A Comparison of Sampling Techniques for Web Graph Characterization

Luca Becchetti Carlos Castillo Debora Donato* Adriano Fazzone

Università di Roma “La Sapienza”
Rome, Italy.

ABSTRACT

We present a detailed statistical analysis of the characteristics of partial Web graphs obtained by sub-sampling a large collection of Web pages.

We show that in general the macroscopic properties of the Web are better represented by a shallow exploration of a large number of sites than by a deep exploration of a limited set of sites. We also describe and quantify the bias induced by the different sampling strategies, and show that it can be significant even if the sample covers a large fraction of the collection.

1. INTRODUCTION

The number of Web pages that can be indexed by search engines is estimated in over 11.5×10^9 pages [22, 7]. The Web represents the greatest endeavour of all times in the field of collecting and sharing knowledge, and introduces significant challenges in retrieving, classifying and ranking its contents.

Web information retrieval techniques often make several assumptions on the properties of the Web, and characterization studies aim at providing a firm basis for those assumptions. Web characterization studies require representative samples of the Web, for instance, for comparing Web ranking or Web crawling techniques. In the case of Web crawling, the subsets obtained by sampling are used by researchers as a benchmark for testing different crawling strategies before employing them in effective crawling.

However, the computational resources for global scale crawling can be prohibitively large for most organizations, with this type of crawl having an estimated cost close of over US $1.5 Million [17], considering only the network connectivity costs.

The large amount of resources needed for performing a large crawl of the Web, and for analyzing the resulting data, can become even more problematic when studying **Web dynamics**. For studying the evolution of a changing Web, most researchers use periodic **snapshots** of the state of the Web at different times. Obtaining each such snapshot takes a large amount of network resources and disk space for storage, and this limits the size and/or frequency of the snapshots.

In this context, processing time, disk space and network usage can be reduced greatly by using adequate sampling methods, able to extract small but representative samples from the Web pages. The research question this paper addresses is: what is the most efficient and effective way of sampling a subset of a the web of a country?

By **efficient**, we mean that we want to sample only a fraction of the graph, by **effective**, we mean that the sample has to be representative of the link structure of the overall graph. The main contributions of this paper are:

- We compare different methods for Web sampling by sub-sampling a large Web collection.
- We evaluate each method by studying several link-based metrics over the resulting sample.
- We show that deep crawling over a limited set of Web sites is not appropriate for studying the characteristics of the macroscopic structure of the Web, and that it is much better to have a representative sample of many different Web sites, even if the exploration has to be more shallow to download the same amount of pages.

The next section discusses previous work on this topic, and Section 3 presents the collection and tools we use. Section 4 compares the sampling strategies by presenting a detailed statistical analysis of the obtained sub-samples. Finally, Section 5 presents our conclusions.

2. PREVIOUS WORK

Our work is concerned mostly with link analysis and the macro structure of the link graph and the “bow-tie” structure depicted by Broder et al. [15]. This structure repeats at smaller scales, as observed in [18], and it can be further decomposed into smaller structures [20, 6].

There are basically two types of sampling techniques that are used for analyzing the Web. Sampling by random walks, and vertical sampling.

Sampling by random walks implies starting at a given page and follow out-links at random; this type of sampling is biased towards pages with high connectivity, as they are more likely to be reachable by a random walk. A technique for countering that bias is due to Henzinger et al. [24]; their method involves a second phase in which the nodes found

*Corresponding author, donato@dis.uniroma1.it

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

LinkKDD'06, August 20, 2006, Philadelphia, Pennsylvania, USA.
Copyright 2006 ACM 1-59593-446-6/06/0008 ...$5.00.

during the random walk are re-sampled. The sampling probability for the second pass is inversely proportional to the PageRank the nodes get in the first pass, so this approximates a uniform sampling. The bias induced by a random walk like this on a Web graph has also been studied by Boldi et al. [12].

Vertical sampling involves delimiting previously a set of Web pages to be crawled, and then studying the obtained samples. This delimitation is typically done considering a restriction on the domain name of the hosts crawled. This can be done either by considering only a large first-level domain (such as `brown.edu` [31], `nd.edu` [2], etc.), or by taking the web domain of a country, this is the set of pages under a common country-code top-level domain. There have been several studies of vertical sampling of countries including [11, 21, 5] among many others (for a recent survey of several of them, see [3]). A related study that considers links among national domains is presented in [10].

A study that is related to ours is [32], in which the authors study how several Web collections taken from different subsets of the Web differ in terms of their macro structure. Those collections were basically disjoint, while here we compare subsets of a larger sample.

