MINING THE INNER STRUCTURE OF THE WEB GRAPH

DEBORA DONATO*, STEFANO LEONARDI*, STEFANO MILLOZZI*, PANAYIOTIS TSAPARAS+

ABSTRACT. The *bow-tie* picture, presented by Broder et al. [3] in 2000, has been up to now the only strong characterization of the well defined structure of the World Wide Web, namely the hyperlinked graph induced by the links among the static html pages. This evocative picture is a clear abstraction of the macroscopic arrangement of the different subsets that comprise the Web graph but, nevertheless, it is quite uninformative with respect to its finer details. In this paper we mine the inner structure of the Web graph. We have discovered that the scale-free properties permeate all the components of the bow-tie which exhibit the same macroscopic properties as the Web graph itself. However, close inspection reveals that their inner structure is quite distinct. We show that the Web graph does not exhibit self similarity within its components, and we propose a possible alternative picture for the Web graph, as it emerges from our experiments.

keywords: Web graph, measurements, statistical properties, bow-tie structure

1. INTRODUCTION

Since the end of the '90s, the World Wide Web has been the subject of an intensive research work in various disciplines. Its unexpected and rapid growth has attracted the attention of the scientific community, interested from one side in the study of the structural properties of the Web and, from the other, in models able to predict the behavior of this evolving "organism".

The first large-scale study of the Web graph was performed by Broder et al. [3] and it revealed that the Web graph contains a giant component that consists of three distinct components of almost equal size: the CORE, made up of a single strongly connected component; the IN set, comprised by nodes that can reach the CORE but cannot be reached by it; the OUT set, consisting of nodes that can be reached by the CORE but cannot reach it. These three components form the well known *bow-tie* structure of the Web graph, shown in Figure 1.

The bow-tie picture describes the macroscopic structure of the Web. However, very little is known about the inner structure of the components that comprise it. Broder et al. [3] pose it as an open problem to study further the structure of those components. Understanding the finer details of the Web graph is an interesting problem on its own, but it is also important in practice in order to improve the performance of algorithms that rely on the link structure of the Web. Furthermore, it could be useful for refining the existing stochastic models for the Web [1, 12, 8].

Partially supported by the EU under contract 001907 (DELIS) and 33555 (COSIN), and by the Italian MIUR under contract ALINWEB..

* Dipartimento di Informatica e Sistemistica, Universitá di Roma "La Sapienza", Via Salaria 113, 00198 Roma, Italy. { donato,Stefano.Leonardi,millozzi}@dis.uniroma1.it

+ Department of Computer Science P.O. Box 68 (Gustaf HŁllstrmin katu 2b), FIN-00014, University of Helsinki, Finland. tsaparas@cs.helsinki.fi.

