AN INTRODUCTION TO THE WORLD WIDE WEB

Debora Donato
Yahoo! Labs, Sunnyvale, California

Keywords: WWW, Web Graph, Power law properties, microscopic structure, macroscopic structure.

Contents

Summary

The Web is one of the most complex artifacts ever conceived by human beings. Its considerable dimensions are due to tools that have minimized the users' effort for content publishing. This characteristic is at the basis of a social phenomenon that has engendered the progressive migration of a good portion of the human activities toward the Web. The Web is the location where people meet, chat, get information, share opinions, transact, work or have fun.

The key feature that has attracted the interest of scientists with different background is that, despite being the sum of decentralized and uncoordinated efforts by heterogeneous groups and individuals, the World Wide Web exhibits a well defined structure, characterized by several interesting properties. This structure was clearly revealed by the study of Broder et al.(2000) who presented the evocative bow-tie picture of the Web. Although, the bow-tie structure is a relatively clear abstraction of the macroscopic picture of the Web, it is quite uninformative with respect to the finer details of the Web graph. In this chapter we present an overview of the main characteristics of the graph induced by his hyperlinked structure. We present a series of measurements on the Web, which offer a better understanding of the Web Graph both at microscopic and macroscopic level.

1. Introduction

The birth of the World Wide Web dates back to March 1989, when its father Tim Berners-Lee presented to the European Organization for Nuclear Research (CERN) a proposal for "a large hypertext database with typed links" [Berners-Lee T. (1989)]. Only at the end of 1990 Berners-Lee and Robert Cailliau implemented the system from then on known as the *World Wide Web* [Berners-Lee T. (2000)].

Since the day in which the first web site was published at the CERN, the Web has been evolving at an incredible rapid pace evaluated, in 2005, to be in the order of seven million new pages a day [Gullí A.(2005)]. The size of the Web is currently estimated to be around 22 billion pages (*http://www.worldwidewebsize.com*). The Web is by far the most used application in the world: in few decade it has deeply revolutionize human behavior impacting on the way people study, communicate, make business, deliver services, have fun.

How can we describe this object able to dramatically influence the core institutions of our daily life such as academia, industries and governments? In practice the Web is a distributed and completely decentralized *information media* comprised by million of computers spread over the world. Given its inherent multidisciplinary the Web has been capturing the interest of scientists coming from very different academic disciplines or professional specializations: physicists, mathematicians, computer scientists but also sociologists, psychologists, biologists and economists. Despite the plethora of research articles and books devoted to characterize, model and improving the Web, it remains largely unstudied [Hendler J. (2008)]. The need to anticipate future developments and encourage multidisciplinary collaborative research has recently motivated the Web Science Research Initiative (WSRI) which aims to the creation of the new discipline of Web Science *(http://www.webscience.org*).

In this chapter we give an overview of the principal characteristics of the World Wide Web, highlighting the main experimental and theoretical findings derived from the study of the Web graph, i.e. the graph whose nodes are the (static) html pages and whose (directed) edges are the hyperlinks among them. A graph is an abstraction for describing complex objects in which single items, in this case the pages, are related together, in this case using hyper-textual links. The Web graph is the starting point of the large amount of research activity that has recently been focused on studying the properties of the Web. The reason of such large interest is primarily given to search engine technologies. Remarkable examples are the algorithms for ranking pages such as PageRank [Brin S. (1998)] and HITS [Kleinberg J.(1997)] whose main goal is to sort the set of results on the basis of their relevance or closeness to the information need of the user.

Studying the properties of the Web is an ambitious task. The initial step to unearth the topological structure of the Web is to collect the hyperlinked structure of large crawls spanning geographical (i.e.,.uk,.es,.pt,.it and so on) or academic sub-domains (i.e.,.edu). These crawls are hardly comprised by more than few hundred million nodes (see Table 1) against estimates of dozens of billions nodes [Gullí A. (2005)].

