# *The Web Graph: an Overview*

Jean-Loup Guillaume and Matthieu Latapy

*LIAFA - Université Paris 7, 2, place Jussieu, 75005 Paris, France.* guillaume@liafa.jussieu.fr
*Hipercom - INRIA Rocquencourt, F-78153 Le Chesnay.* Matthieu.Latapy@inria.fr

We propose here an overview of the questions which arise about the Web graph, the methods used to (try to) answer them, and the kind of answers usually obtained. We first discuss some general facts about the Web graph: its size, its evolution rate, and other problems which make it hard to study. After this, we give an overview of the different statistics known about the Web graph, we give a survey of what is known about its structure at microscopic and macroscopic levels, and finally we describe the main statistical models introduced to study it.

**Keywords:** Web graph, World Wide Web, Internet, Random graphs, Dynamic graphs.

## 1 Preliminaries

The Web graph is the graph of the Web pages together with the hypertext links between them. Therefore, each vertex is an URL (Unique Ressource Locator), and the outgoing edges of a vertex are the hypertext links contained in the corresponding page. This graph is a key element for many present and future applications, like Web indexing, community detection, enhanced browsing, etc.

The Web graph is very large: in July 2000, it was estimated to contain about 2.1 billions vertices [MM00, LW01] and 15 billions edges. Moreover, about 7.3 millions pages are added every day, and many others are modified or removed. This study shows that the Web graph should contain about four billions vertices in the early 2001. With such an exponential increase, Web graph might currently (february 2002) contain more than six billions vertices (a lower bound is given by Google search engine which roughly uses two billion pages). Therefore, the two first problems which arise when one wants to study the Web graph are its size and its rapid evolution. Even if it was possible to get the whole graph, its storage and its algorithmic manipulation would be a real challenge [BBH+98, GLV02, RSWW01, SY01, WSW00].

The first step to study the Web graph is to get large parts of it. Because of its size, and of many other factors we will discuss below, it is *impossible* to get the whole graph, but it is a key challenge to be able to obtain parts of it as large as possible. The method used to compute such parts is called *crawl* and is performed by softwares refered as crawlers, spiders, robots... [HN99, LM01, Pag]. Basically, it can be viewed as a breadth-first search in a graph: one starts with a given set of initial pages, follows all the outgoing links, and iterates the process from the newly discovered pages. However, the real procedure is much more intricate: many technical factors make it impossible to process a real breadth-first search. For example, the frequency of the requests to a given server is limited, in order to avoid its overload. Therefore, any server which contains many pages will never be entirely crawled. These limitations make it hard to crawl more then three hundred millions pages in less than a month. Moreover, the rapid evolution of the graph implies that it changes *during* the crawl. Therefore, the obtained part is not really a part of the whole graph at a given time; it may be seen as a radar scan.

Because of all these considerations, the studies about the Web graph always deal with *parts* of the whole graph, generally from several millions to several hundreds of millions vertices, which are supposed to be representative. Moreover, one must keep in mind that these parts are biaised by the crawl method used, and the influence of this on the mesures is not clearly understood until now.

## 2 Statistical studies

The first results published concerning the Web graph are statistical studies made on large parts of the whole graph. Many static or dynamic parameters have been studied. Their distributions generally follow power laws: the probability of the parameter to have value $i$ is proportional to $i^{-\beta}$ for a given positive real $\beta$. This means in particular that many pages have a low value for the parameter, and that very few have a high value.

The main statistical studies deal with the incoming and outgoing degrees of the vertices [BKM+00, KKR+99, KRRT99, AJB99]. In other words, how many pages have a link to a given page, and how many links does a page contain? These two values follow power laws with exponent $\beta = 2.7$ for the outgoing degree, and $\beta = 2.1$ for the incoming degree, see Figure 1. Therefore, for example, the number of pages containing nine links is five times lower than the number of pages containing one link. Moreover, the average number of links in a page is 7. Some other studies show that the distribution of the sizes of the strongly connected components in the graph also follows a power law, with $\beta = 2.54$ [BKM+00]. Indeed, it was shown that the biggest strongly connected component of a 200 millions vertices crawl of the Web, is of size 56 millions nodes, while the second one is of size 150 thousands. We will discuss this in greater detail in the next section.