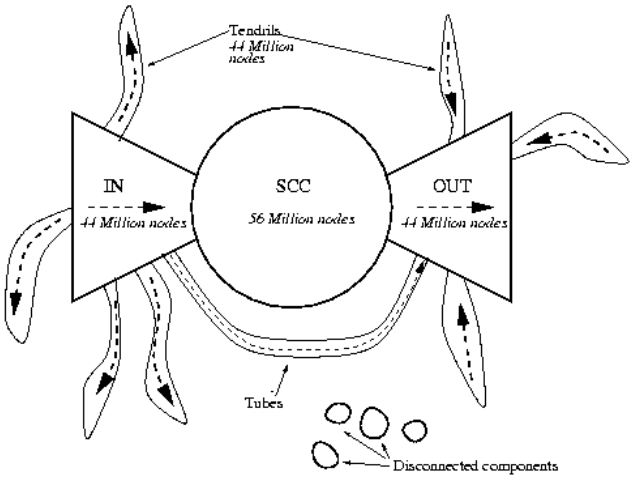


FIGURE 1. The bow-tie structure of the Web graph

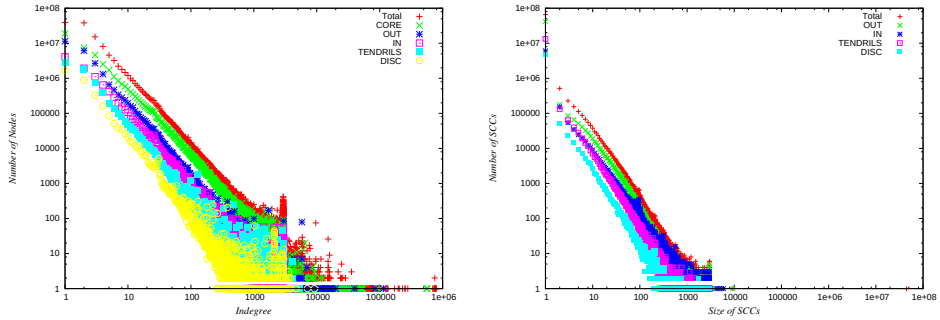
	Italy	Indochina	UK	WebBase
nodes	41.3M	7.4M	18.5M	135.7M
edges	1.15G	194.1M	298.1M	1.18G
CORE	29.8M (72.3%)	3.8M (51.4%)	1.2M (65.3%)	44.7M (32.9%)
IN	13.8K (0.03%)	48.5K (0.66%)	312.6K (1.7%)	14.4M (10.6%)
OUT	11.4M (27.6%)	3.4M (45.9%)	5.9M (31.8%)	53.3M (39.3%)
TENDRILS	6.4K (0.01%)	50.4K (0.66%)	139.4K (0.8%)	17.1M (12.6%)
DISC	1.25K (0%)	101.1K (1.4%)	80.2K(0.4%)	6.2M (4.6%)

TABLE 1. Sizes and bow-tie components for the different crawls and the Alta Vista graph

2. EXPERIMENTS AND RESULTS

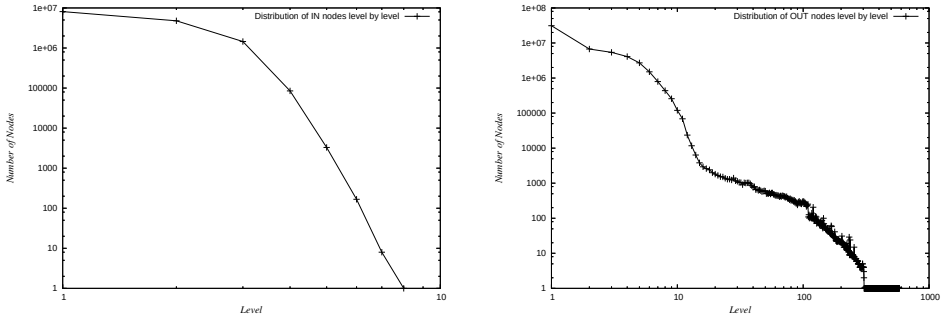
We experiment with four different crawls. The first three crawls are samples from the Italian Web (the `.it` domain), the Indochina Web (the `.vn, .kh, .la, .mm,` and `.th` domains), and the UK Web (the `.uk` domain) collected by the "Language Observatory Project" and the "Istituto di Informatica e Telematica" using UbiCrawler [2]. The fourth crawl is a sample of the whole Web, collected by the WebBase project at Stanford in 2001. The sizes of the crawls are shown in Table 1.

2.1. **Macroscopic measurements.** We have repeated the experiments of Broder et al. [3] and we have observed the same macroscopic properties previously reported: the in-degree, the out-degree, the SCC size distributions follow a power-law, and the graphs have a bow-tie structure. The relative sizes of the components of the bow-tie are shown in Table 1, where we can observe that they vary from crawl to crawl. These discrepancies between the crawls can most likely be attributed to different crawling strategies and capabilities, rather than to the evolution of the Web. The first three crawls are relatively recent, and all crawls are generated using a small number of starting points. Unfortunately, large-scale crawls are not publicly available.



(a) In-degree distributions (c) SCC size distributions

FIGURE 2. Macroscopic measures for all components



(b) Distribution of IN nodes level by level (c) Distribution of OUT nodes level by level

FIGURE 3. Characteristics of the IN and OUT components

2.2. **The inner structure of the bow-tie graph.** As a first step in the understanding of the individual components we compute the same macroscopic measures as for the whole Web graph. We compute the in-degree, out-degree and SCC size distributions for each of the IN, OUT, TENDRILS and DISC graphs. Figure 2 shows the plots of the distributions for each component and for the whole graph, for the case of the WebBase crawl. It is obvious that the same macroscopic laws that are observed on the whole graph are also present in the individual components.

2.2.1. *The structure of the IN and OUT components.* Given the fact that the in-degree, out-degree, and SCC size distributions in the IN and OUT components are the same as for the whole Web graph, we wonder if the Web has a *self-similar* structure [6, 12], that is if the bow-tie structure repeats itself inside the IN and OUT components.