This approach assumes that large samples of the Web will faithfully reproduce the properties of the whole graph and that it is possible to drawn, from their study, relevant conclusions on such properties. This assumption is sustained by the important results of Dill et al. (2001) who considered *thematically unified clusters* (TUCs), that is, sets of pages that are brought together due to some common trait. Dill showed that TUCs exhibit, at a small scale, the same properties of the whole Web. For such a reason the Web is said to be self-similar and it can be viewed as the outcome of a number of similar and independent stochastic processes. Another partial confirmation of the soundness of this approach comes from the work of Becchetti et al. (2006). This work was meant to study the characteristics of different sampling techniques, i.e., procedures for selecting a subset of items or nodes from an initial set. The authors compare many techniques included the *Breath First Search* (BFS) which is the strategy used by crawlers to explore the Web. A more detailed definition of this algorithm is presented in section 2. The comparison was done considering two different classes of properties: (i) *microscopic properties,* which characterize each single node as indegree and outdegree; (ii) *macroscopic properties,* which explore the connectivity of large and dense components. Becchetti et al.(2006) showed that the BFS allows to preserve many of the microscopic characteristics of the entire Web. The macroscopic properties are still captured by BFS provided that the initial seed is sufficiently large.

Even if the crawls used are considerably smaller than the entire Web, large-scale Web characterization studies poses several algorithmic challenges since the huge dimensions of the data sets that can reach several GB. A possible solution is to exploit modern compression techniques [Boldi P. (2004)] that provide simple ways to manage very large graphs. An alternative approach consists of implementing algorithms in semi-external or fully-external memory [Sibeyn J. (2002), Vitter J. (1998)]. These algorithms are successfully used for computing disjoint bipartite cliques of small size, PageRank, and strongly connected components. An overview of these algorithms and the experimental evaluation of their time performances are presented in [Donato D. (2006)].

Web characterization is just one of the possible research lines. A second important research line has been the development of models able to generate graphs which reproduce the same properties of the Web.

Random graphs are graphs generated accordingly to some random process [Bollobás B. (2001)]. The study of random graph was initiated by the seminal work of Erdös and Rényi (1960) in which the ER model was introduced: each edge between two nodes is set with a probability independently from the other edges in the graph.

However the topological properties observed in the Web graph, as for instance the indegree distribution are not visible in ER graphs. Moreover, the ER model is a static model, while the Web graph evolves over time when new pages are published or are removed from the Web. Albert *et al* (1999) initiated the study of 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 *preferential attachment* rule, i.e. with probability proportional to the indegree. The intuition behind the model is that newly generated nodes will preferably link to high popular nodes. This model, known as the *Evolving network*,

shows a power law distribution over the indegree of the vertices with exponent roughly 2 when the number of edges that connect every vertex to the graph is 7. The power law distribution is presented in Section 2.1.

A second popular model is the *Copying model* proposed by Kumar *et al.* (2000). Such a model allows us to explain other relevant properties observed in the Web graph. For every new vertex entering the graph a prototype vertex v is selected at random. A constant number *d* of links connect the new vertex to previously inserted vertices. The model is parameterized on a *copying factor*. The end-point of a link is either copied with probability *p* from a link of the prototype vertex, or it is selected at random with probability *1-p*. The copying event aims to model the formation of a large number of bipartite cliques in the Web graph.

A detailed description of these models goes beyond the scope of this chapter. A good survey of generative models for power-law distributions is given in [Mitzenmacher M. (2003)].

Preliminaries

In order to allow a separate reading of the various section we just recall some basic graph theoretic definitions and algorithms we need in this chapter, even if they have already been presented in the Introduction to this contribute. A complete overview can be found in [Cormen T. H. (1992)].

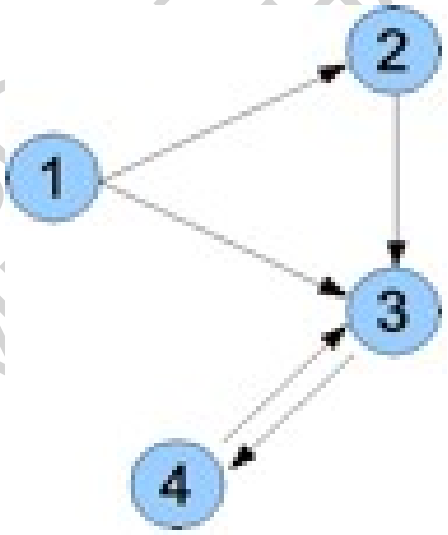

Figure 1. A direct graph

Graph: A *graph* is an abstract representation of a set of objects, some of which are connected by a relation. The set of the objects is called *vertices* (or *nodes)* and the connections are represented by *edges* (or *links*). If we denote with V the set of the vertices and with E the set of the edges, a graph is given by $G=(V,E)$. If the relation among vertices is non-directional, the graph is said to be *undirected*, in the

