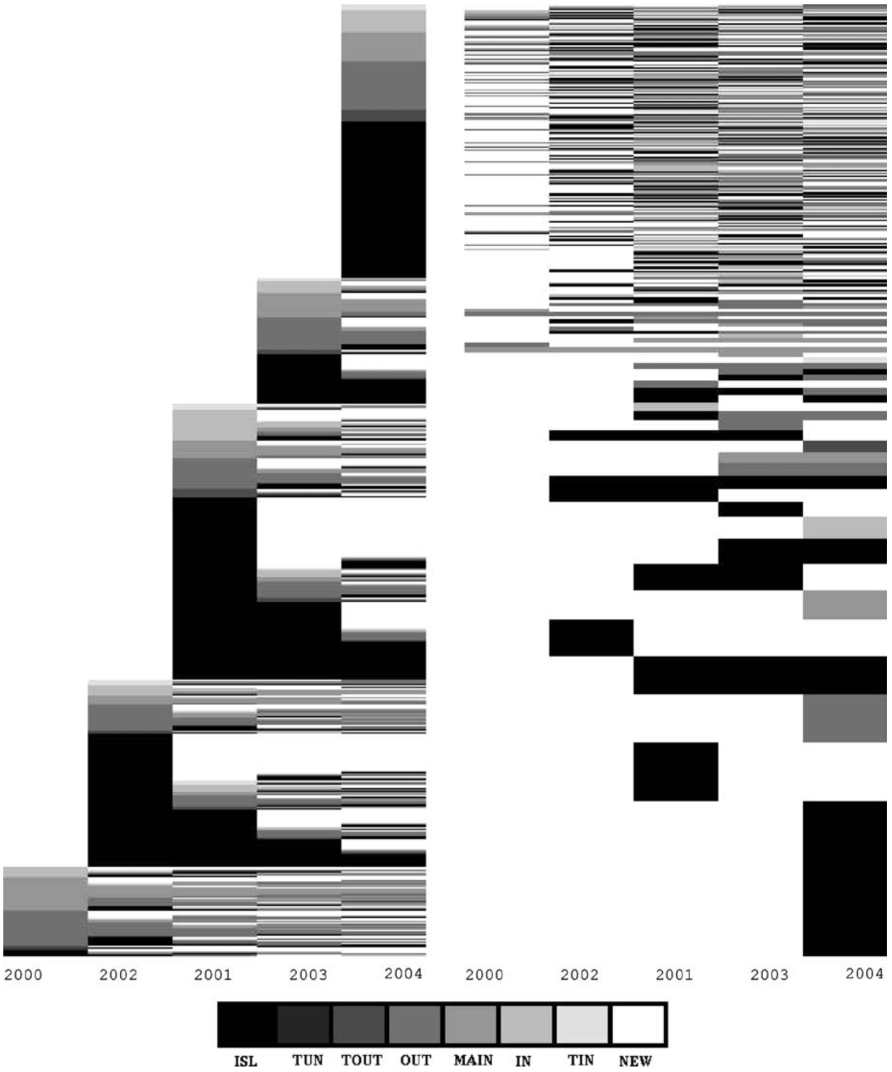

Fig. 5. Migrations of Web sites in the structure (one column per year, one line per site, one grey level per component). The left side is sorted by grey level order, right side by case frequency.

is a year from 2000 to 2004 and each Web site is a horizontal line with segments having gray levels depending on the component that the site belonged to each year. The left visualization has the horizontal lines sorted by gray level and the right visualization is sorted by case frequency.

From the possible 16,807 migration patterns, we found only 2954 (17.6%) in the 78,477 sites. Still, this is quite large and shows the dynamism of the Web. We can clearly see the growth in the white space at the left, the transition NEW to ISLANDS being the most frequent. The white space to the right are the UNKNOWN or DEAD cases.

Fig. 6 shows the same, but keeping only the Websites that were always found (that is, they were never in the NEW, UNKNOWN, or DEAD state). This subset is interesting because is independent of our crawling seeds and policies, and also because represents the core of the Chilean Web. This subset is a zoom on the bottom part of the figure removing all sites having at least one white line and comprising 3395 sites (4.3%). Here we found 704 (9.1%) of the 7776 possible migration patterns, which is consistent with the fact that they should have more component stability. Here we can see that the most frequent cases are to remain in MAIN or OUT or to switch between those components. These cases account for 50.1% of all cases, not including the fifth most frequent case, which are sites that are in OUT but one year were ISLANDS. That is, 50% of the core of

