Dynamics of the Chilean Web Structure*

Ricardo Baeza-Yates
Barbara Poblete

Center for Web Research
Dept. of Computer Science
University of Chile
E-mail: {rbaeza,bpoblete}@dcc.uchile.cl

Abstract

In this paper we present further results on the evolution of the structure of the Chilean Web between 2000 and 2003, focusing on the Web sites transitions in the structure. Our results show that there are many stable Web sites, but also chaotic changes. We also expand our own results on the death behavior of Web sites.

1 Introduction

The Web is highly dynamic and not too much is known about its evolution. There have been some work on page evolution, obtaining models that predict when a page will change, but that differs a lot from site to site. There are also generative models for Web growth, but they do not include Web death. In fact, new websites appear and others disappear, but little is know on how this happens. Less work has been done when sites or domains are used as granularity for the study of the evolution. In [BYP03] 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 web domain, TodoCL.cl, between years 2000 and 2002. In this paper we include data of 2003, extending our previous results. In addition, we focus not only on macro statistics, but also on the transitions of Web sites among different structure components. That is, are the changes in the size of the components due to small transitions in one direction or to large transitions 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 the .cl sites plus all other sites found by crawling that have an IP belonging to a Chilean ISP. 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 two crawls contain more dynamic pages, which in general do not change the Web structure. In addition, the 2003 crawl, although larger in pages compared to 2002, may not reflect the actual growth of the Chilean Web as the number of sites did not increase. Table 1 shows the data gathered for these years. Although our results depend on our crawling policies, we have used always the same crawler, changing only the seeds. Obviously, each year our seed set is larger.

*This paper was written while the first author was visiting the University of Canterbury, Christchurch, New Zealand, thanks to an Erskine fellowship.

Year	2000	2001	2002	2003
Pages	613,415	794,218	2,214,253	3,135,089
Sites (crawled)	7,483	21,207	38,965	38,277
Sites (known)	7,483	22,898	46,277	56,699
Domains (crawled)	6,288	19,389	35,390	33,981
Domains (known)	6,288	20,660	41,717	49,790

Table 1: TodoCL collections.

Our results present how the structure evolves, how sites migrate from one component to another component, and where sites appear and disappear. The changes are dramatic, corroborating that perhaps we are trying to study a process that is still in a transient phase, or that cannot be modeled in detail. This is a first step to measure and follow the evolution of part of the Web structure, as well as try to understand the process behind the changes. To the best of our knowledge there are no other studies on Web composition as specific as ours. Most statistical studies deal with global attributes such as language or size. We would have liked to separate the Chilean Web in commercial, educational, governmental, etc. sites, but Chile does not use a subdomain level indicating this, so the classification is not trivial.

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. The last section has some concluding remarks.

2 Web Structure

The most complete study of the Web structure [BKM$^+$00] focuses on page connectivity. One problem with this is that a page is not a logical unit (for example, a page can describe several documents and one document can be stored in several pages.) Hence, we study the structure of how websites were connected, as websites are closer to being real logical units. Not surprisingly, we found in [BYC01] that the structure at the website level was similar to the global Web, and hence we use the same notation of [BKM$^+$00]. 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 [BYC01] we analyzed the data for 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;

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.

Figure 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 [BYSJC02] 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.

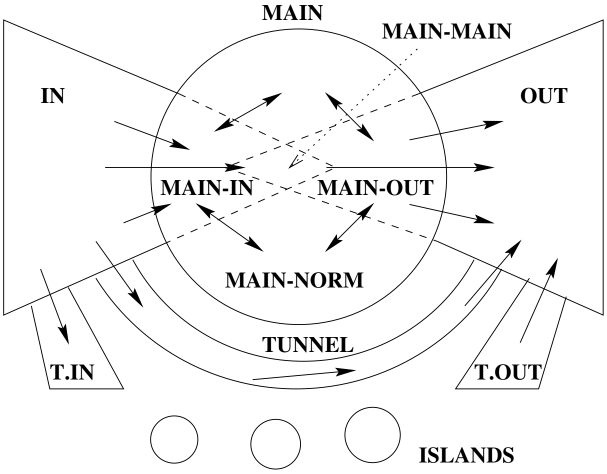

Figure 1: Structure of the Web.

