**EPJ manuscript No.**
(will be inserted by the editor)

# Large scale properties of the Webgraph*

Debora Donato, Luigi Laura, Stefano Leonardi, and Stefano Millozzi

Dipartimento di Informatica e Sistemistica, Universitá di Roma "La Sapienza", Via Salaria 113, 00198 Roma Italy. E-mail: {donato, laura, leon, millozzi}@dis.uniroma1.it

Received: date / Revised version: date

**Abstract.** In this paper we present an experimental study of the properties of web graphs. We study a large crawl from 2001 of 200M pages and about 1.4 billion edges made available by the WebBase project at Stanford [17]. We report our experimental findings on the topological properties of such graphs, such as the number of bipartite cores and the distribution of degree, PageRank values and strongly connected components.

**PACS.** 89.20.Hh World Wide Web, Internet – 89.75.Fb Structures and organization in complex systems

## 1 Introduction

The *Webgraph* is the graph whose nodes are (static) Web pages and edges are (directed) hyperlinks between pages. The study of the *Webgraph* has recently attracted a large interest in the scientific community, primarily motivated by the developing of innovative Web search technologies. The analysis of the link structure of the Web is indeed at the basis of important algorithms such as PageRank [3] and HITS [8] for ranking Web documents returned by a query to a search engine.

The research in this area has primarily focussed on the study of the statistical and topological properties of the Webgraph through the experimental analysis of large crawls of the Web. This analysis has shown the ubiquitous presence of *power law* distributions in he Webgraph, considered by statistical physicists as a typical signature of the presence of scale-free properties in the structure. The first observation of this nature was made by Barabasi and Albert [2] and by Kumar et al. [10] that studied the distribution of the indegree of the Webgraph. This observation has been confirmed by several later experiments, as for example Broder et al. [4] on a crawl of 200M pages from 1999 by Altavista. More formally, the probability that the indegree of a vertex is $i$ is distributed as $Pr_u$[in-degree(u)= $i$]$\propto 1/i^{\gamma}$, for $\gamma \approx 2.1$. In [4], the outdegree of a vertex was also shown to be distributed according to a power law with exponent roughly equal to 2.7 with the exception of the initial segment of the distribution. The number of edges observed in the samples of the Webgraph is about equal to 7 times the number of vertices.

A second important research line has concentrated on the development of new probabilistic models able to generate synthetic graphs holding the properties observed in practice. These properties cannot be recognized in the classical random graph model of Erdös and Rényi (ER) [7] that for instance does not show any *power law* distribution on the degree. Moreover, the ER model is a static model, while the Webgraph evolves over time when new pages are published or are removed from the Web.

Albert, Barabasi and Jeong [1] initiated the study of random evolving networks by presenting a model in which at every discrete time step a new vertex is inserted in the graph. The new vertex connects to a constant number of previously inserted vertices chosen according to the so called *preferential attachment* rule, i.e. with probability proportional to the in-degree, . This model shows a power law distribution on the in-degree of the vertices with exponent roughly equal 2, when the number of edges that connect every vertex to the graph is 7.

Broder et al. [4] presented a fascinating picture of the Web's macroscopic structure, a *bow-tie* shape composed of 5 main regions including a large strongly connected component (SCC) spanning about 28% of the vertices. The study of a large sample from Alexa has also shown the existence of a surprising large number of dense subgraphs, specifically bipartite cliques, of moderately small size [10]. The study of such structures was aimed to trace the emergence of hidden *cyber-communities*. A bipartite clique is interpreted as the core of a community interested in a specific subject, defined by a set of fans, all pointing to a set of centers/authorities, and the set of centers, all pointed to by the fans. Over 100,000 such communities have been

* Partially supported by the Future and Emerging Technologies programme of the EU under contracts number IST-2001-33555 COSIN "Co-evolution and Self-organization in Dynamical Networks" and IST-1999-14186 ALCOM-FT "Algorithms and Complexity in Future Technologies", and by the Italian research project ALINWEB: "Algorithmica per Internet e per il Web", MIUR – Programmi di Ricerca di Rilevante Interesse Nazionale.