converse case, we have a *directed graph* or *digraph*. Figure 1 shows a direct graph with 4 nodes and 5 edges. A graph can be described by the so called *adjacency matrix* $\mathbf{A}$ which is a square matrix whose number of rows and edges is given by V. The element a_{ij} is 1 if there is a direct edge from the node i pointing to the node j, 0 otherwise. For undirected graphs, the adjacency matrix is symmetric.

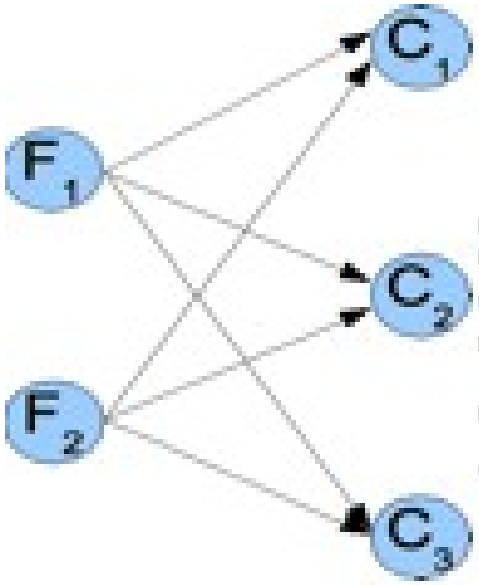

Figure 2. A bipartite cliques

Degree: The degree of a vertex is the number of edges incident to it. In a directed graph the indegree refers to the number of incoming edges to the node and the outdegree to the number of outgoing ones. In Figure 1, the node 3 has indegree 3 and outdegree 1.

Bipartite Cliques: A *bipartite cliques* is made of two sets of nodes; all the nodes in the first set (the *fan* set) point to each node of the second one (the *center* set). An example is shown in Figure 2: we have on the left side the set of the fan nodes (labeled F_1 and F_2), all of them pointing to all center nodes on the right side (labeled $\mathrm{C}_1, \mathrm{C}_2$ and C_3 C3).

Walk: A walk is an interleaved sequence of vertices and edges $v_0, e_0, v_1, e_1, \ldots, v_{n-1}, e_{n-1}, v_n$ in which each edge has the property to be incident to the two nodes immediately preceding and after.

Connected components: A subset of nodes S forms a *connected component (CC)* in a undirected graph G, if there is a walk between any two vertices $u,\ v \in S$.

Weakly Connected components: A *weakly connected component (WCC)* is a maximal group of vertices S that are mutually reachable by violating the edge directions in a directed graph G.

-
-
-

Bibliography

Auerbach F. (1913) "Das Gesetz der Bevolkerungskonzentration". *Petermanns Geographische Mitteilungen* **59**, 74–76. [First observation of the power laws]

Albert R., Jeong H., and Barabási A.-L. (1999) "Diameter of the World Wide Web" *Nature*, **401** 130. [One of the first papers on scale-free networks in information technology]

Baeza-Yates R., Castillo C. and Jean F. S. (2004) "Web Dynamics", *Web Dynamics, Structure and Page Quality*, 93–109. Springer [A paper on the evolution of the topology of the World Wide Web]

Barabási A.-L., Albert R., and Jeong H. (1999) "Mean-Field Theory for Scale-Free Random Networks" *Physica A* **272**. 173–189 [The definition and theoretical analysis of the Barabási-Albert model]

Barabási A.-L., Albert R. (1999) "Emergence of scaling in random networks", *Science* **286,** 509. [The first paper on the widespread occurrence of scale-free networks]

Becchetti L., Castillo C., Donato D., Fazzone A. (2006) "A Comparison of Sampling Techniques for Web Characterization" *Procs. of LinkKDD*, (Philadelphia, Pennsylvania), 8. [A paper on the minimum set of quantities necessary to describe the WWW]

Bianchini M., Gori M., Scarselli F. (2005) "Inside PageRank" *ACM Trans. Internet Technol.*, **5**, 1, 92-128. [An introduction to the PageRank algorithm]