Because any crawling is incomplete (for example, dynamic pages can be unbounded), any Web graph is the 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. On the other hand, our Web subset is a very coherent one and 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. However, many sites, sometimes because of ignorance, do not allow crawlers to enter. For example, in 2001, 56% of the domains and 54% of the sites had only one page. However, 25% of them (14% of the total) was because they had an initial Flash page or called a similar kind of program.

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 56,699 different sites belonging to 49,790 domains, crawled at some point, 56,020 of them being .cl. As of March 1st, 2004, there were 75,758 registered domains with a name server in .cl, with approximately 56,100 with a Web server, our coverage was around 67% in 2003 and higher in 2002. The three last rows represent the new sites (NEW), the sites that were not crawled but exist (UNKNOWN), and the sites that disappeared (DEAD), respectively. 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 to it (this might be incorrect if the site changes its name, but then it is considered as a new site and there are few of 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].

	Sites			
Year	2000	2001	2002	2003
CRAWLED	7,483	21,207	38,965	38,277
NEW	—	15,445	23,379	10,422
UNKNOWN	—	869	1,768	4,142
DEAD	—	822	5,165	8,210

Table 2: Growth and death of sites (2000-2003).

In table 3 we give the relative size of each component. Notice the size of ISLANDS, which is near 40% of the Chilean Web sites. These sites are usually recent, and the main growth of the Web is in that component. We can also observe the growth of MAIN, which may indicate a more mature Web. As our collection is not complete, the percentages for MAIN are lower bounds while for ISLANDS they are 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.

Component Size (%)	2000	2001	2002	2003
MAIN	36.45	9.25	12.02	18.37
IN	10.79	5.84	10.03	8.21
OUT	39.36	20.21	16.90	26.24
TUNNEL	0.37	0.22	0.22	0.21
TENTACLE-IN	1.32	3.04	3.12	1.97
TENTACLE-OUT	4.01	1.68	3.15	3.74
ISLANDS	7.68	59.73	54.54	40.86
MAIN-MAIN	3.88	3.43	4.10	4.65
MAIN-OUT	8.86	2.49	2.79	6.28
MAIN-IN	4.76	1.16	2.23	2.20
MAIN-NORM	18.95	2.15	2.90	5.24

Table 3: Relative size of the components of the Chilean Web (2000-2003).

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

4 Analysis of Web Site Migration

In table 4 we show the migration of sites among the components. There are two ways of reading these tables. By columns we have from which component comes the sites in each component a given year. By rows, we can see where are today the sites of the components in the previous year. 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 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!).

2001 2000	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	959	724	140	305	11	61	24	286	218
OUT	195	1151	39	749	5	96	48	338	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
NEW	742	2128	901	10955	27	437	225	0	0

2002 2001	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	1209	315	105	39	1	8	4	133	148
OUT	896	1679	181	528	15	128	43	358	458
IN	232	96	281	188	1	22	16	127	277
ISLANDS	417	1346	714	5129	23	360	299	1053	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	94	171	36	158	1	22	8	0	0
NEW	1697	2672	2524	15023	41	592	830	0	0

2003 2002	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	UNKNOWN	DEAD
MAIN	2500	863	148	123	7	20	39	589	396
OUT	1006	2923	98	690	9	81	70	907	803
IN	675	327	910	483	6	15	197	482	814
ISLANDS	498	2316	796	9242	20	241	501	1862	5777
TUNNEL	20	31	1	7	0	0	3	14	9
TIN	102	514	28	183	10	50	15	165	150
TOUT	64	150	97	292	4	11	227	123	261
UNKNOWN	189	371	86	528	2	27	39	0	0
NEW	1976	2703	979	4091	24	308	341	0	0

Table 4: Component changes of sites from 2000 to 2003.

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. Figure 2 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 figure 3, shows what really happened, aggregating all the transitions in our data (blue arrows are sites that dissappear). 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. Taking the first choice in all three cases, as the most probable, we get the right side of figure 3. This means that a finer time granularity on the Web snapshots is needed to understand 3.4% of the transitions.

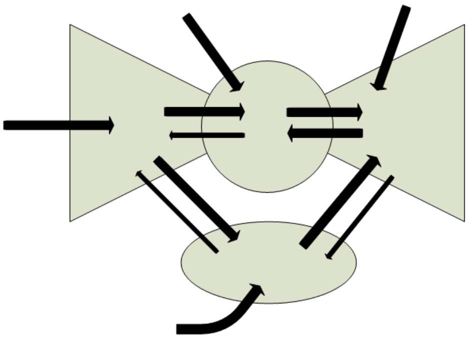

Figure 2: Expected migrations of websites in the Web structure.