The first indication that the self-similarity conjecture is not true comes from the fact that there exists no sizable SCC in the IN and OUT components that could play the role of the CORE in a potential bow-tie. Moreover we surprisingly discovered that there is no giant weakly connected component (WCC) in either of the two components. In fact, there is a large number of WCCs per component and their sizes follow a power law distribution. Statistics for all graphs are reported in Table 2.

In order to better understand how the nodes in IN and OUT are arranged with respect to the CORE, we performed the following experiment. We condensed the CORE in a single node and we performed a forward and a backward BFS. This allows us to split the nodes in the IN and OUT components in *levels* depending on their distance from the CORE. The depths of the components are shown in

	Italy	Indochina	UK	WebBase
The IN component				
nodes in IN	13.8K (0.03%)	48.5K (0.66%)	312.6K (1.69%)	14.4M (11%)
max SCC	1,590	7,867	4,171	5,876
number of WCCs	1,633	117	62K	3.68M
max WCC	4,085 (29.5%)	13.2K (27.2%)	8,246 (2.7%)	197.5K (1.3%)
singleton WCCs	1,543 (11.15%)	63 (0.13%)	56K (17.89%)	3.2M (22.46 %)
The OUT component				
nodes in OUT	11.4M (27.6%)	3.4M (45.9%)	5.9M (31.8%)	53.3M (39%)
max SCC	19,170	39,283	26,525	9,349
number of WCCs	3.73M	729,6K	1.97M	25.4M
max WCC	1.43M (12.52%)	335.9K (9.85%)	457.4K (7.75%)	14.94M (28.01%)
singleton WCCs	3.49M (30.6%)	672K (19.71%)	1.84M (31.11%)	24.48M (45.91%)
The CORE component				
nodes in CORE	29.8M (72.3%)	3.8M (51.4%)	1.2M (65.28%)	44.7M (33%)
entry points	10.2K (0.03%)	2.3K (0.06%)	106.3K (0.88%)	2.6M (5,87%)
exit points	15.6M (52.2%)	2.3M (59.6%)	4.8M (39.8%)	29.6M (72.03%)
bridges	6.25K(0.02%)	1.5K (0.04%)	61.8K (0.51%)	2M (4.58%)
connectors	1.7M (5.71%)	164.2K (4.32%)	537.9K (4.45%)	2.96M (6.63%)
petals	325.3K (1.09%)	52.5K (1.38%)	138K (1.14%)	1.4M (3.14%)

TABLE 2. Statistics for the IN, OUT and CORE components for each crawl

	Italy	Indochina	UK	WebBase
depth IN	2	11	15	8
depth OUT	26	21	25	580

TABLE 3. IN and OUT depth

Table 3. In all graphs, the depths of the components are relatively small. Furthermore, most nodes are concentrated close to the CORE. Typically, about 80-90% of the nodes in the OUT component are found within the first 5 layers. For the WebBase graph, although the OUT is much deeper, with 580 levels, more than 58% of its nodes are at distance 1 from the CORE, and 93% are within distance 5. Furthermore, after level 305 there exists only a single chain of nodes that extends until level 580, making the effective depth of the OUT 305. The node distributions, level by level, for the WebBase graph are shown in Figure 3(b) and 3(c), for the IN and OUT sets respectively. The plots are in logarithmic scale.

Therefore, we conclude that the IN and OUT components are shallow and highly fragmented. They are comprised of several sparse weakly connected components of low depth. Most of their volume consists of nodes that are directly linked to the CORE.

2.2.2. *The structure of the CORE.* We concentrate the study of the CORE on two main aspects:

(1) its relation with the IN and OUT components
(2) its connectivity properties

We address the first question measuring the *entry points* to the CORE (nodes that are pointed to by at least one node in the IN component), the *exit points* (nodes that point to at least one node in the OUT component and the *bridges* (nodes that are both entry and exit points). In Table 2 we can note that

in-deg	del	out-deg	del	total del	max SCC	max SCC %	SCC num
4,000	1.1K	233	1,154	2,263	42.2M	94.4%	595K
2,600	9.9K	185	10K	20.6K	39.8M	89.0%	1.75M
1,750	26K	158	25K	51K	37M	82.9%	3M
1,000	52K	130	54K	105K	33.7M	75.5%	4.75M
500	112K	105	108K	219K	29.4M	66.1%	7M
225	259K	82	227K	487K	23.5M	53.3%	10M
120	518K	62	499K	949K	17.8M	40.8%	13M