Bollobás B. (2001) "*Random Graphs"*, Cambridge University Press. [The basic text on Graph theory]

Berners-Lee T. (1989) "Information Management: A proposal", CERN. [The first document describing the World Wide Web]

Berners-Lee T. (2000) *Weaving the Web*, 23 Harper Collins. [Presentation of the first implementation of the World Wide Web]

Boldi P., Codenotti B., Santini M. and Vigna S. (2004) "Ubicrawler: scalable fully distributed web crawler", *Software: Practice & Experience*, **34** (8), pp. 711–726. [A paper describing a piece of software made available by the authors to collect web data]

Boldi P. and Vigna S. (2004a). The Web Graph Framework I: Compression Techniques. In *Proc. of the Thirteenth International World Wide Web Conference (*Manhattan, USA), pp. 595-601. [A review of the compression techniques used to store www data]

Bordino I. and Donato D. (2009) "Dynamic characterization of a large Web graph" *1st International Conference On Web Science.* (Athens, Greece. 2009) [A study on the evolution of a subset (.uk) of the WWW]

Brewington B. E. and Cybenko G. (2000) "Keeping up with the changing web" *Computer*, **33**, 52–58. [An estimate of the average time during which a collection of web dataset remains valuable]

Brin S., Page L. (1998) "The anatomy of a large-scale hypertextual web search engines*" Computer Networks and ISDN Systems*, **30**, 107–117. [Seminal paper on PageRank]

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata S., Tomkins A. and Wiener J. (2000). Graph structure in the web. *Comput. Netw.* **33**, pp. 309–20. [First extensive study of the

characteristic features of the graph structure induced by the Web]

Caldarelli G. (2007) "*Scale Free Networks"* Oxford University Press. [Textbook on networks]

Cho J. and Garcia-Molina H. (2003) "Estimating frequency of change", *ACM Transactions Internet Technology*, **3**, 256–290. [An estimate of the change frequency of data to improve web crawlers]

Cormen T. H., Leiserson C. E. and Rivest R. L. (1992) "Introduction to Algorithms". Cambridge, MA: MIT Press. [A textbook on algorithms on data manipulation]

Dill S., Kumar R., McCurley K., Rajagopalan S., Sivakumar D. and Tomkins A. (2001) "Self-similarity in the Web", *Proceedings.of the 27th International Conference On Very Large Data Bases* (Rome, Italy, Sept.11-14) Morgan Kaufmann Publisher. [A paper on the self-similar properties of the WWW]

Donato D., Laura L., Leonardi S., Meyer U. Millozzi, Sibeyn J.F. (2006) "Algorithms and Experiments for the Webgraph.", *Journal of Graph Algorithms and Applications* **10**, 219 – 236. [An analysis of the WWW based on the largest crawl available at that time]

Donato D., Laura L., Leonardi S. and Millozzi S. (2007) "The Web as a graph: how far we are", *ACM Transactions of Internet Technology* **7**, 23. [A review on the WWW studies]

Donato D., Leonardi S., Millozzi S. and Tsaparas P. (2008) "Mining The Inner Structure of the Web Graph." *Journal of Physics A*, **41,** 224017. [A comparison between different crawls of the WWW together with a proposal of a model]

Donato D., Leonardi S. and Tsaparas P. (2008 b). "Stability and Similarity of Link Analysis Ranking Algorithms.", *Special Issue of Internet Mathematics devoted to the ANAW workshop*, **3**, 479-507. [A paper on the invariance of PageRank related techniques in a dynamical WWW].

Erdös P. and Rényi R. (1960) " On the Evolution of Random Graphs", *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5,** 17-61. [The first paper on Random Graphs]

Fetterly D., Manasse M., Najork M. and Wiener J. (2003) "A large-scale study of the evolution of web pages" *Proceedings of the 12th International Conference on WWW*, 669–678. [A paper on the dynamic of the WWW]

Gomes D. and Silva M. J. (2006) "Modelling information persistence on the web", *Proceedings of the 6th International Conference on Web engineering*, 193–200. [A paper on the possibility to track information during web evolution]

Gulli A. and Signorini A. (2005) "The indexable Web is more than 11.5 billion pages" *Proceedings. of the 14th International Conference on WWW* (Chiba, Japan), 902-903. [An estimate of the Web size at that time]

Gupta S., Anderson R. M. and May R. M. (1989) "Networks of sexual contacts: implications for the pattern of spread of hiv" *AIDS*, **3**,807–817. [A paper using the geometry of social networks to understand the better policy to avoid pandemias of sexually-transmitted diseases].

