

A BEGINNER'S GUIDE TO THE WEBGRAPH: PROPERTIES, MODELS AND ALGORITHMS.

Debora Donato

Luigi Laura

Stefano Millozzi

Dipartimento di Informatica e Sistemistica
Università di Roma “La Sapienza”
{donato,laura,millozzi}@dis.uniroma1.it

When I took office, only high energy physicists had ever heard of what is called the Worldwide Web.... Now even my cat has its own page.

– BILL CLINTON, *announcement of Next Generation Internet initiative* (1996)

Abstract

How many mouse clicks separates your web page from Julia Roberts's one? Probably few. And how many pages have exactly all the links your page has? Probably more than you think. And, if only few pages point your one, will Google still be able to rank it high? Probably yes.

These findings derive from the study of the Webgraph, i.e. the graph whose nodes are the (static) html pages and whose (directed) edges are the hyperlinks among them. We discuss its main properties, the stochastic graph models aimed to capture them and the algorithmic challenges that such a huge structure poses.

1. Introduction

An extensive study of “the Web as a graph” appeared, in 1999, in the work of Kleinberg et al. [19]. Here the authors, for the first time, explicitly focused on the directed graph induced by the hyperlink structure of the World Wide Web and several previously appeared results, together with new ones, were listed in an homogeneous framework. From then on the term *Webgraph* addresses the graph whose nodes are the (static) html pages and edges are the hyperlinks among them.

A large amount of research recently studied the properties of the Webgraph by collecting and measuring samples spanning a good share of the whole Web. A second important research line has been the development of stochastic models generating graphs that capture the properties of the Web. A stochastic model of the Webgraph can be helpful for several reasons, including (i) testing of web applications against synthetic benchmarks, (ii) formally proving of properties of web algorithms and (iii) monitoring the real evolution of the Webgraph against the model's projections. These studies require the development algorithmic tools to deal with graph of several billion edges.

In the following section we recall useful definitions. In Section 3 we discuss the major properties and models of the Webgraph, and in Section 4 related algorithmic issues are considered. We conclude by raising some open problems in Section 5.

2. Preliminaries

We recall that, in a directed graph, the in-degree (out-degree) of a node is the number of incoming (outgoing) edges. For example, if we refer to the simple directed graph shown in Figure 1, 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).

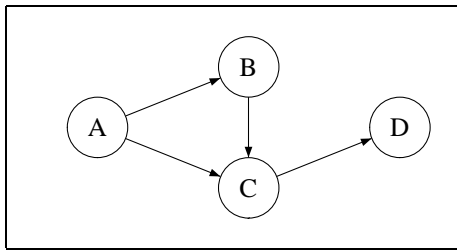


Figure 1: A directed graph.

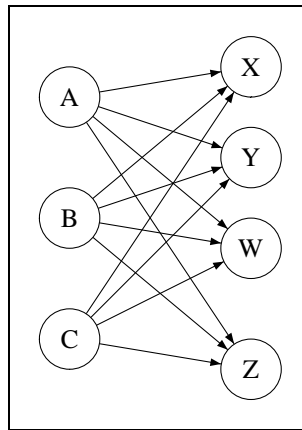


Figure 2: A (3,4) bipartite clique.

A *bipartite clique* is made of two sets of node; all the nodes in the first set (the *fan set*) point to each node of the second one (the *core set*). An example is shown in

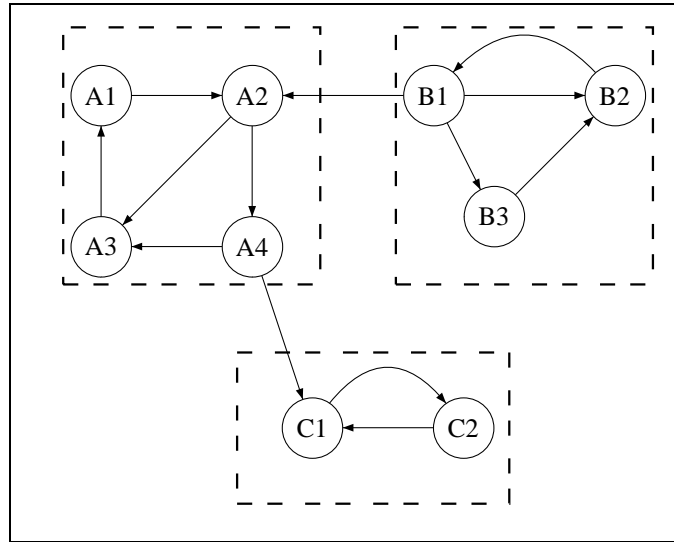


Figure 3: An example of strongly connected components of a graph.