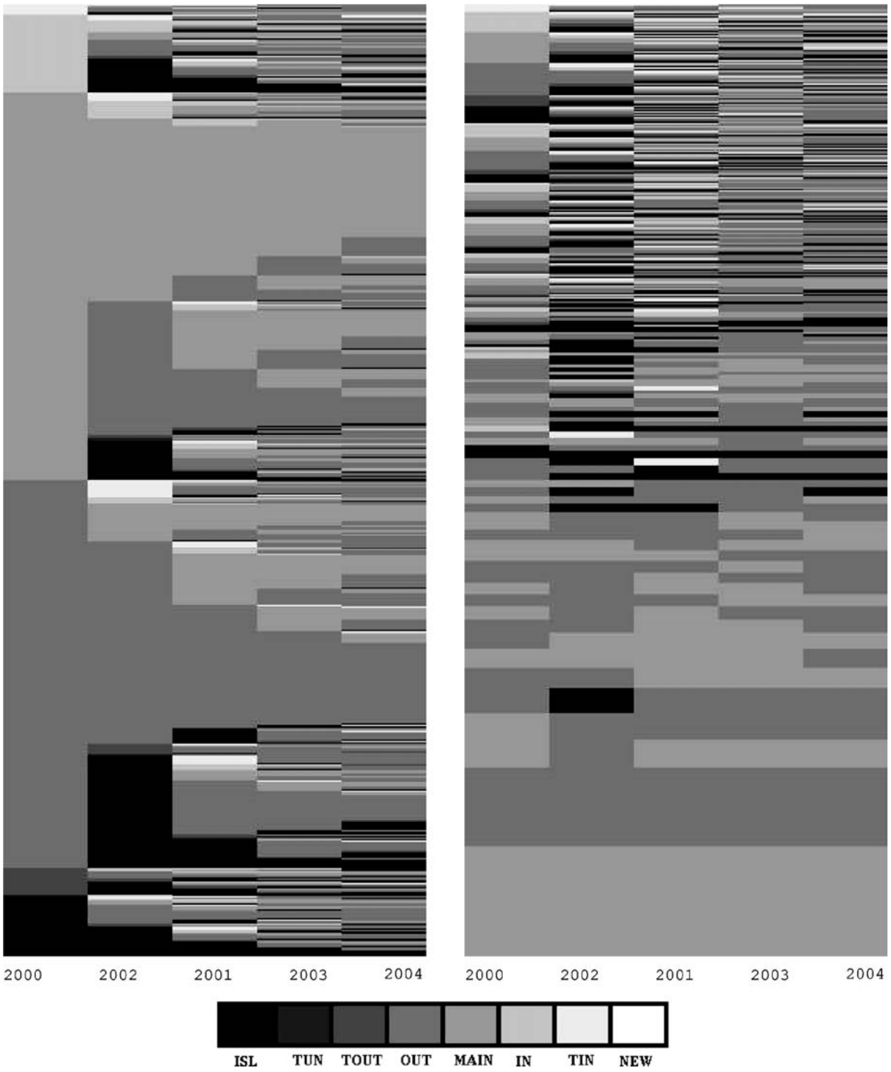

Fig. 6. Migrations of Web sites in the structure considering only stable Web sites (one column per year, one line per site, one grey level per component). The left side is sorted by grey level order, right side by case frequency.

the Web is quite stable, only 2.2% overall. We can notice also that there is almost no migration from IN to MAIN in opposition to what our intuition predicted. Also, there are Web sites that appear directly in MAIN or OUT. This means that a good site seems to be linked from a site in MAIN in less than a year, or that sites obtain links from portals in MAIN (for example, a banner).

5. Web size dynamics

Another issue is the dynamics of the sites' contents, which is far more difficult and complex. One first estimation is to look at the changes in the number of pages. For example, the largest 100 sites (in pages) per year, involve 408 sites for all years (so there are many changes in page size), and only 10 and 72 sites were in the top for 3 and 2 years, respectively. Fig. 7 shows the number of pages of the 10 largest Web sites per year from 2000 to 2003 (in total 39 different Web sites). Although the number of pages depends on crawling policies, we have used more or less the same policies all the time and the changes are quite radical.

