A Teapot Graph and Its Hierarchical Structure of the Chinese Web

Jonathan J. H. Zhu
Dept of Media & Communication
City University of Hong Kong
j.zhu@cityu.edu.hk

Tao Meng, Zhengmao Xie, Geng Li
School of EECS, Peking University
mengtao@net.pku.edu.cn
xzm@net.pku.edu.cn
ligeng@net.pku.edu.cn

Xiaoming Li
State Key Laboratory of Advanced Optical Communication Systems & Networks, Peking University
lxm@net.pku.edu.cn

ABSTRACT
The shape of the Web in terms of its graphical structure has been a widely interested topic. Two graphs, Bow Tie and Daisy, have stood out from previous research. In this work, we take a different approach, by viewing the Web as a hierarchy of three levels, namely page level, host level, and domain level. Such structures are analyzed and compared with a snapshot of Chinese Web in early 2006, involving 830 million pages, 17 million hosts, and 0.8 million domains. Some interesting results have emerged. For example, the Chinese Web appears more like a teapot (with a large size of SCC, a medium size of IN and a small size of OUT) at page level than the classic bow tie or daisy shape. Some challenging phenomena are also observed. For example, the INs become much smaller than OUTs at host and domain levels. Future work will tackle these puzzles.

Categories and Subject Descriptors
H.1.1 [**Models and Principles**]: Systems and Information Theory

General Terms:
Experimentation, Measurement, Verification

Keywords:
Bow Tie Graph, Daisy Graph, Teapot Graph, Self Similarity

1. INTRODUCTION
How does the World Wide Web look like as a graph? The question is not only important for information scientists in networking traffic, search engine optimization, and other areas of information technology but also interesting for social scientists who are concerned about the diffusion, use, and impact of the technology. The pioneering work by Broder et al. [1] suggests that the Web looks like a Bow Tie of four distinct components, each in a roughly equal size, including a strongly connected component (SCC, which accounts for 29% of the total web pages), an IN component (24%), an OUT component (24%), and a disconnected component (DISC and tendrils, 24%). While highly heuristic, the Bow Tie model has been considered an oversimplified representation of the complex Web structure.

A major revision is the Daisy Graph proposed by Donato et al. [2], in which the IN and OUT components are described as a large number of "small and shallow petals" hanging from a disproportionately larger and denser SCC in the center. Donato et al. analyzed separately a *global* Web (based on data from Alta Vista and WebBase) and three *national* Webs (including Italy, UK and five Indochina countries[1]). Sharp differences between the two types of Web were found. The global Web largely resembles Bow Tie. However, the three national Webs all show only two components: a dominant SCC (51-72%) and a visible OUT (28-46%) (columns 1-3 of Table 1). In short, national Webs appear to be more connected (with a larger SCC) and polarized (with two competing components) at the same time. Liu et al. found a more dominant SCC (80%) in 140M web pages from China in 2003 [3]. However, the OUT (7%) was nearly negligible in the graph.

In this paper, we report an empirical test of the Daisy Graph, against the backdrop of Bow Tie, based on a larger-scale set of more recent web pages from China. Our aim is twofold: to replicate the Daisy Graph with a much larger national Web and to explore the hierarchical structure of the national Web. Donato et al. found no evidence for self-similarity between global and national Webs [2]. We move one step further by examining whether there is a self-similarity across different layers within a national Web.

2. THE EXPERIMENT METHOD

2.1 Crawl of Web Pages
We crawled over 830M web pages from all web sites within China, as identified by the IP addresses assigned to the country, between January and February 2006. This number is close to the official estimate (947M) of Chinese web pages at that time.

2.2 Hierarchical Structure
We consider the Web a hierarchical system in which web pages are the bottom layer, which is operationally defined by a complete URL, such as http://net.pku.edu.cn/~xzm/index.html. Host is considered the second layer, which is defined as the collection of web pages hosted on a web server. More precisely, a host corresponds to all the pages under the address of http://.../ (i.e., the part between "http://" and the first "/" from the left).

We then regard institutional web sites or domains to be the third layer. Intuitively, a domain corresponds to a registered URL from a national authority (e.g., CNNIC), for instance, http://pku.edu.cn, http://sina.com, and http://infomall.cn, etc. In practice, we maintain two sets of higher level "familiar domain names". To form a domain URL from a host URL, the program scans the

WWW 2008, April 21--25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

[1] Including Vietnam, Thailand, Lao, Cambodia, and Myanmar.

relevant URL from right to left and stops after seeing an unknown component.

If we wish to go further, we may say national Webs as the fourth layer,[2] and the global WWW as the top layer. In an analogue, individual pages are bricks, hosts buildings, domains street blocks, and national Webs cities of the global kingdom of WWW. It is both necessary and informative to examine the structure of each layer of the Web. In the current study, we confine to the three layers ranging from pages to domains.