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

3. Properties and Models of the Webgraph

Despite being the sum of a decentralized and uncoordinated effort of a huge number of heterogenous groups and individuals, the Webgraph exhibits a well defined structure, characterized by several properties. Kleinberg et al. [19] and Albert et al. [2] observed independently that, if we plot, on a double logarithmic scale, the number of nodes with a given in-degree against the in-degree values, we obtain a negative slope straight line:

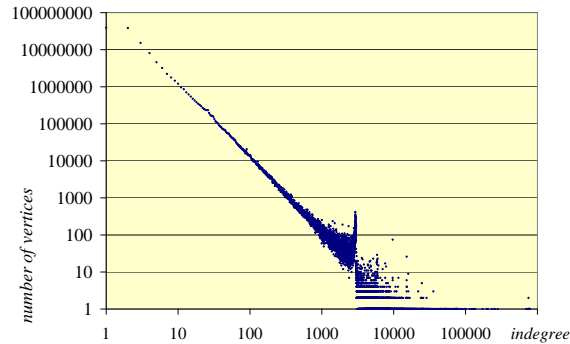


Figure 4: In-degree distribution of the WebBase [32] crawl.

this is an example of *power law*¹ distribution. It is interesting to notice that this finding came at the same time from exponents of two well distinct research communities, i.e. on one side computer scientists focusing on creating and improving web algorithms, and on the other physicians fascinated by the complexity and dynamical nature of the Web. In Figure 4 is shown the in-degree distribution of a sample of the Webgraph.

In order to capture the Webgraph evolving nature Albert, Barabasi and Jeong [2] presented the *Evolving Networks* model in which at every discrete time step a new vertex is inserted into the graph. The new vertex connects to a constant number of previously inserted vertices chosen according to the *preferential attachment* rule, i.e. with probability proportional to the in-degree. This model shows a power law distribution over the in-degree of the vertices with exponent roughly 2. This value has been measured (and formally proved) when the number of edges that connect every vertex to the graph is 7, that is equal to the average value observed in several samples of the Webgraph [19, 9, 13].

The evidence of a well defined structure of the Webgraph was emphasized by Broder et. al. [9] that presented a suggesting picture (shown² in Figure 6): a *bow-tie* shape with a core made by a large strongly connected component and four sets of vertices distinguishable from their relation to the core: upstream nodes, that can reach

¹By a *power law* in-degree distribution we mean that the percentage of web pages with in-degree d is proportional to $1/d^\alpha$ for some constant α and large enough d .

²Source: IBM Almadem Research Center website [4]

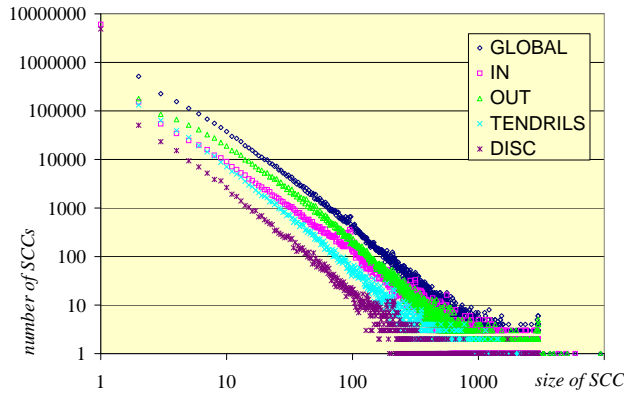


Figure 5: SCC distribution of the Web Base crawl.

it but not be reached from; viceversa, downstream set can reach the core but not be reached by it; the “tendrils” set is made up by those nodes that neither can reach nor be reached from the core; the last set is made of “islands”, i.e., small groups of vertices not connected to the bow-tie.