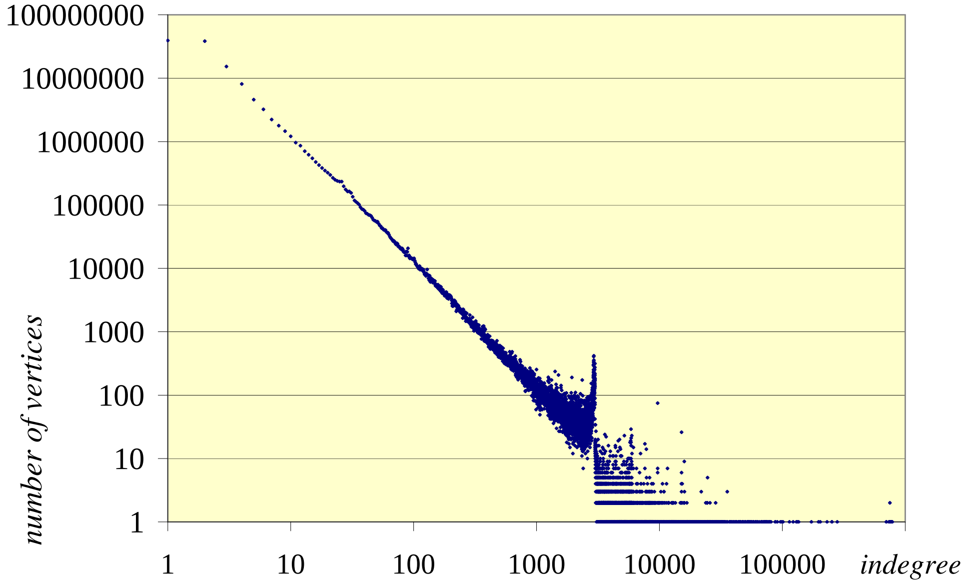


**Fig. 1.** In-degree distribution of the Web Base crawl.

recognized [10] on a sample of 200M pages crawled by Alexa in 1997.

The Copying model has later been proposed by Kumar et al. [9] in the attempt to model the formation in the Webgraph of a large number of bipartite cliques. The Copying model selects a random prototype vertex $p$ for every new vertex entering the graph. A constant number $d$ of links connect the new vertex to previously inserted vertices. The model is parameterized on a *copying factor* $\alpha$. The end-point of a link is either copied with probability $\alpha$ from a link of the prototype vertex $p$, or it is selected at random with probability $1 - \alpha$.

As pointed out at the beginning of this paper, the analysis of the linked structure of the Web is at the basis of important Web search algorithms such as the popular PageRank algorithm introduced by Brin and Page [3]. This algorithm has a simple interpretation in terms of a random walk in the Webgraph. Assume the walk has reached page $p$. The walk then continues either by following with probability $1 - c$ a random link in the current page, or by jumping with probability $c$ to a random page. The rank of every page is given by the probability that the random walk stops at that specific page.

The correlation between the distribution of PageRank and in-degree has been studied by Pandurangan, Raghavan and Upfal [15] motivated by the fact that PageRank is considered a much better strategy than simply ranking pages by indegree. They showed, by analyzing a sample of 100,000 pages of the brown.edu domain, that PageRank and in-degree are similarly distributed with a power law of exponent 2.1. However, it has been observed very little correlation between the two distributions, i.e., pages with high in-degree may well have low PageRank.

Recently Panduragan, Raghavan and Upfal [15] proposed a model that complements the Evolving Network model [1] by choosing the endpoint of a link with probability proportional to the in-degree and to the PageRank of a vertex. The authors show by computer simulation that with an appropriate fitting of the parameters the graphs generated capture the distributional properties of both PageRank and in-degree.

More generative models for the Webgraph have been presented in literature [16,11,15,5]. We refer to [14] for an excellent survey of models generating graphs holding power law distributions.

**Outline of the paper.** In this paper we report an extensive study of the statistical properties of the Webgraph by analyzing a crawl of about 200M pages collected in 2001 by the WebBase project at Stanford [17] and made available for our study. We briefly present our sample in Section 2, and the experimental findings on its structure are presented in the following sections.