One reason for sudden changes could be attributed to the business behind Web evolution. However, there are additional and very different reasons for page count changes. The main one is Web design changes. For example, from static pages to dynamic generation of pages. Even worse, design changes that

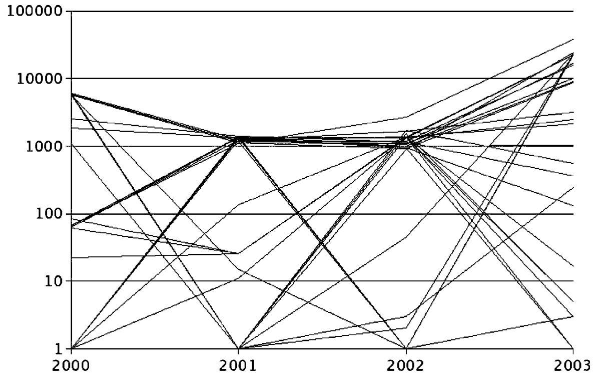

Fig. 7. Changes in the number of pages for the 10 top sites per year (2000–2003).

do not allow crawlers to enter, mostly because of ignorance. For example, in 2001, 56% of the domains and 54% of the sites had only one page. However, in 25% of them (14% of the total) was because they had an initial "binary" page that hides the internal links (Flash pages for example). In 2004, only 40% of the sites had one page, but 31% of them were due to binary pages (13% of the total). Although the percentage is the same, the absolute value of "invisible" sites has more than doubled in three years.

6. Concluding remarks

The Web is very young and in Chile the first Web site appeared at the end of 1993 in our CS department. As we have data for five years, our study covers more than 40% of the main part of the lifetime of the Chilean Web.

The overall number of sites of the Chilean Web almost doubles each year, as we believe that the last year did not reflect the actual growth, mainly due to the prevalence of dynamic pages. This growth is the result of about a 100% increase plus a 20% death. So, one might use a simple model for Web site growth of $f_n = (\alpha - \beta)f_{n-1}$ where α is the growth rate and β the death rate. According to our results we have $\alpha \approx 1.98$ and $\beta \approx 0.17$, obtaining $f_n \approx 1.81f_{n-1}$. While an exponential growth cannot be sustained too long, the Web has been growing exponentially for more than 10 years. On the other hand, the Web grows continuously, and we only have one snapshot per year. Different time granularities for this type of data could be considered to see if a one-year sampling is good enough.

There is still work to do to understand how the composition of the structure changes, but perhaps there are no formal processes driving the situation. Indeed, our results imply that perhaps we are trying to study a process that is still in a transient phase, or that cannot be modeled at such a level of detail.

We plan to extend our study by separating the Chilean Web sites in commercial, educational, governmental, military, etc. categories. Although Chile does not use a subdomain level indicating this, we have the classification made at registration time. Perhaps there will be stability differences among these different classes.

Acknowledgments

We thank the help of Edgardo Krell and Sebastian Castro from NIC Chile for providing the .CL domain data, as well as the support of Millennium Nucleus Grant P04-067-F from Mideplan, Chile.

Appendix A

Tables 5–8 present all the transitions among components from 2000 to 2004. There are two ways

Table 5
Component changes of sites from 2000 to 2001

2000	2001								
	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	959	724	139	304	11	61	24	275	218
OUT	195	1151	39	749	5	96	48	336	323
IN	39	89	118	279	2	31	25	103	122
ISLANDS	18	124	14	213	0	14	19	77	97
TUNNEL	1	1	3	18	0	0	2	2	1
TIN	5	31	0	18	3	3	2	19	17
TOUT	3	38	25	131	0	4	12	44	44
UNKNOWN	0	0	0	0	0	0	0	0	0
DEAD	0	0	0	0	0	0	0	0	0
NEW	741	2128	901	10,955	27	437	225	0	0

Table 6
Component changes of sites from 2001 to 2002

2001	2002								
	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	1209	315	105	39	1	8	4	132	148
OUT	896	1679	181	528	15	128	43	358	458
IN	231	96	281	188	1	22	16	127	277
ISLANDS	417	1346	714	5129	23	360	299	1052	3327
TUNNEL	11	15	3	4	1	2	0	8	4
TIN	78	214	24	127	2	65	5	57	74
TOUT	51	79	41	57	0	18	24	32	55
UNKNOWN	92	171	36	158	1	22	8	0	0
DEAD	0	0	0	0	0	0	0	0	822
NEW	1504	2434	2474	14,923	38	562	789	0	0

Table 7
Component changes of sites from 2002 to 2003

2002	2003								
	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	2494	851	147	123	7	20	39	431	377
OUT	1006	2918	98	689	9	81	69	701	778
IN	674	322	910	481	6	15	196	449	806
ISLANDS	497	2314	796	9239	20	241	501	1780	5765
TUNNEL	20	31	1	7	0	0	3	11	9
TIN	102	512	28	182	10	49	15	141	148
TOUT	64	149	97	291	4	11	226	86	260
UNKNOWN	187	362	86	528	2	27	39	0	0
DEAD	0	0	0	0	0	0	0	0	5165
NEW	1972	2698	979	4090	24	308	341	0	0