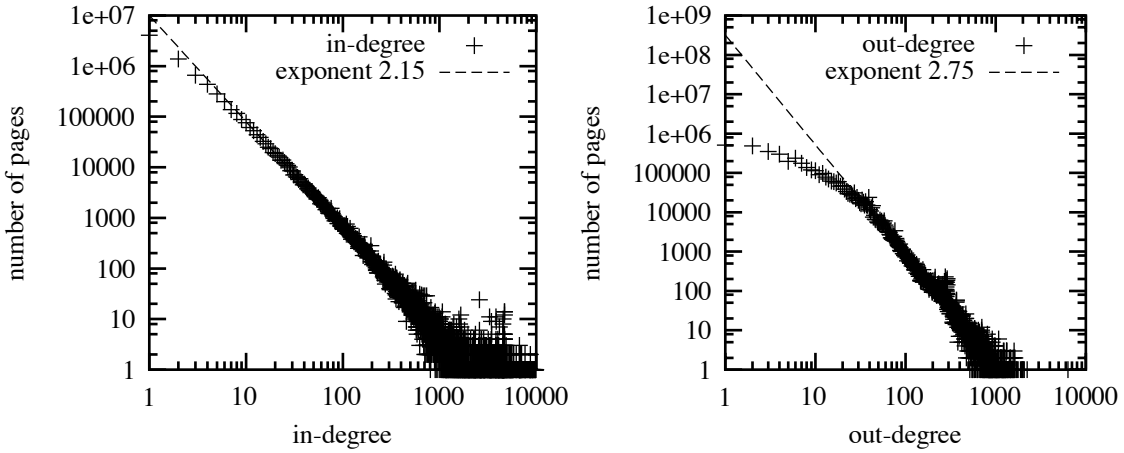


**Fig. 1:** Typical In and Out-degree distribution, obtained on a 27 millions pages crawl

More recently, it appeared that the dynamics of the Web graph plays an important role in the applications. For instance, a search engine must keep an up-to-date vision of Web pages, revisiting them as soon as they change. The hard trick is then to detect or to estimate when a page is going to change. Therefore, some studies on dynamical parameters appeared [BC00a, BC00b, CGM00, DFKM97]. For example, it is shown in [CGM00] that half of the Web is replaced every 50 days. Moreover, some parts of the Web evolve faster than others: 40% of the `.com` domain changed each day while less than 10% of the pages in other domains changed during this time.

## 3 Structure

A natural question which arise when one wants to study the Web graph concerns its structure at a higher level than the statistics: does the graph exhibit some local substructures, does it have a global structure, etc.? These questions can be separated into two parts: the microscopic structure of the Web graph on the one hand, and its macroscopic structure in the other hand.

The local structures which can be found in the Web graph are of great interest for community detection (and thus for search engines) [CDK+99, GKR98, KL01, Kle99, KRRT99]. For example, one can observe

that there are many couples of sets of Web pages such that each page of the first set has a link to all the pages of the second set, while there is no link between pages of the second set. This kind of structure is called a *bipartite clique*. It corresponds to a community centered on a given topic of interest: the pages in the first set are pages of *fans* while page in the second one are pages of *stars*. See Figure 2. Another definition has been recently introduced in which a community is a set of pages which have more links inside the set than outside [FLG00]. Some other kinds of local structures have been discovered in the Web graph, and studied from this point of view [ERC+00].

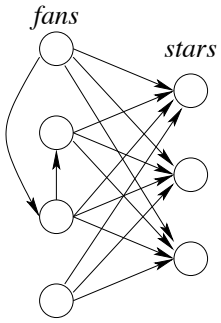


**Fig. 2:** *A structure which often appears in the Web graph, and which is interpretated as a set of* fans *having links to their* stars.

At a macroscopic level, it had long been thought that the Web was a large set of pages highly connected together. In other words, the Web graph would be mainly composed of a large strongly connected component, containing most of the pages. On the countrary, it is shown in the famous study [BKM+00] that the Web graph would have a bow-tie structure: in this paper, a study is made on a 200 millions vertices graph obtained from a crawl of the Web, and it appears that is is composed of four parts of equivalent sizes. See Figure 3. The first part is the largest strongly connected component of the graph (the second largest is much smaller), which composes the *core* of the well connected pages. The second part, called *IN*, is composed of those pages from which the core is reachable, but which are not reachable from the core. Conversly, the third part, called *OUT*, is the set of pages reachable from the core but from which the core is unreachable. Finally, the *dendrites* are the pages reachable from one of the three first parts, or from which one of the three first parts is reachable, but which belong to none of the previous parts. Only ten percent of the whole graph do not belong to one of these four parts which compose the bow-tie.