When characterizing a large graph, many aspects of the topology of the graph can be studied. Reference [16] presents a comprehensive survey on the type of statistics that can be extracted.

3. EXPERIMENTAL FRAMEWORK

This section introduces the test collection we use for extracting the sub-samples and the strategies used for sampling.

3.1 Collection

We are using a collection of pages under the country-code of Slovakia (`SK`), obtained by a breadth-first crawl carried by the Laboratory of Web Algorithmics[1] in June 2005.

The crawling found 13,478 hosts starting from a large set of seed URLs, with limits of up to 100,000 pages per host, and 16 levels of links from the starting pages. Compared to other crawls used for Web characterization [11, 3, 21] this is a very deep crawl; in other crawls, up to 20,000 pages per host and 8 links of depth are typical limits.

The obtained graph has 50.6 million nodes and 1.9 billion arcs, this is roughly 38 links per page. The full graph uses about 7.4 GiB of disk space uncompressed (this is, using 4 bytes per arc).

We note that our view of the Web is itself a sample obtained by BFS, as it is not possible to obtain a copy of the full Web, which may contain, for all practical purposes, an infinite number of pages. A breadth-first crawler with certain limits is the standard tool for studying characteristics of the Web. This paper focuses on finding out how to obtain the same characteristics of a large BFS crawl by sampling less pages, and on measuring the biases that are induced by different sampling methods.

3.2 Tools

We used the web graph compression framework [13] so the graph uses only 728 MiB on disk. Remarkably, this is **less than 2 bits** per arc on average. We obtained the samples by processing the compressed graph. To extract strongly connected components and compute most of the statistics presented here, we used the semi-external algorithms implemented in the COSIN library [27].

With respect to the hardware, the most important aspect was to have a reasonable amount of RAM to speed up the semi-external algorithms of the COSIN library. We used a normal PC with a Pentium-4 processor at 2.8 GHz and 2 GiB of RAM. We stored several partial snapshots of the Web graph and used about 150 GB of disk space in two SCSI disks in RAID-1 (mirror), although we could have used less by studying the partial graphs sequentially and then deleting them. The total running time for all the analysis was roughly two days, and link-based ranking computation was the most expensive part.

3.3 Sampling methods

Our sampling methods pick some nodes according to some schema, and then include an edge in the sampled graph if both its source and destination nodes were picked. We fixed the fraction of nodes to be picked by each sampling method to 0.1, 0.2, 0.5, 0.8 and 0.9.

The schemes used for picking nodes are the following:

Uniform random sampling Pages are chosen uniformly at random with a certain probability. This sampling strategy is actually not possible for a standard Web crawler that must discover pages by following links, but we used it as a baseline for the comparison.

Sampling by selecting entire sites Sites are chosen uniformly at random with a certain probability, and all of the pages inside a site are included in the sample. We continued this process until we have a predefined fraction of the nodes in the graph. This is feasible in practice and the crawler must be instructed not to follow links outside the sampled sites.

Sampling by breadth-first search (BFS) All the initial pages of sites (the starting or home page, located in the root directory of the site and typically named "`/index.*`" or just "`/`") were sampled. We consider those pages to be at depth equal to 1. All of the pages that are linked by those pages are considered to have depth equal to 2, and so on. This strategy simulates a BFS search that stops when a given threshold of nodes is reached.

Sampling by OPIC The OPIC algorithm (online page-importance computation) was introduced by Abiteboul et al. [1] as an algorithm for ranking pages while discovering them. It can be seen as a biased breadth-first search in which the pages that are highly linked are more likely to be chosen. To implement this algorithm in external memory, we approximated it by re-calculating page importance 20 times during the simulated crawl (instead of after inserting every node).

[1] Laboratory of Web Algorithmics, *Dipartimento di Scienze dell'Informazione*, *Universitá degli studi di Milano*, `http://law.dsi.unimi.it/`.

4. EVALUATION

This section presents the empirical evaluation of the sampling methods.

4.1 Overlap

There is some degree of overlap between the sampling strategies. In Table 1 we show the overlap between sampling by breadth-first search (BFS), by OPIC, and by sites. The overlap between two sampling strategies is measured as the fraction of nodes that are sampled by both. In this case, the measurement is done over the largest strongly connected component of the resulting graph.