Broder et. al. [9] also estimated the probability of the existence of a path between a random source and a random destination. They found out that the Webgraph exhibits the *small world phenomenon*³ [31, 17] only if the hyperlinks are considered undirected: almost all pages of a giant connected component, including about 90% of the web documents, are reachable within few hops from every other page; this confirmed, at some extent, the observations made in [2] about the diameter of the Web.

In Figure 5 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.

A surprising number of specific topological structures such as bipartite cliques of relatively small size has been observed in [22]. The study of such structures is aimed to trace the emergence of hidden *cyber-communities*. Over 100,000 such communi-

³With *small world phenomenon* we mean that, despite the huge dimension of the graph, the diameter, i.e. the maximum distance between two connected nodes, is small.

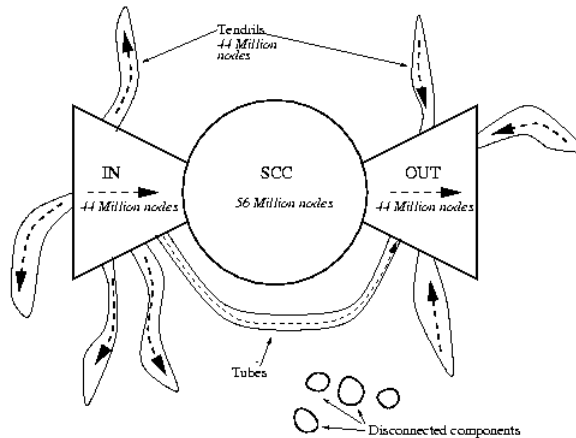


Figure 6: The *bow-tie* structure of the Webgraph.

ties have been recognized on a sample of 200M pages on a crawl from Alexa [3] of 1997 and, more recent estimates [13], based on a sample from WebBase [32] of 2001, indicate that the number grew up to more than 1M (see Figure 7).

To model the formation of such a large number of bipartite cliques Kumar et al. proposed the *Copying* model, parameterized on a *copying factor* α . Here, for every new vertex entering the graph a prototype vertex p is selected at random. A constant number d of links connect the new vertex to previously inserted vertices. The end-point of a link is either copied with probability α from a link of the prototype vertex p , or it is selected at random with probability $1 - \alpha$. The model has been analytically studied and shown to hold power law distributions on both the in-degree and the number of disjoint bipartite cliques for specific values of α . In particular, the in-degree is distributed with a power law with exponent 2.1 when $\alpha = 0.8$.

The Google search engine is based on the popular PageRank algorithm first introduced by Brin and Page [8]. 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, with probability c the surfer jumps to a random page chosen accordingly to some distribution. The

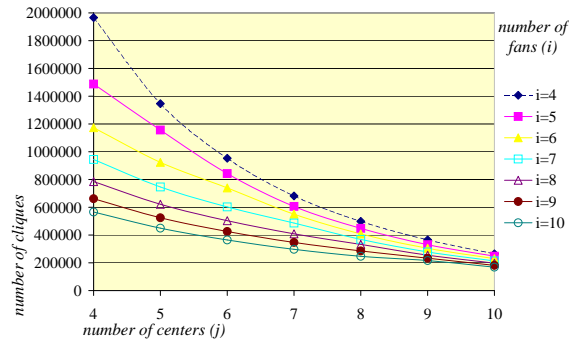


Figure 7: Number of bipartite cliques in the WebBase crawl.

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 .

The correlation between the distribution of PageRank and in-degree has been studied in a work of Pandurangan, Raghavan and Upfal [25]. They showed by analyzing a sample of 100,000 pages of the brown.edu domain that PageRank is distributed with a power law of exponent 2.1. This exactly matches the in-degree distribution, but very surprisingly it is observed very little correlation between these quantities, i.e., pages with high in-degree may have low PageRank. This property was confirmed on much larger scale by Donato et al. [13] on the WebBase crawl, where the statistical correlation value between PageRank and in-degree distribution is $-5.1877E - 6$, on a range of variation in $[-1, 1]$ from negative to positive correlation.

Based on the above observations, Pandurangan et al. [25] proposed a model that complements the Evolving Network model by choosing the endpoint of a link with probability proportional to the in-degree and to the PageRank of a vertex. They also showed, by computer simulation, that with an appropriate fitting of the parameters the graphs generated capture the distributional properties of both PageRank and in-degree.

