The Mobile Web Is Structurally Different

Apoorva Jindal
Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
Email: apoorvaj@usc.edu

Christopher Crutchfield
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: ccrutch@mit.edu

Samir Goel, Ravi Kolluri,
Ravi Jain
Google Inc.
Mountain View, CA 94043
Email: {sgoel, rkolluri,
ravijain}@google.com

***Abstract*—One of the premier applications on the global Internet is browsing the World Wide Web. The advent of advanced browser-enabled cell phones, high-speed wireless networks, and "unlimited-data" pricing plans is fueling the demand for Web access on mobile devices. Further, there is an increasing amount of content in the mobile Web, the set of web pages written in markup languages (CHTML, XHTML, and WML) designed specifically for consumption on mobile wireless devices. Understanding the structural properties of the WWW can be very helpful in a variety of applications, such as crawling the web more efficiently, or performing better search results ranking. So far, however, this line of investigation has been limited to the web consisting of HTML pages. In this study we examine the structural properties of the mobile web graph inferred from a crawl of mobile markup pages.**

We find that the mobile web graph differs in general from the fixed web in several important ways. Its connectivity is sparser than the fixed web and its node degree distributions fall off much more rapidly. We further analyze the web graph in terms of its bow-tie structure, which has been studied previously for the fixed web. The properties of the bow-tie structure for mobile web are quite different from those of the fixed web, such as having a smaller central core strongly connected component (SCC) and more disconnectedness. We also find the CHTML and XHTML/WML subgraphs of the mobile web subgraph differ significantly, indicating the influence of different usage and maturity of the mobile web in Japan compared to other countries. We also consider the domain-level graphs, where all nodes of a domain are collapsed into a single node and all inter-domain edges are hidden, and find notable differences between the fixed and mobile graphs.

To our knowledge this is the first study of the structural properties of the web graph. We briefly comment on the potential implications of the findings, focusing on crawl as an example application.

I. Introduction

One of the premier applications on the global Internet is browsing the World Wide Web, with a key task being searching for information using a search engine. We define the mobile web to be the web of pages written in markup languages like CHTML, XHTML and WML, designed for, or particularly suitable for, consumption on mobile wireless devices such as cellular phones. The fixed web is the set of web pages written in HTML.

The mobile web has been growing steadily. In many countries the rate of growth of mobile devices exceeds that of desktop PCs. With the increased coverage of high-bandwidth wireless networks, and the availability of hand-held devices to browse the web over these networks, the mobile web is expected to become even more important in the future.

In general, mobile phones have a small display size, and are connected to a relatively low-bandwidth cellular network. These factors influence the characteristics of mobile pages. They are often smaller in length, have fewer outgoing links, and have fewer images, making them different in composition from fixed web pages. Further, consider the *web graph* to be the graph formed by considering each page on the web a node and each hyperlink an edge between nodes. These factors also influence the properties of the graph formed by interconnection of these pages. Fewer outgoing links imply that the graph formed by the mobile web is sparse compared to the fixed web. These characteristics have significant implications for search engines that crawl and serve the mobile web.

The structural properties of the fixed web have been well studied [1], [5], [7], and models have been proposed that represent the structure and evolution of the web [9], [10]. In contrast, very little is known the structural properties of the mobile web. In this paper we study the structure of the mobile web and compare it to that of the fixed web. To the best of our knowledge, this is the first study that characterizes the sparseness and the structure of the mobile web graph in statistical terms.

Our contributions in this paper are as follows. We describe first-order statistical characteristics of the mobile web graph, such as average node degrees. We then analyze the graph using public-domain software that has been previously used for analysis of the fixed web [6]. We show that, like the fixed web, the mobile web is composed of a *bow-tie structure*, which is a model proposed by Broder *et al.* [1] to study the fixed web, and we characterize the properties of this structure. We contrast the bow-tie structures for the fixed and mobile web graphs, highlighting differences in relative sizes of the components of the bow-tie and in degree distributions. We also consider the CHTML and XHTML/WML subgraphs separately, since the former subgraph corresponds largely to pages for consumption in Japan, a highly evolved mobile data access market. We show that the CHTML and XHTML/WML subgraphs differ significantly in their structural properties. Since a significant proportion of the edges in both the fixed and mobile graph are among pages belonging to the same domain, we then consider the corresponding domain-level graphs, where all nodes of a domain are collapsed into a single node and all intra-domain edges are hidden. The comparison of the domain-level fixed and mobile web graphs also shows interesting differences.