A recent study [DKM+01] shows that the Web has scale-free properties in the sense that it contains a lot of such bow-ties, when one considers for instance the sub-graphs induced by the content of the pages (all the Web pages dealing with a particular subject), by the localization (on a same server), or by the pages in a given geographic region

# 4 Models of the Web graph

Following classical approches used in statistical physics, some researchers tried to introduce some models which exhibit behaviours similar to the ones mesured on the Web graph. Two families of models have been introduced. The first family contains models whose aim is to explain the power law for degree distribution, while the second one is more based on structural properties of the Web graph.

It was first thought that the Web graph was similar to a random graph with a given distribution of the degrees (the one given above) [AJB99]. This model was used in particular to show that there exists a short path (of length 19 on average) between every pair of pages. However, bow-tie theory claims that there exists a path between two randomly chosen pages only once out of four. Another model was then introduced, based on an idea known as *preferential attachment* [ABJ00, BA99]: the new pages have most of their links towards some pages which already have many incoming links. This model generates non oriented graphs with a power-law distribution for degrees having an exponent of 3. These problems have been considered in a more complex model [DMS00] which also adds oriented links between pre-existing

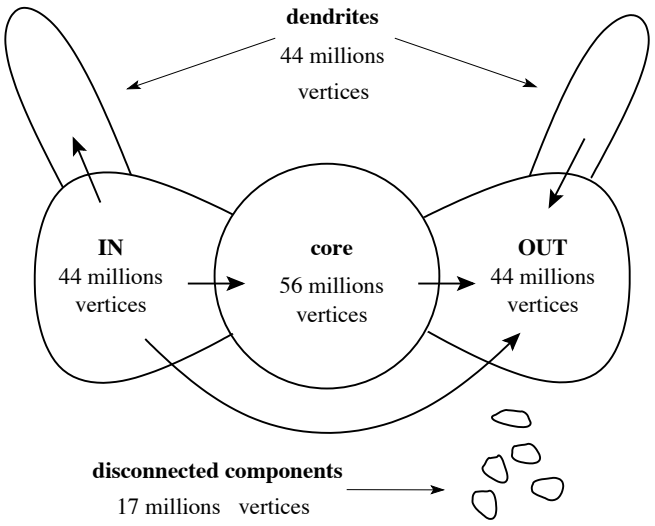


**Fig. 3:** *The bow-tie macroscopic structure of the Web graph [BKM+00]: the core, the IN component, the OUT component and the dendrites. Each of these parts contains around one quarter of the pages, the disconnected part being reduced to less than 10% of the whole.*

vertices at each time step. A last problem adressed by this family of models is to explain the deviation from the power law that appears at small numbers of edges (see for instance the out-degree distribution in Figure 1). Another model [PGF+00] adds links for a newly created vertex with an adjustable mixture of preferential attachment and uniform distribution. This model explains both the power law and the deviation for small values.

All previous models failed to explain the large amount of bipartite cliques observed in the Web graph. Therefore a new model has been introduced which explains this aspect [KKR+99, KRR+00]. In this model, every newly added vertex chooses a prototype vertex and each of its outlinks is choosen whether uniformly at random or by copying corresponding prototype outlink. An exponential variant of this model adds more than one vertex at a time and each of these vertices can only link to older vertices (this captures the fact that a page creator is not aware of newly created pages until these pages are referenced). Both variants of this model generate graphs whose in-degree distribution follows a power law , and which contain a large amount of bipartite cliques.

# 5 Perspectives

As one may notice from this short overview of the studies concerning the Web graph, many important questions still have no answers. For example, the influence of the crawl technique on the studies is not clear, and it may be important. It would for example be interesting to introduce some models of crawlers in order to verify that the bow-tie structure of the Web graph is not an artefact. This could for example be done by crawling graphs obtained using the statistical models introduced in the context of Web graph studies. Moreover, very few results are known about the community structures of the Web. Knowing the communities on the Web and the relations they share would give deep insight in the behaviour of the Web and of its users, which would certainly help to improve searching, crawling and browsing.