In a more recent paper Dill et al. [12] explain how the web shows a fractal structure in many different ways. The Webgraph can be viewed as the outcome of a number of similar and independent stochastic processes. At various scales we have that there are “cohesive collections” of web pages (for example pages on a site, or pages about a topic) and these collections are structurally similar to the whole Web. The central

regions of such collections are called “Thematically Unified Clusters” (TUCs) and they provide a navigational backbone of the Web. Pennock et al. [26] argue that the Web is the sum of stochastic independent processes that share a common (fractal) structure but this structure sometimes can be really different. Indeed they provide examples where the distributions exhibit large deviation from power laws.

Motivated by the above works, Laura et al. [23] propose a *Multi-layer* model of the Webgraph in which every new page that enters into the graph (i) is assigned to a number of regions (layers) it belongs to and (ii) it is allowed to link only to vertices of those regions. Inside each region, the links are chosen according to a combination of the EN [2] and Copying [21] model. This model showed some nice properties such as a distribution of the in-degree with a power law of exponent 2.1 for a wide range of variation of the parameters.

4. Algorithms for the Webgraph

In this section we present an overview of the algorithmic tools related to the study of the webgraph⁴. Despite the fact that in the seminal paper of Kleinberg et al. [19] all the measures have been done using a computer with 12 GB of ram, therefore allowing all the computations to be held in main memory, the size of the Webgraph requires to explicitly deal with massive graphs stored in secondary (slow) memory. A survey of algorithms for (general) graph stored in secondary memory can be found in the work of Chiang et al. [10]; we distinguish between *semi-external* algorithms, that use a small constant amount of memory for each node of the graph, therefore limiting the size of the input, and the “pure” *external* ones, that impose no constraints.

It is important to point out that, in the context of external memory algorithms, problems considered easy in main memory can become tricky; the typical example is the visit of a graph (either breath first or depth first search) that is very easy to be performed in main memory and, usually, due to the lack of locality, can be unfeasible in secondary memory. Indeed, so far there are no worst-case efficient external-memory algorithms to compute DFS trees for general directed graphs. A semi-external algorithm for DFS was developed recently by Sibeyn et al. [29]. It maintains a tentative forest, which is modified by I/O-efficiently scanning of non-tree edges, so as to reduce the number of cross edges (in [24] a sample of the Webgraph has been analyzed with this algorithm).

It is a well-known fact that SCCs can be computed in linear time by two rounds of DFS. Therefore, with the above technique, it is possible to afford this computation in

⁴Note that in a broad sense all the algorithmic issues related with Web Search Engines can be included in this topic. Here we mention only algorithms focusing on the “Web as a Graph”.

secondary memory. However, experimental results seen in [24] show that this computation, among the other measures described in the previous section, remains one of the most difficult. A different approach, partially inspired by the work of Fleischer et al. [15], takes advantage of the bow-tie structure of the Webgraph: instead of computing all the SCCs, the algorithm first looks for the big one (CORE), then it detaches it from the graph and continues the computation over the small remaining ones (see [14]).

In [22] an algorithm for enumerating disjoint bipartite cliques (i, j) of size at most 10 has been presented, with i being the fan vertices on the left side and j being the center vertices on the right side. The algorithm proposed by Kumar et al. is composed of a pruning phase that consistently reduces the size of the graph in order to store it in main memory. A second phase enumerates all bipartite cliques of the graph. A final phase selects a set of bipartite cliques. Every time a new clique is selected, all intersecting cliques are discarded. Two cliques are intersecting if they have a common fan or a common center. A vertex can then appear as a fan in a first clique and as a center in a second clique. A different technique has been presented in [24].

Among the link analysis algorithms, that are devoted to rank web pages, we cite Kleinberg's HITS [18] and Brin and Page's Pagerank [8], that we mentioned in the previous section. In both cases, the idea of authoritative page is "one that is pointed by authoritative pages" and therefore the weights are computed in an iterative manner. An efficient implementation of Pagerank has been proposed in [16].

All the above techniques deal with graphs represented as list of edges. A different approach is to compress the graph so it can fit into main memory where it can be processed [1, 5, 28, 6, 7, 27, 30]. Among these ones we cite the results achieved in the work of Boldi and Vigna [6]. Here the authors, exploiting some observed properties of web addresses such as locality, similarity and consecutivity, are able to compress a web graph at the rate of 3.08 bits per link.