More specifically, we report in Section 3 on the study of the in-degree and the out-degree, in Section 4 on the values computed by PageRank, and in Section 5 we present the study of the bipartite cores of our sample. All frequency distributions are plotted on a loglog scale. We conclude in section 6 with the study of the Bow-Tie structure and of the strongly connected components of the Webgraph.

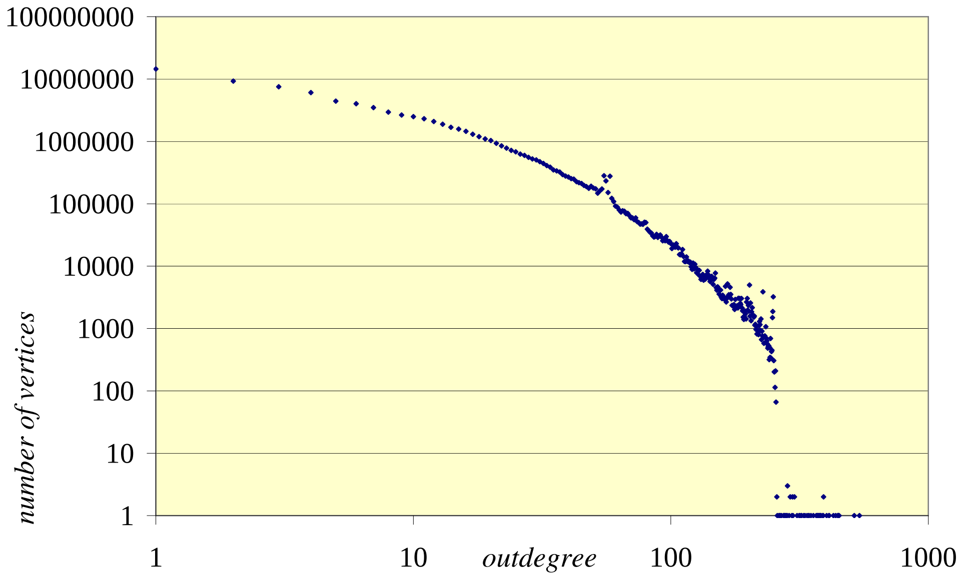


**Fig. 2.** Out-degree distribution of the Web Base crawl.

This work has required the development of several software tools for computing statistical and topological properties of very large graphs. A detailed description of the software tools developed within this project is in [12].

## 2 The WebBase crawl

We conducted our experiments on a $200M$ nodes crawl collected from the WebBase project at Stanford [17] in 2001. The repository makes several crawls available to researchers. The sample we study in our work contains only link information, i.e. no information about URLs is available. There are no recent estimates about the size of the web, but a study made by Cyvellance [6] showed that in July 2000 the web reached 2.1 billion webpages, and the number is growing 7 million pages each days. This means that the WebBase sample, when it was collected, contained about one tenth of the web.

## 3 In-degree and out-degree

We recall that the in-degree (out-degree) of a node is the number of entering (leaving) edges. For example, if we refer to the simple directed graph shown in Figure 3 the in-degree of vertex $C$ is 2 (it is linked from $A$ and $B$) while its out-degree is 1 (it links node $D$).

The in-degree distribution, shown in Figure 1, follows a power law with $\gamma = 2.1$. This confirms the observations done on the crawl of 1997 from Alexa [10], the crawl of 1999 from Altavista [4] and the notredame.edu domain [2].

We note a *bump* between the values 1.000 and $10,000$, that has also been observed by Broder *et al.* [4] and it is probably due to a huge clique created by a single *spammer*. Since our sample contains only structural information and not URLs, we can't propose or deny possible explanations for this phenomena.

In Figure 2 it is shown the out-degree distribution of the WebBase crawl. While the in-degree distribution is fitted with a power law, the out-degree is not, even for the final segment of the distribution. A deviation from a power law for the initial segment of the distribution was already observed in the Altavista crawl [4]. A possible explanation of this phenomena is that writing a scale-free series of hyperlinks is seriously limited by the patience of webmasters.