Figure 4 shows the real migration of each site in the structure using one color for component. The order of the colors, from white to green is (NEW+UNKNOWN+DEAD, TIN, IN, MAIN, OUT, TOUT, TUNNEL, ISLANDS). From the possible 2,401 migration patterns, we found only 1,126 (47%) in the 56,699 sites. We can clearly see the growth in the white space at the left, being the transition NEW to ISLANDS the most frequent. The white space to the right are the UNKNOWN or DEAD cases.

Figure 5 shows the same, but keeping only the web sites 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. This subset is a zoom on the bottom part of the figure 4, that comprises 3,971 sites. Here we found 445 of 1296 possible migration patterns (34%), 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 50.8% of all cases, not including the third most frequent case, which are sites that are in OUT but one year were ISLANDS. We can notice also that there is almost no migration from IN to MAIN in opposition to what intuition predicts. Also, there are websites 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).

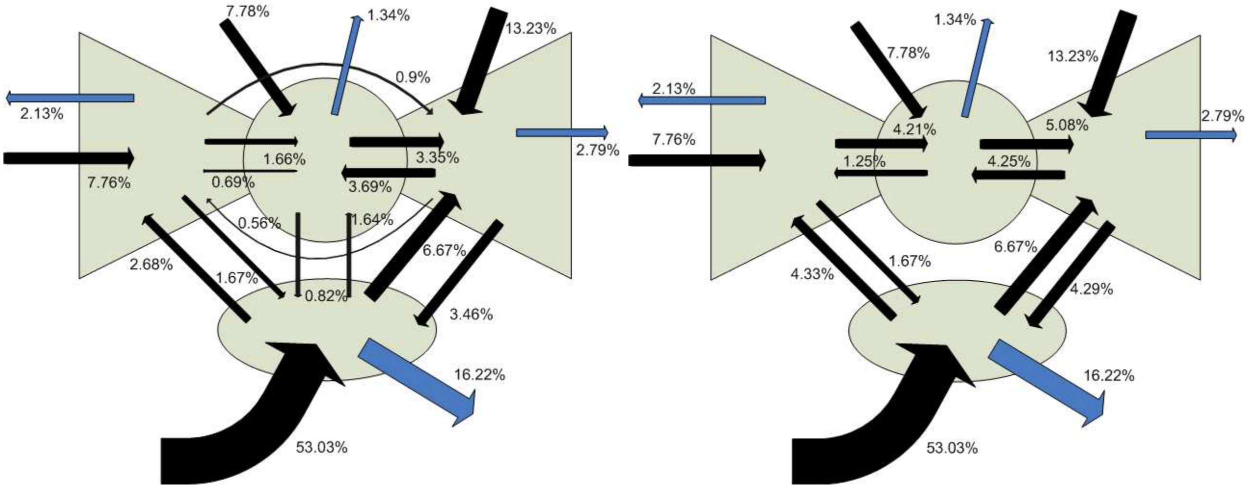

Figure 3: Aggregated real (left) and possible (right) migrations of websites in the Web structure.

5 Concluding Remarks

The overall number of sites of the Chilean Web is almost duplicating each year, as we believe that the 2003 data did not reflect the actual growth. That is the result of about a 100% increase plus a 20% death. So, using a simple model for Web site growth as $f_n = (\alpha - \beta) f_{n-1}$ where α is the growth rate and β the death rate, according to our results we have $f_n \approx 1.8 f_{n-1}$. However, the Web growths 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 a lot to do to understand how the composition of the structure changes, but perhaps there are no formal processes behind and it is just a transient phase. Another problem is the dynamics of the sites content. For example, the largest 100 sites (in pages) per year, involve 328 sites for all years (so there are many changes on content), and only 6 and 60 sites were in the top for 3 and 2 years, respectively. Although page count depends in crawling policies, we have used more or less the same policies all the time and the changes are quite radical.

Acknowledgements

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 P01-029-F from Mideplan, Chile, and Chilean Fondecyt Project 1020803 Grant from CONICYT.

References

[BYC01] Ricardo Baeza-Yates and Carlos Castillo. Relating web characteristics with link analysis. In *String Processing and Information Retrieval*. IEEE Computer Science Press, 2001.