		Strategy	
Strategy	Sample size	by OPIC	by sites
by BFS	10%	60.5%	57.4%
	20%	65.9%	64.9%
	50%	69.3%	71.8%
	80%	59.5%	71.1%
	90%	71.0%	71.7%
by OPIC	10%	-	50.0%
	20%	-	59.0%
	50%	-	66.1%
	80%	-	70.0%
	90%	-	68.2%

Table 1: Overlap of the samples obtained by varying the sampling strategy. The sample sizes are given relative to the entire collection.

If we focus in a particular cut-off value, for instance 50%, we can see that the overlap among our samples is between 2/3 and 3/4, meaning that the obtained samples are neither entirely equivalent nor entirely disjoint. OPIC is more similar to breadth-first search than to sampling by sites. This similarity is also observed when analyzing other measures.

4.2 Microscopic measures

Before studying the global-scale connectivity of the graph, we study link metrics that can be measured in every node. The most natural metric to start with is the degree. We begin by analyzing the total degree, considering the sum of the in-degree and the out-degree. In our sample, the average degree is 38.1 links.

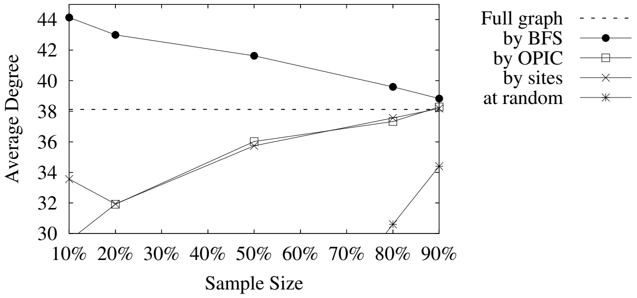

Figure 1: Average degree in the different samples.

As we can see in Figure 1, the sampling by BFS consistently overestimates the average degree, while the other sampling strategies tend to underestimate it. Obviously the random sample generates a very disconnected graph (remember that we keep in the resulting graph and edge if both ends of the edge are sampled) and thus underestimates the degree even more.

The overestimation of the degree observed in the sample obtained by BFS, means that pages that are topologically close to the "root" page of a Web site are more connected than "deeper" pages. This is consistent with the observation that most of the inter-site links (links among pages in different sites) are pointing to the home page of a site [10].

As for the distribution of degree, for both the in-degree and out-degree it is known that it follows a power-law; we used Hill's estimator [25] for computing the exponent. This requires a step in which we plot a cumulative distribution and manually assess which is the range in which the data exhibits a power-law distribution. Table 2 shows the result. In the case of out-degree the distribution is typically log-normal or double-pareto [28] and we are providing the exponent for the tail of the distribution (pages with high out-degree).

Strategy	Sample size	In-degree	Out-degree
Random	10%	1.81	3.91
	20%	1.80	3.08
	50%	1.91	3.76
	80%	2.30	3.31
	90%	2.12	3.44
by BFS	10%	1.81	3.09
	20%	2.02	3.19
	50%	1.80	3.10
	80%	2.15	3.07
	90%	1.80	3.60
by OPIC	10%	2.13	3.19
	20%	2.02	3.61
	50%	1.84	3.26
	80%	1.84	3.10
	90%	2.25	3.73
by sites	10%	1.86	3.97
	20%	2.06	3.05
	50%	1.88	3.48
	80%	2.22	3.46
	90%	2.23	3.36
Full collection	100%	1.83	3.66

Table 2: Exponents of the power-law exponent for the distribution of the in-degree and out-degree in the collection.

In all the sampling strategies, the exponents are kept within a reasonable range of the actual value (except for the out-degree exponent). There is no clear advantage for any of the sampling strategies, and there is a large variability in the obtained exponent for all of them. Even the strategy that samples nodes at random performs reasonably well. The fact that is it relatively easy to obtain a power-law distribution by sampling a large Web graph can also be compared with the observation in [19] that this distribution is also relatively easy to obtain by using synthetic graph models such as preferential attachment [8].

A related measure is **edge reciprocity**, that is, the fraction of edges that are reciprocal. This is measured by computing the overlap between the out-neighbors and in-neighbors for all nodes in the graph. This can be done easily by sequentially scanning both the graph and its transposed version.

In the full graph, the edge reciprocity is 0.12, meaning that on average 12% of the out-neighbors of a node have a link back to it.

As shown in Table 3, all the strategies perform reasonably well except for the small sub-sample (10%) of the OPIC strategy. This overestimates the edge reciprocity by a factor of 50%, while breadth-first search underestimated the edge reciprocity by a factor of 20% when sub-sampling half of the nodes.