Most of the studies realized until now consider the graph as a static object, and work on a picture of it. However, it is now clear that many important properties and applications deeply depend on its dynamics. It is therefore important to put some efforts in the study of this dynamics. For example, what is the structure of the difference between two crawls, computed at some interval in time ? Can it be used to detect communities, to improve search engines, or to study the impact of social events on the Web ? Is an event like the *olympic games* detectable on the evolution of the Web graph ?

## References


[ABJ00] R. Albert, A. Barbasi, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web, 2000.

[AJB99] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[BA99] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[BBH+98] Krishna Bharat, Andrei Broder, Monika Henzinger, Puneet Kumar, and Suresh Venkatasubramanian. The Connectivity Server: fast access to linkage information on the Web. *Computer Networks and ISDN Systems*, 30(1–7):469–477, 1998.

[BC00a] B. Brewington and G. Cybenko. How dynamic is the web. In *Proceedings of the Ninth International World Wide Web Conference*, May 2000.

[BC00b] Brian E. Brewington and George Cybenko. Keeping up with the changing web. *IEEE Computer*, 33(5):52–58, 2000.

[BKM+00] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.

[CDK+99] S. Chakrabarti, B. E. Dom, S. Ravi Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[CGM00] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *26th International Conference on Very Large Data Bases*, September 2000.

[DFKM97] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: A live study of the world wide web. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, December 1997.

[DKM+01] Stephen Dill, Ravi Kumar, Kevin McCurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. Self-similarity in the web. VLDB 2001, 2001.

[DMS00] S. Dorogovtsev, J. Mendes, and A. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett. 85*, pages 4633–4636, 2000.

[ERC+00] Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, and Seyda Ertekin. The shape of the web and its implications for searching the web, 2000.

[FLG00] Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.

[GKR98] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[GLV02] Jean-Loup Guillaume, Matthieu Latapy, and Laurent Viennot. Efficient and simple encodings for the web graph. In *Proceedings of the 11-th international conference on the World Wide Web, 2002.*, 2002.

[HN99] Allan Heydon and Marc Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

[KKR+99] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: Measurements, models, and methods. In T. Asano, H. Imai, D. T. Lee, S. Nakano, and T. Tokuyama, editors, *Proc. 5th Annual Int. Conf. Computing and Combinatorics, COCOON*, number 1627. Springer-Verlag, 1999.

[KL01] Jon Kleinberg and Steve Lawrence. The structure of the web. *Science*, 294:1849–1850, november 2001.

[Kle99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. Proceedings of the 41th IEEE Symp. on Foundations of Computer Science, 2000.

[KRRT99] S. Ravi Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31(11-16):1481–1493, 1999.

[LM01] Sbastien Ailleret Pierangelo Veltri Laurent Mignet, Vincent Aguilera. Xyro : The xyleme robot architecture. First DIWeb Workshop, 2001.

[LW01] Mark Levene and Richard Wheeldon. Web dynamics. In *Software Focus, 2*, pages 31–38, 2001.

[MM00] B. Murray and A. Moore. Sizing the internet. White paper, Cyveillance, 2000.

[Pag] The Web Robots Pages. http://www.robotstxt.org/wc/robots.html.

[PGF+00] David M. Pennock, C. Lee Giles, Gary W. Flake, Steve Lawrence, and Eric Glover. Winners don't take all: A model of web link accumulation. Technical Report 2000-164, 2000.

[RSWW01] Keith Randall, Raymie Stata, Rajiv Wickremesinghe, and Janet L. Wiener. The link database: Fast access to graphs of the web. Technical Report 175, Compaq Systems Research Center, 130 Lytton Avenue - Palo Alto, CA 94301, november 2001.

[SY01] T. Suel and J. Yuan. Compressing the graph structure of the web. In *Proceedings of the IEEE Data Compression Conference (DCC)*, march 2001.

[WSW00] Rajiv Wickremesinghe, Raymie Stata, and Janet Wiener. Link compression in the connectivity server. Technical report, Compaq systems research center, 2000.