Hendler J., Shadbolt N., Hall W., Berners-Lee T., and Weitzner D. (2008) "Web Science: An Interdisciplinary Approach to Understanding the Web", *Communications of the ACM*, **51**, 60-69. [A review of different approaches used for the study of the WWW]

Kleinberg J. (1997) "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, **46**, 604–632. [Introduction to the HITS algorithm]

Koehler W. (2002) "Web page change and persistence—a four-year longitudinal study", *Journal of American Society of Information Science Technology*, **53**, 162–171. [A long term study on the effect of dynamics to the search of information in the WWW]

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. "Crawling the web for emerging cyber communities" *Proceedings of the 8th WWW Conference*., 403–416 (1999). [A paper on the method to extract communities in the WWW]

Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., and Upfal E. (2000) "Stochastic models for the web graph" *Proceedings of 41st FOCS*, 57–65. [An analysis of the dynamical evolution of about 150 websites during a year]

Mitzenmacher M. (2003) "A brief history of generative models for power law and lognormal distributions", *Internet Mathematics*, **1.** [A nice review on all the models that produce self-similar structures as scale-free networks]

Newman M. E. J. (2002) "Assortative mixing in networks", *Physical Review Letters*, **89**, 208701. [The paper where is introduced the concept of assortativity]

Newman M. E. J. (2005) "Power Laws, Pareto Distributions and Zipf's Law", *Contemporary Physics* **46,** 323–351. [A paper on the theory of scale-invariant distributions as the power laws]

Ntoulas A., Cho J. and Olston C. (2004) "What's new on the web?: the evolution of the web from a search engine perspective" *Proceedings of the 13th International Conference on World Wide Web*, 1–12. [A paper on the analysis of the Web dynamics from search queries]

Pennock D., Flake G., Lawrence S., Glover E. and Giles C. (2002) "Winners don't take all: Characterizing the competition for links on the web" *Proceedings of the National Academy of Science (USA)*, **99**, 5207–5211. [An analysis of the WWW and the various ways to assess theimportance of a page]

Sibeyn J., Abello J. and U. Meyer. (2002) "Heuristics for semi-external depth first search on directed graphs" *Proceedings of the 14th annual ACM symposium on Parallel algorithms and architectures*, 282–292. [A paper showing fast algorithms to compute the components of a large graph as the WWW]

Vitter J. S. (1998) "External memory algorithms" *Proceedings of the 6th Annual European Symposium on Algorithms (Lecture Notes in Computer Science)*, **1461**, 1–25. [A review of this field useful to handle massive datasets]

Zipf G. K. (1949) *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, MA. [Seminal book on scale-invariance in Social Systems]

Biographical Sketch

Debora Donato is currently a scientist at Yahoo! Labs in Sunnyvale. She obtained a Ph. D. in Computer Engineering in 2005 from the University of Rome "La Sapienza". She has been a visiting scientist at Basic Research Unit of the University of Helsinki, Finland, and at the Max Planck Insitut fur Informatik, Saarbrucken, Germany. Her research interests include Web Information Retrieval, Link Analysis, Algorithms for the Characterization of the Web, Complex Networks, Social networks and P2P networks.

Her scientific results appeared in prestigious venues such as Journal of the IEEE Computing in Science and Engineering, the ACM Transactions on Internet Technology (TOIT), the ACM Transactions of the Web (TWEB), the European Journal of Physics A and B. She is also author of a number of book chapters describing the algorithms, the techniques and the challenges in Web Search.

She has been member of a number of European and Italian projects, including Dynamically Evolving Large Scale Information Systems (DELIS), COevolution and Self-organization In dynamical Networks (COSIN), Algorithms for the Next Generation Internet and Web (ALGO-NEXT), and Algorithms for Internet and the Web (ALINWEB). She has actively participated to all the activities of the above projects and she has been in charge of a number of deliverables and technical reports.