	Sample size			
Strategy	10%	50%	90%	100%
Random	0.13	0.13	0.12	
by BFS	0.13	0.10	0.13	
by OPIC	0.18	0.11	0.13	
by sites	0.11	0.15	0.13	
Full collection				0.12

Table 3: Edge reciprocity in the different sampling strategies.

Assortativity: the degree of the nodes in a large scale-free network induces a natural "hierarchy" that can be used to define different classes of nodes. A network in which most nodes are connected to other nodes in the same class (for instance, most of the connections of highly-linked are to other highly-linked nodes) is called "**assortative**" and a network in which the contrary occurs is called "**disassortative**". This distinction plays an important role in the propagation of epidemics [23].

We measured the correlation coefficient between the in-degree of a page and the average in-degree of its neighbors. We also measured the assortativity of the out-degree. As per Table 4, the Web graph is slightly disassortative in the in-degree, a phenomenon that has been observed in other scale-free networks [29].

		Sample size			
Strategy	Type	10%	50%	90%	100%
by BFS	Indeg.	-0.045	-0.021	-0.018	
	Outdeg.	0.279	0.140	0.922	
by OPIC	Indeg.	-0.012	-0.023	-0.018	
	Outdeg.	0.279	0.173	0.921	
by sites	Indeg.	-0.055	-0.073	-0.050	
	Outdeg.	0.061	0.970	0.920	
at random	Indeg.	-0.055	-0.044	-0.017	
	Outdeg.	0.920	0.911	0.916	
Full collection	Indeg.				-0.017
	Outdeg.				0.917

Table 4: Assortativity obtained with different sampling strategies. This includes the assortativity of the in-degree and of the out-degree.

Interestingly, even if we sample 90% of the nodes by sites we still obtain an overestimation of the assortativity coefficient of the in-degree by a factor of 3. The sampling by sites actually gives the worst approximation in our experiments when compared to the other techniques.

On the other hand, if we observe the correlation coefficient of the out-degree of pages that are linked, we observe an assortative behavior (> 0.9). This means that pages that are connected to each other are very likely to have a similar out-degree. This fact is better captured by the random sampling and the sampling by sites than with the other techniques. All the sampling methods detect the fact that the network is assortative in its out-degree even with a small sample, but they differ greatly in their estimation of the actual coefficient.

PageRank distribution: the tail of the distribution of PageRank [30] scores in the graph follows a power-law. A power-law has also been observed by some authors in the distribution of the values of the authority score given by a static (global) version of the HITS [26] algorithm.

Table 5 shows the results of calculating the exponent in the tail of the distribution of PageRank in the different samples. Surprisingly, even the random strategy achieves a good performance in approximating this exponent, and the worst approximations are given by (small) sub-samples created with the strategy that samples entire sites, and with the strategy based on OPIC.

Strategy	Sample size	θ	Iterations	Residual $[\times 10^{-5}]$
Random	10%	2.23	30	60
	20%	2.29	30	30
	50%	2.28	19	9.629
	80%	2.30	18	9.587
	90%	2.31	15	9.869
by BFS	10%	2.13	12	7.436
	20%	2.36	13	8.934
	50%	2.27	15	8.981
	80%	2.25	15	9.494
	90%	2.28	15	9.804
by OPIC	10%	2.56	12	9.799
	20%	2.61	14	8.891
	50%	2.33	15	9.321
	80%	2.27	15	9.548
	90%	2.27	15	9.745
by sites	10%	1.88	17	9.871
	20%	2.23	16	9.564
	50%	2.27	20	9.028
	80%	2.26	17	9.720
	90%	2.27	15	9.393
Full collection	100%	2.31	30	1.668

Table 5: Power law exponent θ of the tail of the distribution of PageRank, number of iterations required for convergence and residual after the computation. The max. number of iterations was set to 30 and the maximum residual to 10^{-6}.

The iterations required for the convergence of PageRank are also depicted in Table 5. We imposed two limits: 30 iterations or less than 10^{-6} of L2-norm in the difference between two consecutive iterations. The more disconnected the graph is (as in the strategy that samples nodes at random and the strategy that goes by sites), the longer the computation takes.

PageRank/Degree correlation: in general, there is no correlation between in-degree and out-degree in our sample, but there is clearly a correlation between in-degree and PageRank (this is an indication that despite its drawbacks, in-degree is also reliable as a measure for ranking [33, 14]). Sampling at random or by sites provides a poorer estimator of the correlations between PageRank, indegree and outdegree in this graph, as shown in Table 6.