3. RESULTS

3.1 Teapot Graph

Of the 837M pages crawled, 43 billion links are found, which amounts to 52 links per page, or almost twice as much as found in Italy (28 links/page) and Indochina (27 links/page) or 3 times as much as in UK (16 links/page). However, as shown below, the larger number of links per page in the Chinese Web does not necessarily result in a denser graph, though.

Table 1. Components of Chinese Web Graph

	Italy[1]	UK[1]	Indochina[1]	China (page-level)	China (host-level)	China (domain-level)
SCC	72.3%	65.3%	51.4%	44.1%	50.7%	63.3%
IN	0.03%	1.7%	0.7%	25.5%	1.4%	0.7%
OUT	27.6%	31.8%	45.9%	14.6%	47.4%	34.9%
DISC /Tendrils	0.01%	1.2%	2.1%	15.8%	0.5%	1.1%
Total	100%	100%	100%	100%	100%	100%
N of Pages	41.3M	18.5M	7.4M	836.7M	16.9M (hosts)	0.79M (domains)
N of Links	1.15G	194.1M	298.1M	43.28G	43.28G	43.28G

[1] Taken from [2].

As shown in Table 1 (column 4), the overall graph of Chinese Web departs significantly from the Bow Tie shape because the SCC accounts for a much larger share (44%) and the OUT a smaller share (15%) than the counterparts in Bow Tie. On the other hand, the Chinese Web has not become a Daisy yet because its SCC is still smaller than half of the graph and both OUT and Disc/Tendrils are still sizable. It appears that the Chinese Web looks more like a Teapot (Figure 1). The Teapot graph also differs from an earlier Chinese web graph [3], most noticeable in the size of SCC (44% vs. 80%), which might be attributed to differences in time, crawling strategy, and other factors.

3.2 Hierarchical Structure

Also shown in Table 1 (columns 5 and 6), the Chinese Web becomes increasingly closer to the Daisy shape when it is progressively aggregated from pages to hosts and domains. When links from individual pages are pooled together into hosts/domains, the latter expectedly become more connected, resulting in a larger SCC (from 44% at the page-level to 51% at the host-level and 63% at the domain level). The finding differs again from Liu et al. [3], in which the SCC became, surprisingly, smaller (from 80% to 67%) when aggregated from pages to hosts.

What cannot be predicted in advance is how the three other components realign after the aggregation. Interestingly, it turns out that the OUT component becomes a "winner" as it takes almost all the remaining share (47% at the host-level and 35% at the domain-level), which is similar to Liu et al. (from 7% to 30%) [3]. Nevertheless, it is necessary to note that what has changed here is the *relative proportion*, not the absolute size, of OUT. In fact, the absolute size of OUT has reduced as well in the aggregation (from 213M pages to 8M hosts and 277K domains).

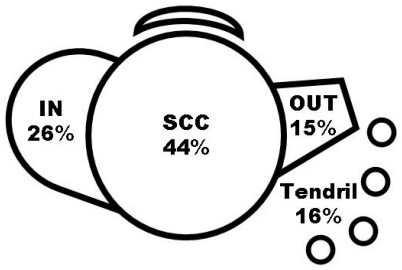

Figure 1. A Teapot Graph of Chinese Web

4. CONCLUSION AND FUTURE WORK

In this paper, we present a large-scale experiment on the graph properties of Chinese Web. A Teapot Graph is constructed as alternative to the classic Bow Tie or Daisy Graphs. A three-layer structure is further considered for the national Web. The most unexpected finding is the absence of self similarity between page-level and host/domain levels. In addition, the existence of a large number of hosts with only single page in the current data set may have introduced extraneous influences to the resulting graphs across all three levels.

In future work, we will examine the reasons behind the dramatic change in the relative proportions of IN and OUT and identify content, technical, and geographic features of the web pages and sites appearing in different components of the structures.

5. ACKNOLWEDGMENTS

The work was funded in part by NSFC (60573166), HKSAR CERG (CityU 1456/06H) and City University of Hong Kong SRG (7001882).

6. REFERENCES

[1] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the web. Computer Networks, 33(1-6), 309-320.

[2] Donato, D. Leonardi, S., Millozzi, S., & Tsaparas, P. Mining the inner structure of the Web graph. Eighth International Workshop on the Web and Databases (WebDB 2005), June 16-17, 2005, Baltimore, Maryland.

[3] Liu, G., Yu, H., Han, J. & Xue, G. (2005). China web graph measurements and evolution. In Y. Zhang et al. (Eds.): *APWeb 2005*, LNCS 3399, 668– 679.

[2] In large countries, there may be local and/or regional layers between institutional and national layers. We are currently exploring the role of provincial layer in the Chinese Web.