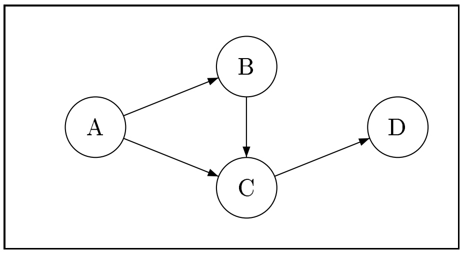


**Fig. 3.** A directed graph.

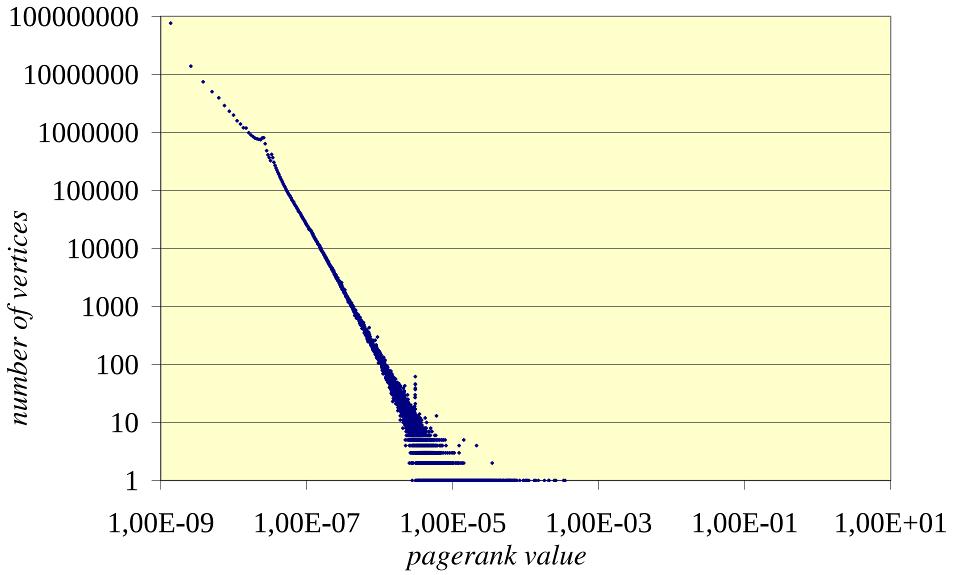


**Fig. 4.** PageRank distribution of the Web Base crawl.

## 4 PageRank

The *PageRank* algorithm is at the basis of the ranking operated by the Google Web search engine. The idea behind link analysis ranking is to give higher rank to documents pointed by many Web pages. Brin and Page [3] extend this idea further by observing that links from pages of high quality should confer more authority. It is not only important which pages point to a page, but also what is the quality of the pages. They propose a weight propagation algorithm in which a page of high quality is a page pointed by many pages of high quality.

The PageRank algorithm performs a random walk on the graph $G$ that simulates the behavior of a "random surfer". The surfer starts from some node chosen according to some distribution, usually the uniform distribution. At each step the surfer proceeds as follows: with probability $1-c$ an outgoing link is picked uniformly at random, and the surfer moves to a new page, and with probability $c$ the surfer jumps to a random page chosen accordingly to some distribution, usually the uniform distribution. The authority weight $Rank(i)$ of a node $i$ (called the page rank of node $i$) is the fraction of time that the surfer spends at node $i$.

More formally, the computation of PageRank is (expressed in matrix notation) as follows. Let $N$ be the number of vertices of the graph and let $n(j)$ be the out-degree of vertex $j$. Denote by $M$ the square, stochastic matrix whose entry $M_{ij}$ has value $1/n(j)$ if there is a link from vertex $j$ to vertex $i$. Denote by $[\frac{1}{N}]_{N\times N}$ the square matrix of size $N\times N$ with entries $\frac{1}{N}$. Vector $Rank$ stores the value of PageRank computed for the $N$ vertices. A matrix $M'$ is then derived by adding transition edges of probability $(1-c)/N$ between every pair of nodes to include the possibility of jumping to a random vertex of the graph:

$$M' = cM + (1-c)\times[\frac{1}{N}]_{N\times N}$$

A single iteration of the PageRank algorithm is

$$M'\times Rank = cM\times Rank + (1-c)\times[\frac{1}{N}]_{N\times 1}$$

The matrix $M'$ is the matrix of the Markov chain that corresponds to the random walk performed by the PageRank algorithm. The addition of the jump matrix guarantees that the Markov chain is irreducible and aperiodic, then there is an equilibrium steady state distribution for the states of the Markov chain. The PageRank is the stationary distribution of the Markov chain, that is the left eigenvector of the matrix $M'$.