Strategy	Size	Correlation In/Out	In/PR	Out/PR
by BFS	10%	0.060	0.709	0.024
	20%	0.036	0.704	0.019
	50%	0.023	0.746	0.008
	80%	0.032	0.746	0.005
	90%	0.037	0.745	0.005
by OPIC	10%	0.088	0.734	0.022
	20%	0.051	0.706	0.015
	50%	0.025	0.745	0.005
	80%	0.032	0.754	0.005
	90%	0.038	0.745	0.005
by sites	10%	0.031	0.523	0.005
	20%	0.043	0.676	0.020
	50%	0.073	0.561	0.005
	80%	0.033	0.725	0.005
	90%	0.049	0.609	0.006
at random	10%	0.050	0.599	0.010
	20%	0.047	0.567	0.007
	50%	0.040	0.616	0.006
	80%	0.033	0.739	0.004
	90%	0.034	0.747	0.004
Full collection	100%	0.034	0.733	0.004

Table 6: Correlation between PageRank, in-degree and out-degree.

4.3 Macroscopic measures

The web graph has a well defined structure that was made evident by the study in [15], where its bow-tie shape was depicted. The Web graph nodes are organized in five different sets. The first set is an unique large strongly connected component, known as CORE. Starting from this set, we can identify a set of nodes that can reach the nodes in the CORE but that can not be reached from them. This set is called IN. Conversely, we can identify the OUT set comprised by the nodes that can be reached by the ones in the CORE but can not reach them.

A fourth set, called TENDRILS, consists of nodes not in the CORE that are reachable from the nodes in IN, or can reach the nodes in OUT. "TUBES" is the intersection of TENDRILS-IN and TENDRILS-OUT. The last set, DISC, is comprised by all the remaining nodes organized in a number of independent SCCs. All the components we have described are depicted in Figure 2.

The dimensions of the bow-tie components are computed using a simple algorithm that, exploiting the fact that the largest strongly connected component has a large size compared to the other strongly connected components, allows to detect the CORE easily by using sequential forward and backward traversals of the graph.

In our graph, the CORE represents roughly 71% of the graph and the OUT component about 29% of the graph. The other components are very small, as shown in Table 7. Does this mean that our graph is significatively different than the one analyzed in [15] and other Web characterization studies? We think not, for the following reason: we observe that, as the crawl goes deeper and deeper (in relation to the number of starting points), all the new nodes that are found lie in the CORE and OUT components (possibly also moving some nodes from OUT to CORE as new connections are discovered), so the relative size of the IN component goes to zero. In the case of [15] the crawl is quite deep but the number of starting points is also very large, this explains the size of their IN component. In Figure 4 we can see the evolution of the relative sizes of the components.

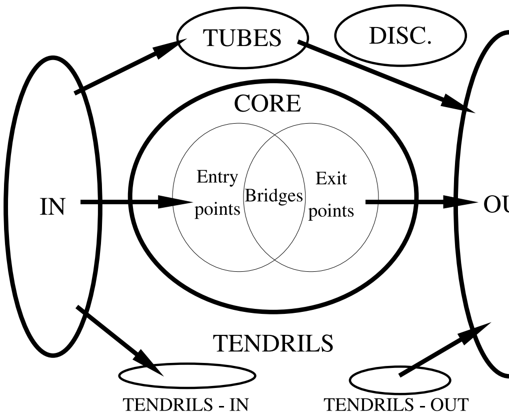

Figure 2: Bow-tie structure of the Web.

Component	Size	%
CORE	35,874,391	70.8%
IN	65,570	0.1%
OUT	14,668,250	29.0%
TENDRILS	21,545	0.0%
TUBES	0	0.0%
DISC.	6,398	0.0%
Total	50,636,154	100.0%

Table 7: Relative sizes of the bow-tie components in the full graph.

The best approximation of the relative sizes of the CORE, IN and OUT components is given by BFS, and the strategy that samples sites performs quite poorly in this task. Interestingly, the strategy that samples by OPIC provides a better approximation of the size of the OUT component, possibly because it is biased towards high-quality nodes that have many in-links. Table 8 at the end of this paper details how the components evolve as the sampling size is increased in the different strategies.