TABLE 4. Deleting nodes with high in-degree and out-degree

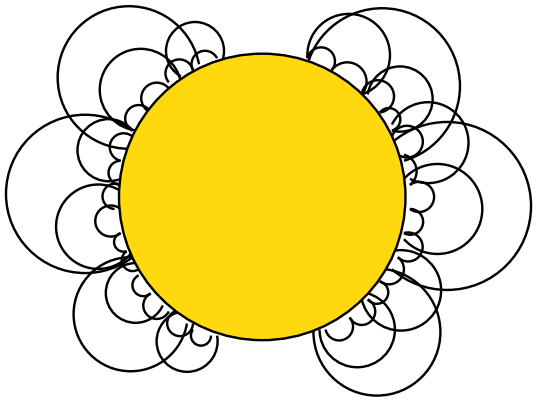


FIGURE 4. The daisy structure of the Web

the majority of the nodes in the CORE is connected to the "outside" world. In the WebBase crawl, this number is around 80% of the whole CORE, while the "deep CORE" consists of a little more than 20%.

Regarding the connectivity, we observe that there are few nodes with just one in and out link that could make the CORE weakly connected. We define this kind of nodes *connectors* (or *petals* if the source of the incoming link, and the target of the out-going link are the same node). Moreover the CORE seems resilient to targeted attacks performed by deleting not only nodes with total degree bigger than a prefixed threshold but also the k nodes with the highest in-degree and k nodes with the highest out-degree. In the first case, we observe that the threshold on the total degree must become as low as 100 in order to obtain an SCC of size less than 50% of the CORE. For the second kind of attack, the results are reported in Table 4.

There are two ways to interpret these results. The first is that there are no obvious *failure points* in the CORE, that is, strong hubs or authorities that pull the rest of the nodes together, and whose removal from the graph causes the immediate collapse of the network. In order to disconnect the CORE you need to remove nodes with sufficiently low degree. On the other hand, note that we managed to reduce the largest SCC to 35-40% of the original by removing about 1M nodes. However this is less than 1% of the total nodes. In that sense the CORE is vulnerable to targeted attacks.

3. Discussion and Future Work

In this paper we undertook a study of the Web graph at a finer level. We observed that the ubiquitous presence of power laws describing several properties at a macroscopic level does not necessarily imply self-similarity in the individual components of the Web graph. Indeed, the different components have quite distinct structure, with the IN and OUT being highly fragmented, while the CORE being well interconnected.

Our work suggests a refinement of the bow-tie pictorial view of the Web graph [3]. The bow-tie picture seems too coarse to describe the details of the Web. The picture that emerges from our work can better be described by the shape of a *daisy* (Figure 4): the IN and OUT regions are fragmented into large number of small and shallow *petals* (the WCCs) hanging from the central dense CORE.

A deeper understanding of the structure of the Web graph may also have several consequences on designing more efficient crawling strategies. The fact that IN and OUT are highly fragmented may help in splitting the load between different robots without much overlapping. Moreover, the fact that most of the vertices are at few hops from the CORE may explain why breadth first search crawling is more effective than other crawling strategies [11].

As a concluding remark, we observe that we are still very far from devising a theoretical model that is able to capture the finer connectivity properties of the Web graph.

References

[1] A.L. Barabasi and A. Albert. Emergence of scaling in random networks. *Science*, (286):509–512, 1999.

[2] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711–726, 2004.

[3] 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, 2000.

[4] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to algorithms*. MIT Press and McGraw-Hill, 1992.

[5] S. Millozzi D. Donato, S. Leonardi and P. Tsaparas. Mining the inner structure of the web graph. Technical report, DELIS-TR-157, http://delis.upb.de/docs/, 2005.

[6] 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.

[7] P. Erdös and A. Rèny. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[8] 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*, 1999.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber communities. In *WWW*, 1999.

[10] L. Laura, S. Leonardi, S. Millozzi, U. Meyer, and J.F. Sibeyn. Algorithms and experiments for the webgraph. In *European Symposium on Algorithms (ESA)*, 2002.

[11] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *WWW Conference*, 2001.

[12] 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.

[13] J.F. Sibeyn, J. Abello, and U. Meyer. Heuristics for semi-external depth first search on directed graphs. In *SPAA*, 2002.