We computed the PageRank distribution on the WebBase crawl. Here, we confirm the observation of [15] by showing this quantity distributed according to a power-law with exponent $\gamma = 2.109$. We also computed the statistical correlation between PageRank and in-degree. We obtained a value of $-5.1877E-6$, on a range of variation in $[-1,1]$ from negative to positive correlation. This confirms on much larger scale the observation done by [15] on the brown.edu domain of 100,000 pages, that the correlation between the two measures is very weak.

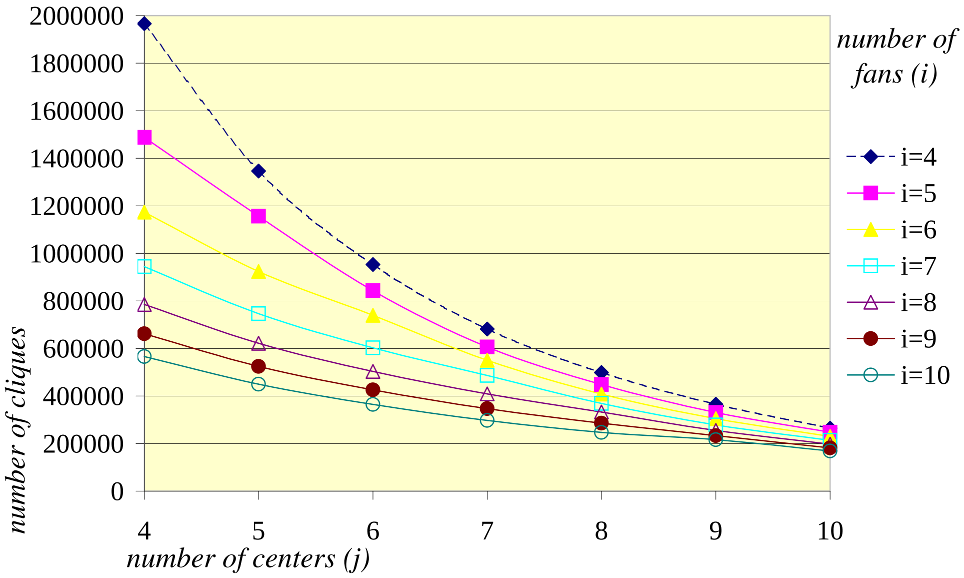


**Fig. 6.** Number of bipartite cores in the Web Base crawl.

## 5 Bipartite cliques

As we mentioned in the introduction, a large number of bipartite cliques in the web has been observed in [10]. A simple bipartite clique is shown in Figure 5: we have on the left side the set of the *fan* nodes (labelled $A$,$B$ and $C$), all of them pointing to all *center* nodes on the right side (labelled $X$,$Y$,$W$ and $Z$).

In Figure 6 the graphic of the distribution of the number of bipartite cliques $(i, j)$, with $i, j = 1, \ldots, 10$ is shown. The shape of the graphic follows that one presented by Kumar et al. [10] for the 200M crawl by Alexa. However, we detect a much larger number of bipartite cliques. For instance the number of cliques of size $(4, j)$ differs from the crawl from Alexa for more than one order of magnitude. A possible (and quite natural) explanation is that the number of *cyber-communities* has consistently increased from 1997 to 2001. We also recall that the longevity of *cyber-communities*' website is bigger as compared to other websites [10]. A second possible explanation is that our algorithm for finding disjoint bipartite cliques, which is explained in [13], is more efficient than the one implemented in [10]. We will try to get access to the Alexa sample [10] and execute on it our algorithm for disjoint bipartite cliques.