Extended bow-tie: Refinements of the bow-tie model were proposed by [20, 4]; these models search for sub structures of the largest SCC. As shown in Figure 2, in the CORE component we call a page an *entry point* if it is directly reachable from IN and an *exit point* if it can reach OUT di-

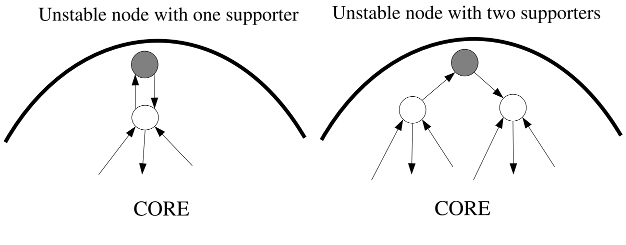
Unstable node with one supporter
Unstable node with two supporters
CORE
CORE

Figure 3: Left: depiction of an unstable node that forms a "petal". Right: depiction of an unstable node that forms a "connector".

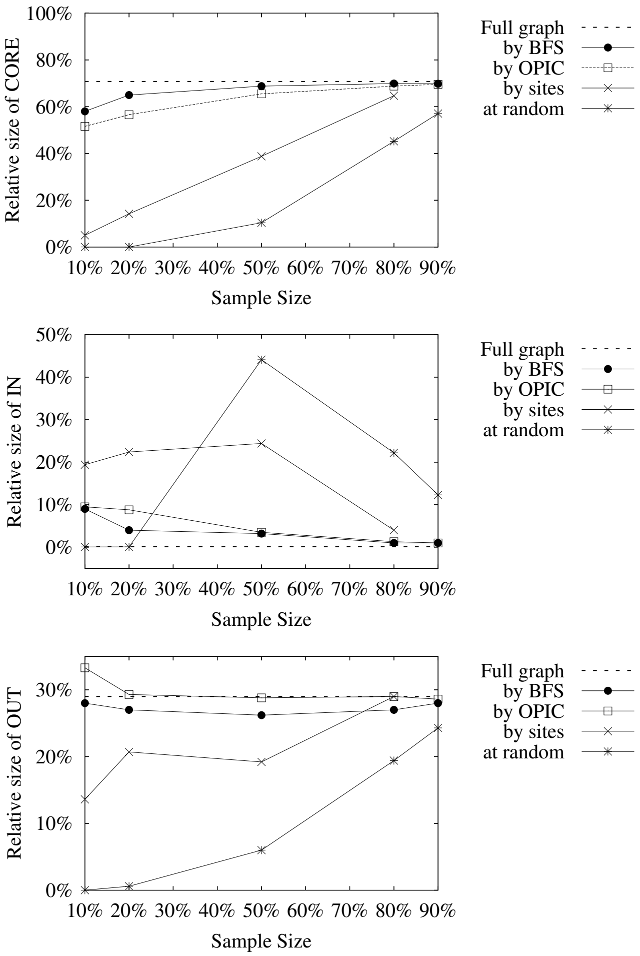

Figure 4: Relative sizes of the CORE, IN and OUT components of the bow-tie structure.

rectly. We also observe several nodes that we call *unstable*, that belong to the CORE but are only attached to it by two links. As shown in Figure 3, we observe two types of unstable nodes, that we call *petals* and *connectors*.

In general, all the strategies make a good work at capturing these components with some advantage for OPIC. Table 9 at the end of the paper details these findings.

Levels in IN, OUT: We say that nodes that are at 1-click distance from any node in the CORE are in the level 1, at 2-clicks in the level 2 and so on. We measured the number of nodes in each level in the IN and OUT components. This feature is really important since it provides a direct criteria to decide how deep we have to go in order to collect a predefined amount of nodes in the IN and OUT sets.

Table 10 (at the end of the paper), shows that sampling by sites performs very bad in finding the distribution of pages per levels, and also that BFS is better than OPIC in estimating the distribution of pages into levels in the OUT component of the graph.

5. CONCLUSIONS

After analyzing these results the conclusion seems robust: for many measures the BFS and OPIC strategies perform much better than sampling by sites. This seems to indicate that many characteristics of the connectivity of the Web arise from the interaction among many different sites, presumably under the control of different Web site administrators.

Even a very deep crawl fails to capture some important characteristics of the Web graph if it is done over a limited set of sites. On the other hand, the most significant problem we observed when sampling by BFS is that it tends to overestimate the average degree of pages.

Our findings seem to indicate that for large-scale Web characterization studies, **the set of starting pages must be as large as possible**. During this study, we have observed again and again that it is much more important to have pages from many different Web sites than to crawl thousands of pages from every Web site in the sample.

We have also observed an interesting phenomenon in the distribution of the relative sizes of the components in the bow-tie structure: that **the IN component shrinks as the crawling goes by**. This happens because the crawler discovers more new pages in the CORE and OUT component than in IN. In general, the existence of a relatively large IN component in the bow-tie structure as depicted in [15], may be just a direct consequence of the sampling strategy used. This merits further study.