Table 8
Component changes of sites from 2003 to 2004

2003	2004								
	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	3671	1483	300	207	15	44	40	796	460
OUT	1010	5473	133	1108	26	167	132	1180	928
IN	412	231	593	755	11	47	99	488	506
ISLANDS	231	1799	337	7431	14	240	435	2518	2625
TUNNEL	6	21	0	15	4	3	8	15	10
TIN	39	226	17	180	2	77	11	103	97
TOUT	49	186	90	459	11	11	192	176	255
UNKNOWN	184	462	216	1116	1	53	78	0	593
DEAD	66	161	57	566	0	15	42	919	11,482
NEW	2417	3940	1817	12,869	39	457	920	0	0

of reading these tables. In each column, we have the percentage of sites in a component that come from components of the previous years. In each row, we have how the sites of a component one year were disturbed in the components of the following year.

References

[1] R. Baeza-Yates, C. Castillo, Relating Web characteristics with link analysis, in: String Processing and Information Retrieval, IEEE Computer Science Press, Silver Spring, MD, 2001.

[2] R. Baeza-Yates, F. Saint-Jean, C. Castillo, Web dynamics, structure, and link ranking, in: String Processing and Information Retrieval, Lecture Notes in CS, Springer, Berlin, 2002.

[3] R. Baeza-Yates, B. Poblete, Evolution of the Chilean Web structure composition, in: First Latin American World Wide Web Conference, November, IEEE CS Press, Santiago, Chile, 2003.

[4] R. Baeza-Yates, B. Poblete, Dynamics of the Chilean Web structure, in: 3rd Workshop on Web Dynamics, New York, USA, May 2004.

[5] Z. Bar-Yossef, A. Broder, R. Kumar, A. Tomkins, Sic transit Gloria Telae: Towards an understanding of the Web's decay, in: 13th World Wide Web Conference, New York, USA, 2004.

[6] K. Bharat, B-W. Chang, M. Henzinger, M. Ruhl, Who links to whom: mining linkage between Web sites, in: IEEE International Conference on Data Mining, 2001.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, Graph structure in the Web: Experiments and models, in: 9th World Wide Web Conference, Amsterdam, Netherlands, 2000; Also published in Computer Networks.

Ricardo Baeza-Yates received his Ph.D. in CS from the University of Waterloo, Canada, in 1989. In 1992, he was elected president of the Chilean Computer Science Society (SCCC) until 1995, being elected again in 1997. In 1993, he received the Organization of American States award for young researchers in exact sciences. In 1997 with two Brazilian colleagues obtained the COMPAQ prize to best Brazilian research article in CS. He was international coordinator of CYTED (Iberoamerican cooperation in S&T) on applied electronics and informatics from 2000 to 2004. During 2002–2004, he was a member of the Board of Governors of the IEEE Computer Society. In 2003, he was incorporated to the Chilean Science Academy, being the first computer scientist to achieve that status. Currently he is professor and director of the Center for Web Research at the CS department of the University of Chile, where he was the chairperson in the periods 1993–1995 and 2003–2004. He is also ICREA Professor at the Department of Technology of the Pompeu Fabra University at Barcelona, Spain. His research interests include information retrieval, algorithms, and information visualization. He is co-author of the book Modern Information Retrieval (Addison-Wesley, 1999), as well as co-author of the second edition of the Handbook of Algorithms and Data Structures (Addison-Wesley, 1991); and co-editor of Information Retrieval: Algorithms and Data Structures, (Prentice-Hall, 1992), among other publications in journals published by ACM, IEEE or SIAM. He has been visiting professor or invited speaker at several conferences and universities all around the world, as well as referee of several journals, conferences, NSF, etc. He is member of the ACM, EATCS, IEEE (senior), SCCC (distinguished) and SIAM.

Barbara Poblete is currently a second year Ph.D. student at the University Pompeu Fabra (UPF) in Barcelona, Spain. She obtained a B.Sc. and M.Sc. in Computer Science and a Computing Engineering professional degree from the University of Chile in Santiago, Chile. She is a member of the Web Research Group at the UPF, and administrator of the Chilean vertical search engine TodoCL (http://www.todocl.cl). She obtained the second place in the XII Latin American Master's Thesis Contest in 2005. Her current research interests are Web mining, Information Retrieval and Web dynamics.