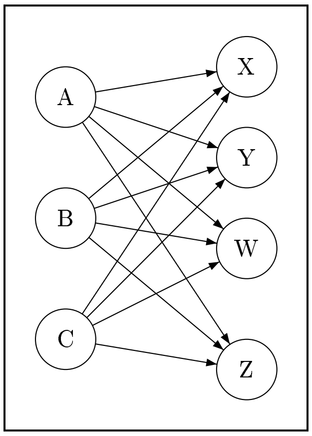


**Fig. 5.** A (3,4) bipartite clique

## 6 Strongly connected components

In a directed graph we say that a set of nodes $S$ is a *strongly connected component (scc)* if and only for every couple of nodes $A, B \in S$ there exists a directed path from $A$ and $B$ and from $B$ to $A$. The number of nodes of $S$ is the size of the scc. For example, in the graph shown in Figure 7, there are 3 distinct strongly connected components, respectively of size 4, 3 and 2.

Broder et al. [4] identified a very large strongly connected component of about 28% of the entire crawl, and shown a picture of the whole Web as divided in five distinct regions: SCC, IN, OUT, TENDRILS and DISCONNECTED. The SCC set is the set of all the nodes in the single large strongly connected component; in the IN (OUT) region we find all the nodes that can reach the

| | SCC | IN | OUT | TENDR. | DISC. |
|---|---|---|---|---|---|
| Altavista (1999) [4] | 28% | 21% | 21% | 22% | 9% |
| WebBase (2001) | 33% | 11% | 39% | 13% | 4% |

**Table 1.** Size of regions in both Altavista and WebBase crawl

SCC set (are reached from the SCC). TENDRILS are either nodes that leave the IN without entering the SCC or enter the OUT without leaving the SCC. In Table 1 we report the relative size of the 5 regions. We can still observe in the WebBase crawl a large SCC, however the biggest component is the OUT region, and both IN and TENDRILS have a reduced relative size if compared to the Altavista crawl. We also observe a huge difference between the size of the largest SCC of about 48 millions nodes, and the size of the second largest scc that is less than 10 thousands nodes.

In Figure 8 it is shown the scc distribution of the Webbase sample and of the different regions (of course the SCC region is a single scc). All distributions follow a power law whose exponent is 2.07, very close to the value observed for both the in-degree and the PageRank distribution.

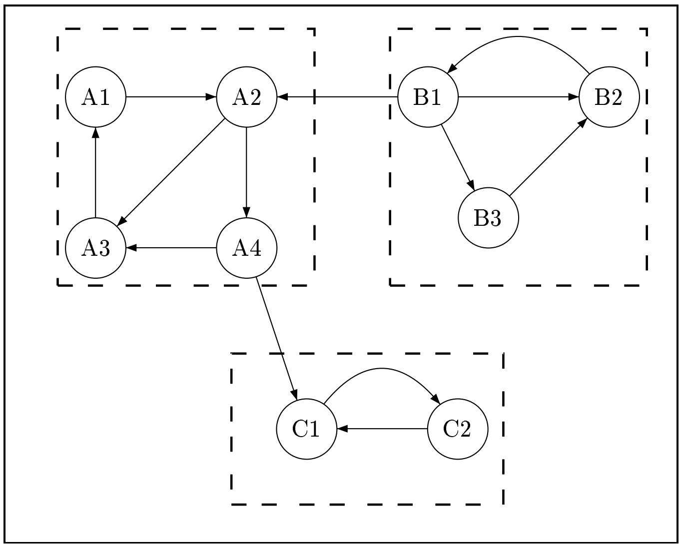


**Fig. 7.** An example of strongly connected components of a graph.

## 7 Conclusions

In this work we have presented an experimental analysis of the statistical and topological properties of a large sample of the Webgraph. We plan in the near future to compare these results with the ones obtained on more recent crawls of the Webgraph in order to assess the temporal evolution of its topological properties.