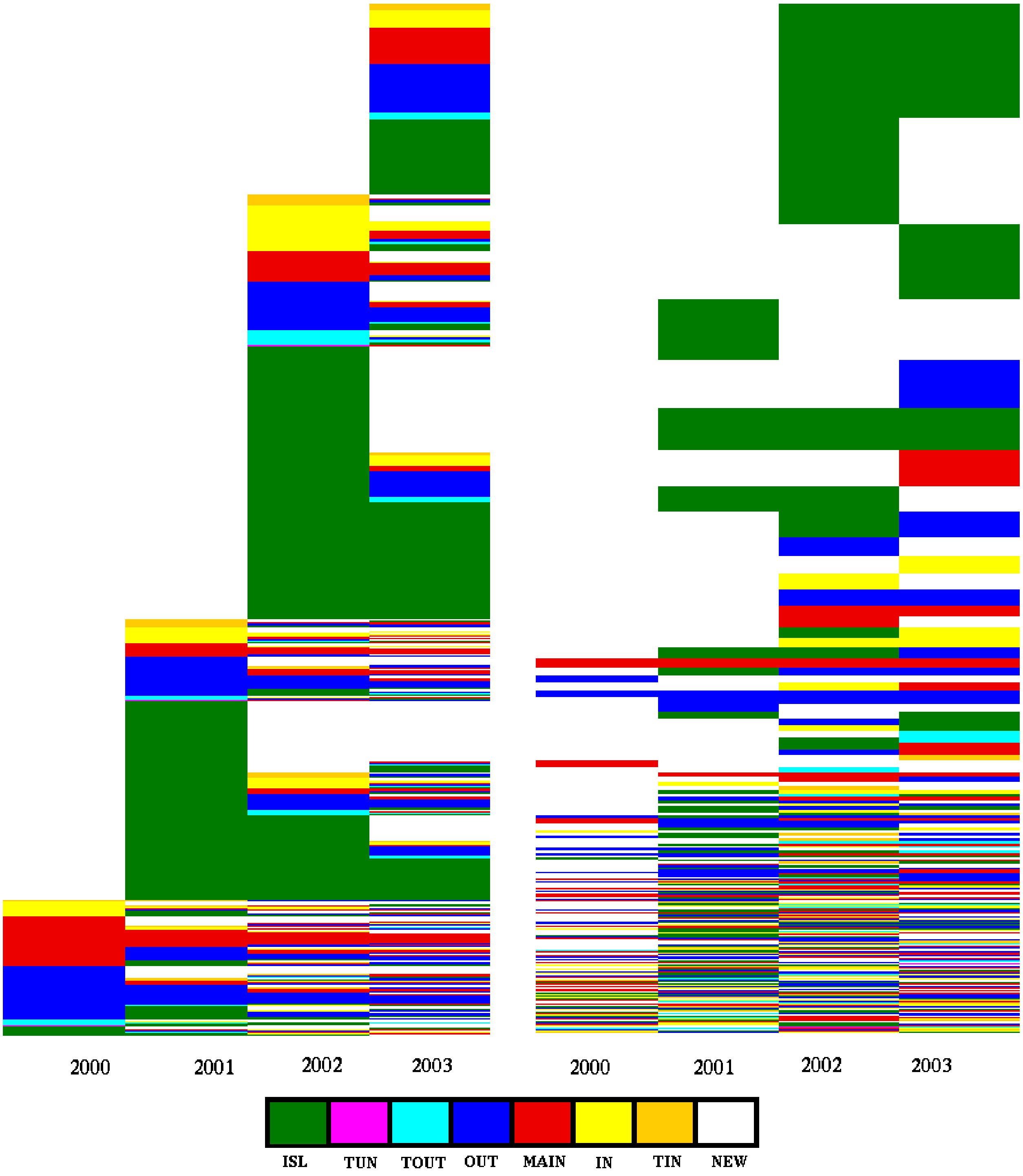

Figure 4: Migrations of websites in the structure (one column per year, one line per site, one color per component). Left side is sorted by color order, right side by case frequency.

