# The Web as a graph: *measurements, models, and methods*

Michelle Ackermann <mackerma@student.ethz.ch>

Seminar of Distributed Computing WS 03/04

## 1. Introduction

The pages and hyperlinks of the Web may be viewed as nodes and edges in a directed graph, the Web graph. The purpose of the paper [1] is to give a survey of the achievements concerning the Web graph.
The Web graph is an interesting subject to study: It consists of more than two billion pages [4] created by the uncoordinated actions of tens of millions of individuals. Therefore the Web has been widely believed to lack structure and organization as a whole. Analyses of the Web graph have revealed a structure that is valuable for organizing information, improving search algorithms and understanding the Web in a social context.

## 2. Summary

The paper begins by describing two algorithms that operate on the Web graph, one for finding high-quality pages on a topic query, the other for enumerating all topics of the Web, which may be useful to define and categorize topics for Web portals. There seem to be two types of pages in the Web, the hub pages and the authoritative pages. The algorithms take that into account.
The second part deals with measurements and properties of the graph that were discovered by running the algorithms. The main observation is that the in- and out-degree of the nodes seem to be Zipfian distributed. The result is that traditional random graph models do not explain the observations.
The last section presents a family of random graph models. It proposes that random copying is a mechanism to create Zipfian degree distributions.

## 3. Related Work

The first step to study the Web graph is to get large parts of it. It is impossible to get the whole graph. The method to compute such parts is called crawl. It's not possible to get a snapshot, because a crawl can take months.
Next the graph has to be stored efficiently, so that traversal and searching can be done. Naive graph representation schemes can significantly increase query execution time. [5] proposes an efficient way to structure and store Web graphs.
The structure of the Web has been analyzed by [2]. The Web has the shape of a bowtie. It has long been thought that there exists a short path between every pair of pages in the Web. But this is only possible in the knot of the bowtie.
Another study [3] shows that the Web has scale-free properties. That means the Web consists of many bowties, induced by a particular topic, region, etc.

## 4. Perspectives

Many important questions still have no answers:
The influence of the crawl technique on the studies is not clear, although it may be important. For example it has to be verified that the bowtie structure is not an artefact of the crawl technique.
Very few results are known about the community structures of the Web. Knowing the communities on the Web and the relation they share would give deep insight in the behaviour of the Web and of its users, which would help to improve searching, crawling and browsing.
Moreover the dynamics of the Web have to be better included in the studies. The Web is very dynamic, but most of the studies analyzed a static graph. For example one could investigate the difference between two crawls. Most Web portals have a static set of topics, which is adapted only from time to time. Algorithms could be designed to discover new topics dynamically and organize them automatically.
There are other interesting subgraphs for enumeration. For example webrings, cliques and directed trees, for which algorithms could be developed.
The relationship between copying and Zipfian distributions should be studied deeper. It's a challenge to efficiently analyze such graphs, because copying generates myriads of dependencies. The results could also be used outside the Web, i.e. scientific citations, term frequencies or the genome.

## References


[1] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: *The Web as a graph: measurements, models, and methods*, 1999
[2] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. L. Wiener: *Graph structure in the web*, 2000
[3] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, A. Tomkins: *Self-Similarity in the Web*, 2001
[4] J.-L. Guillaume, M. Latapy: The *Web Graph: an Overview*, 2002
[5] S. Raghavan, H. Garcia-Molina: *Representing Web Graphs*, 2003