**Acknowledgements:** We are very thankful to the WebBase project at Stanford and in particular Gary Wesley for their great cooperation. We also thank James Abello, Guido Caldarelli, Paolo De Los Rios, Camil Demetrescu and Alessandro Vespignani for several helpful discussions.

## References

1. R. Albert, H. Jeong, and A.L. Barabasi. *Nature*, (401):130, 1999.
2. A.L. Barabasi and A. Albert. Emergence of scaling in random networks. *Science*, (286):509, 1999.
3. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engines. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
4. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, S. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, June 2000.
5. C. Cooper and A. Frieze. A general model of undirected web graphs. In Springer-Verlag, editor, *Proc. of the 9th Annual European Symposium on Algorithms(ESA)*, LNCS 2161, pages 500–511, 2001.
6. Cyvellance. `www.cyvellance.com`.
7. P. Erdös and Renyi R. *Publ. Math. Inst. Hung. Acad. Sci*, 5, 1960.
8. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1997.
9. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Random graph models for the web graph. In *Proc. of 41st FOCS*, pages 57–65, 2000.
10. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber communities. In *Proc. of the 8th WWW Conference*, pages 403–416, 1999.
11. L. Laura, S. Leonardi, G. Caldarelli, and P. De Los Rios. A multi-layer model for the webgraph. In *On-line proceedings of the* 2nd International Workshop on Web Dynamics., 2002.
12. L. Laura, S. Leonardi, and S. Millozzi. A software library for generating and measuring massive webgraphs. Technical Report 05-03, DIS - University of Rome La Sapienza, 2003.
13. L. Laura, S. Leonardi, S. Millozzi, U. Meyer, and J.F. Sibeyn. Algorithms and experiments for the webgraph. In Springer-Verlag, editor, *Proc. of the 11th Annual European Symposium on Algorithms (ESA)*, volume 2461 of *Lecture Notes in Computer Science*, 2002.
14. Michael Mitznmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2003.
15. G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. In *Proc. of the 8th Annual International Conference on Combinatorics and Computing (COCOON)*.
16. D.M. Pennock, G.W. Flake, S. Lawrence, E.J. Glover, and C.L. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proc. of the National Academy of Sciences*, 99(8):5207–5211, April 2002.
17. The stanford webbase project. `http://www-diglib.stanford.edu/~testbed/doc2/WebBase/`.

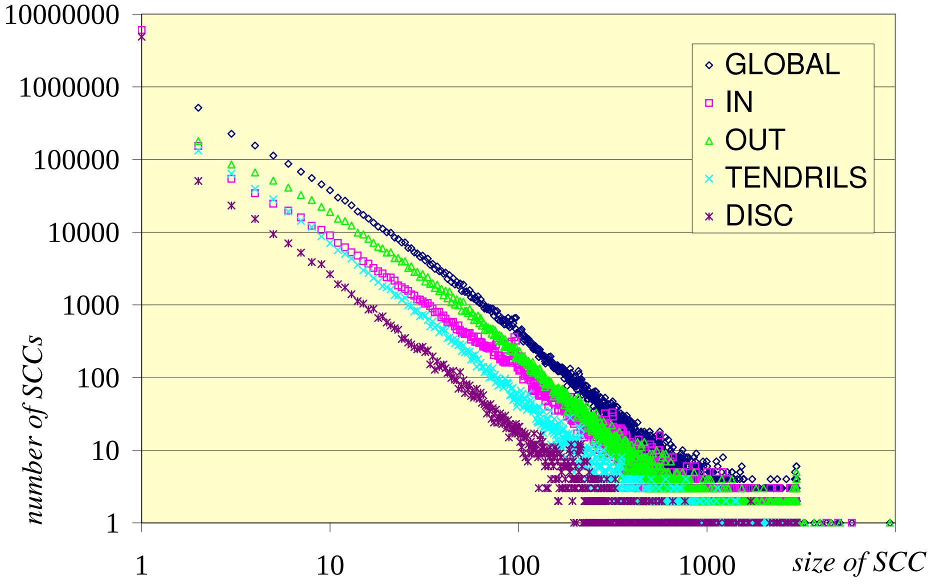


**Fig. 8.** SCC distribution of the Web Base crawl.