For future work, we plan to extend both the coverage and scope of this research, in terms of considering more metrics (such as clustering coefficient) and to test our findings by sampling other large Web graphs. We are also interested in the study of the dynamics of the Web.

Acknowledgments

We thank the Laboratory of Web Algorithmics at the Università degli Studi di Milano for providing several Web collections for research.

6. REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the twelfth international conference on World Wide Web*, pages 280–290, Budapest, Hungary, 2003. ACM Press.

[2] R. Albert, H. Jeong, and A. L. Barabási. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[3] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of national web domains. Technical report, Universitat Pompeu Fabra, July 2005.

[4] R. Baeza-Yates, C. Castillo, and F. S. Jean. *Web Dynamics*, chapter Web Dynamics, Structure and Page Quality, pages 93–109. Springer, 2004.

[5] R. Baeza-Yates, C. Castillo, and V. López. Characteristics of the Web of Spain. *Cybermetrics*, 9(1), 2005.

[6] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web structure, dynamics and page quality. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, volume 2476 of *Lecture Notes in Computer Science*, Lisbon, Portugal, 2002. Springer.

[7] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 367–376, New York, NY, USA, 2006. ACM Press.

[8] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[9] A. A. Benczúr, K. Csalogány, D. Fogaras, E. Friedman, T. Sarlós, M. Uher, and E. Windhager. Searching a small national domain – a preliminary report. In *Poster Proceedings*

		Sample Size				
Strategy	Component	10%	20%	50%	80%	90%
by BFS	CORE	58,4%	65,3%	68.8%	69.9%	69.8%
	IN	9,3%	4,8%	3.2%	1,6%	1,1%
	OUT	28,1%	27,4%	26.2%	27,4%	28,3%
by OPIC	CORE	51.6%	56.6%	65.5%	68.8%	69.6%
	IN	9.5%	8.8%	3.5%	1.3%	1.0%
	OUT	33.3%	29.3%	28.8%	29.0%	28.6%
by sites	CORE	5.0%	14.2%	38.8%	64.7%	67.1%
	IN	19.4%	22.4%	24.4%	4.0%	0.7%
	OUT	13.6%	20.7%	19.2%	29.0%	31.8%
Random	CORE	0.04%	0,01%	10.4%	45.2%	57.1%
	IN	0.06%	0.08%	44.1%	22.2%	12.3%
	OUT	0.01%	0.6%	6.2%	19.4%	24.3%
		Sample size = 100%				
Full collection	CORE	70.8%				
	IN	0.1%				
	OUT	29%				

Table 8: Relative sizes of the components obtained by the sampling strategies.

Strategy	Sample size	EntryPoints	ExitPoints	Bridge	Connectors	Petals
Random	10%	34%	17%	9%	35%	34%
	20%	48%	15%	7%	0%	0%
	50%	28%	69%	19%	1%	0%
	80%	15%	69%	11%	3%	0%
	90%	10%	71%	7%	4%	0%
by BFS	10%	1%	7%	1%	2%	0%
	20%	1%	72%	0%	3%	0%
	50%	0%	68%	0%	4%	0%
	80%	0%	68%	0%	4%	0%
	90%	0%	68%	0%	4%	0%
by OPIC	10%	1%	57%	1%	3%	0%
	20%	1%	61%	0%	3%	0%
	50%	0%	65%	0%	4%	0%
	80%	1%	67%	0%	4%	0%
	90%	0%	68%	0%	4%	0%
by sites	10%	3%	63%	1%	2%	0%
	20%	2%	73%	2%	3%	0%
	50%	1%	73%	1%	3%	0%
	80%	1%	67%	1%	3%	0%
	90%	0%	69%	0%	4%	0%
Full collection	100%	0%	68%	0%	4%	0%

Table 9: Results of measurements in the extended bow-tie model.

of Conference on World Wide Web, Budapest, Hungary, May 2003.

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

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

[12] P. Boldi, M. Santini, and S. Vigna. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. In *Proceedings of the third Workshop on Web Graphs (WAW)*, volume 3243 of *Lecture Notes in Computer Science*, pages 168–180, Rome, Italy, October 2004. Springer.

[13] P. Boldi and S. Vigna. The webgraph framework: Compression techniques. In *Proceedings of the 13th conference on World Wide Web*, New York, NY, USA, 2004. ACM Press.

[14] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297, February 2005.