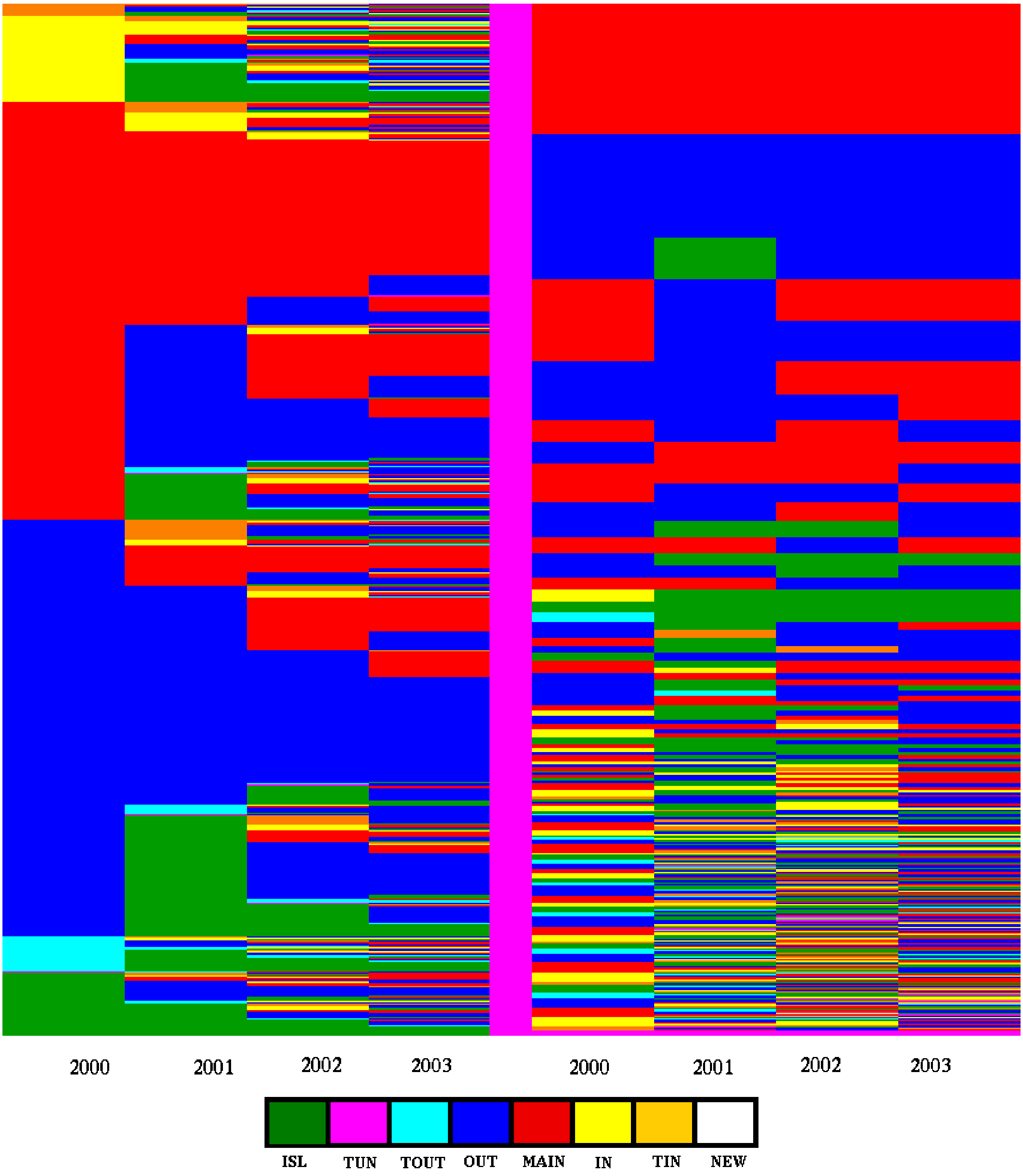

Figure 5: Migrations of websites in the structure considering only stable websites (one column per year, one line per site, one color per component). Left side is sorted by color level order, right side by case frequency.

[BYSJC02] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web dynamics, structure, and link ranking. In *String Processing and Information Retrieval*. Lecture Notes in CS, Springer, 2002.

[BYP03] Ricardo Baeza-Yates and Barbara Poblete, Evolution of the Chilean Web Structure Composition. In *First Latin American World Wide Web Conference*. IEEE CS Press, Santiago, Chile, November 2003.

[BKM$^+$00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, and A. Tomkins. Graph structure in the Web: Experiments and models. In *9th World Wide Web Conference*, 2000.