A software library that implements some of the above algorithms, as well as algorithms able to generate webgraphs according to some of the models discussed in the previous section, is presented in [14] and is freely available from the COSIN website [11].

5. Open problems

Despite its dynamic nature, the Webgraph has been studied so far from a static point of view: snapshots of it have been analyzed but it is still missing the projection of its properties against a temporal axis. A first step towards this direction has been presented in [20], where each edge is labelled with the dates of its first and last appearance in the web. This new data, of course, poses several challenges; among them we cite (i) the problem of efficiently representing dynamic graphs in secondary memory, (ii)

whether it is possible to adapt the web graph compression techniques to it and (iii) if it is possible to design algorithms able to deal explicitly with the time labels without the need of generating multiple snapshots from it.

Acknowledgements

The authors thank Luciana S. Buriol, Antonella Poggi, Alessandro Termini and Andrea Vitaletti.

This work has been partially supported by the “Progetto ALINWEB: Algoritmica per Internet e per il Web” (MIUR Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale) and the FET Open Project IST-2001-33555 (COSIN [11]).

References

- [1] M Adler and M Mitzenmacher. Towards compressing web graphs. Technical Report 00-39, U.of Mass.
- [2] R. Albert, H. Jeong, and A.L. Barabasi. *Nature*, (401):130, 1999.
- [3] Alexa, inc. <http://www.alexa.com>.
- [4] IBM Almadem Research Center. <http://www.almadem.ibm.com>.
- [5] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: fast access to linkage information on the web. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 469–477. Elsevier Science Publishers B. V., 1998.
- [6] P. Boldi and S. Vigna. The webgraph framework i: Compression techniques. In *Proceedings of the 7th WWW conference*, 2003.
- [7] P. Boldi and S. Vigna. The webgraph framework ii: Codes for the worldwide web. Technical Report 294-03, Universit di Milano, Dipartimento di Scienze dell’Informazione, 2003.
- [8] 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.
- [9] 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.
- [10] Y. Chiang, M. T. Goodrich, E. F. Grove, R. Tamassia, D. E. Vengroff, and J. S. Vitter. External-memory graph algorithms. In *Symposium on Discrete Algorithms*, pages 139–149, 1995.
- [11] European Project COSIN - COevolution and Self-Organisation In dynamical Networks. <http://www.cosin.org>.
- [12] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. In *Proceedings of the 27th VLDB Conference*, 2001.

- [13] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *European Journal of Physics B*, 2004. DOI: 10.1140/epjb/e2004-00056-6.
- [14] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. A software library for generating and measuring massive webgraphs. Technical Report D13, COSIN European Research Project, 2004.
- [15] L. Fleischer, B. Hendrickson, and A. Pinar. On identifying strongly connected components in parallel. In *Proceedings of IRREGULAR 2000*, number 1800 in LNCS, pages 505–511, 2000.
- [16] Taher H. Haveliwala. Efficient computation of pagerank. Technical report, Stanford University, 1999.
- [17] J. Kleinberg. The small world phenomenon: an algorithmic perspective.
- [18] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1997.
- [19] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models and methods. In *Proc. Intl. Conf. on Combinatorics and Computing*, pages 1–18, 1999.
- [20] R. Kraft, E. Hastor, and R. Stata. Timelinks: Exploring the link structure of the evolving web. In *Second Workshop on Algorithms and Models for the Web-Graph (WAW2003)*.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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)*.
- [26] 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.
- [27] S. Raghavan and H. Garcia-Molina. Representing web graphs. In *Proceedings of the 19th International Conference on Data Engineering*, 2003.

- [28] K. Randall, R. Stata, R. Wickremesinghe, and J. Wiener. The link database: Fast access to graphs of the web, 2001.
- [29] J.F. Sibeyn, J. Abello, and U. Meyer. Heuristics for semi-external depth first search on directed graphs. In *Proceedings of the fourteenth annual ACM symposium on Parallel algorithms and architectures*, 2002.
- [30] T. Suel and J. Yuan. Compressing the graph structure of the web. In *Data Compression Conference*, pages 213–222, 2001.
- [31] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, (393):440, 1998.
- [32] The stanford webbase project. <http://www-diglib.stanford.edu/~testbed/doc2/WebBase/>.