[15] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. 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. ACM Press.

[16] L. Costa, F. A. Rodrigues, and G. a. Travieso. Characterization of complex networks: A survey of measurements, Jun 2005.

[17] N. Craswell, F. Crimmins, D. Hawking, and A. Moffat. Performance and cost tradeoffs in web search. In *Proceedings of the 15th Australasian Database Conference*, pages 161–169, Dunedin, New Zealand, January 2004.

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

[19] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Simulating the webgraph: a comparative analysis of models. *Computing in Science & Engineering [see also IEEE Computational Science and Engineering]*, 6(6):84–89, 2004.

[20] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. In *Eigth international workshop on the Web and databases WebDB*, Baltimore, USA, June 2005.

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

Levels in the IN component

Strategy	Sample size	Levels	level 1	level 2	level 3	level 4	level 5
Random	10%	4	87%	11%	2%	0%	-
	20%	3	99%	0%	0%	-	-
	50%	34	85%	7%	4%	2%	1%
	80%	14	94%	4%	1%	0%	0%
	90%	23	95%	5%	1%	1%	0%
by BFS	10%	7	70%	17%	3%	0%	0%
	20%	7	81%	5%	1%	11%	1%
	50%	8	87%	10%	1%	0%	0%
	80%	8	86%	12%	1%	0%	0%
	90%	7	85%	12%	1%	0%	0%
by OPIC	10%	10	71%	23%	4%	1%	1%
	20%	8	77%	16%	2%	5%	0%
	50%	9	87%	12%	1%	0%	0%
	80%	7	84%	14%	2%	0%	0%
	90%	5	84%	14%	1%	0%	0%
by sites	10%	30	47%	23%	7%	7%	2%
	20%	15	24%	40%	8%	17%	3%
	50%	38	73%	15%	9%	2%	1%
	80%	106	78%	19%	2%	0%	0%
	90%	5	86%	12%	2%	0%	0%
Full collection	100%	7	86%	14%	0%	0%	0%

Levels in the OUT component

Strategy	Sample size	Levels	level 1	level 2	level 3	level 4	level 5
Random	10%	3	90%	9%	0%	-	-
	20%	24	0%	0%	2%	5%	13
	50%	462	28%	6%	13%	20%	14%
	80%	339	47%	6%	9%	14%	9%
	90%	230	21%	2%	4%	6%	4%
by BFS	10%	10	34%	5%	25%	26%	7%
	20%	11	50%	4%	17%	20%	7%
	50%	14	51%	6%	11%	18%	11%
	80%	17	52%	6%	8%	14%	9%
	90%	18	52%	6%	8%	14%	8%
by OPIC	10%	19	22%	9%	25%	30%	10%
	20%	17	28%	7%	18%	27%	14%
	50%	18	41%	6%	12%	18%	12%
	80%	18	49%	7%	9%	14%	9%
	90%	18	51%	6%	8%	13%	9%
by sites	10%	23	17%	1%	3%	9%	13%
	20%	38	25%	2%	7%	16%	14%
	50%	24	22%	2%	3%	7%	5%
	80%	20	50%	7%	8%	13%	8%
	90%	20	49%	6%	8%	13%	9%
Full collection	100%	19	51%	7%	8%	13%	8%

Table 10: Levels in the IN and OUT components.

[22] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan, 2005. ACM Press.

[23] S. Gupta, R. M. Anderson, and R. M. May. Networks of sexual contacts: implications for the pattern of spread of hiv. *AIDS*, 3(12):807–817, December 1989.

[24] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near–uniform url sampling. In *Proceedings of the Ninth Conference on World Wide Web*, pages 295–308, Amsterdam, Netherlands, May 2000. Elsevier Science.

[25] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174, 1975.

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

[27] S. Millozzi, D. Donato, L. Laura, and S. Leonardi. Cosin tools: a library for generating and measuring massive webgraphs. Available online at `http://www.dis.uniroma1.it/~cosin/html_pages/COSIN-Tools.htm`, 2003.

[28] M. Mitzenmacher. Dynamic models for file sizes and double pareto distributions. *Internet Mathematics*, 1(3):305–333, 2003.

[29] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701+, October 2002.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

[32] A. M. Serrano, A. Maguitman, M. Boguna, S. Fortunato, and A. Vespignani. Decoding the structure of the www: facts versus sampling biases, Nov 2005.

[33] T. Upstill, N. Craswell, and D. Hawking. Predicting fame and fortune: Pagerank or indegree? In *Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pages 31–40, Canberra, Australia, December 2003.