In addition, we illustrate the value of understanding the web structure by considering a particular application, web crawl. Web crawl is a fundamental web operation that, while appearing deceptively simple, is challenging and resource-intensive. It can be made more efficient and effective when the structure of the graph that is being crawled is taken into account.

This paper is organized as follows: Section II summarizes the background information required to understand the paper while Section III reviews prior work. We have used multiple corpora in this study to analyze and contrast the structural properties of mobile web graph. Each corpus consists of set of crawled web pages that share a common characteristics. Section IV describes each corpus and the processing steps used to generate extract the structural properties. Section V compares the structural properties of the fixed web graph and the mobile web graph. Sections VI and VII study the structural properties of language subgraphs and domain-level graphs. Then, Section VIII studies the impact of these structural differences on the design of a crawling algorithm for the mobile web. Finally, we conclude in Section IX and discuss some future directions.

II. Background

We recall some definitions on directed graphs. The out-degree of a node is the number of edges originating from it, and the in-degree is the number of edges pointing to it). A strongly connected component is a set of nodes such that there exists a path between any pair of distinct nodes in the set. A directed graph may have many strongly connected components.

Broder *et al.* [1] proposed a bow-tie model to visualize the structure of the web. The bow-tie model (figure 1) consists largely of three major components: SCC, OUT, and IN. The SCC component consists of a set of nodes comprising the *largest* strongly connected component in the graph. The SCC is at the core of the bow-tie model. Some of the most prominent pages in the web graph would be part of the SCC. All the other components of the bow-tie model are defined with respect to the SCC. The OUT component refers to the set of nodes that satisfy two properties: (a) Each node in this set is reachable from at least one node in the SCC, and (b) no node in the SCC is reachable from any node in this set. IN is the set of nodes such that: (a) There is a path from each node in this set to at least one node in the SCC, and (b) no node in this set is reachable from any node in the SCC.

Apart from these three major components, the bow-tie model consists of three minor components: Tubes, Tendrils, and Disconnected components, all of which do not belong to SCC, IN or OUT. Tubes consist of nodes that are on a directed path from a node in IN to a node in OUT. Tendrils consist of nodes that are either reachable from nodes in the IN component or can reach nodes in the OUT component. The remaining nodes are not reachable from any node SCC, IN, and OUT, and belong to the Disconnected component. For simplicity, we use Tendrils in this paper to mean the union of the Tendrils and Tubes set.

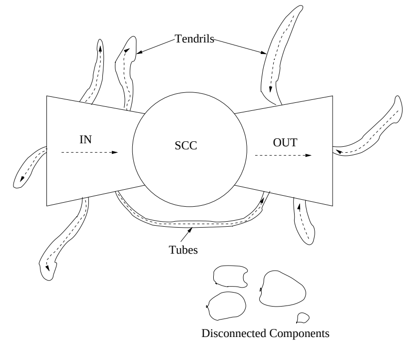


Fig. 1. Bow-tie structure of the web-graph proposed by Broder *et al.* [1].

III. Related Work

The structure of the fixed web graph has been been studied in detail [1], [5], [7], [8], [11]. These studies have looked at the distributions of the in-degree, out-degree, size of strongly connected components and Pagerank values and found them all to follow a power-law distribution[1]. Broder *et al.* [1] introduced the bow-tie model for the fixed web graph. Later studies [5], [8] have also used this model to characterize the structure of the fixed web graph. Also, the subgraphs of the fixed web graph formed by clusters of pages sharing a common trait were observed to retain the structural properties of the fixed web graph [5].

We are not aware of any work that statistically characterizes the properties of the mobile web graph.

IV. Methodology

This section describes the corpora we have used in this study. It also describes the processing steps used in deriving the graph properties.

A. Corpora used in the study

XHTML, WML and CHTML are the three predominant markup languages used to develop web pages for cell phones. CHTML was developed by NTT DoCoMo and is used predominantly in Japan, whereas XHTML and WML are dominant markup languages in the rest of the world. We have analyzed the the corpus of XHTML/WML pages separately from that for CHTML pages because of significant differences in their properties. We discuss these differences in more detail in section V. These differences stem from the highly evolved nature of mobile web in Japan compared to the rest of the world. Both these corpus are based on the Google's mobile web index of June 2007.

In this paper, we do not consider the situation where a mobile web page links to another mobile web page only via a fixed web page. In that case the two pages are considered

[1]Recall that a discrete random variable X has a distribution with a power-law tail if $Pr[X = x] \propto x^{-\alpha}$, for some constant $\alpha > 1$ and $x \geq x_{min}$, where α is known as the power-law coefficient.

to not be connected in the mobile web graph. We believe this case is somewhat uncommon, and studying this case further is outside the scope of the present paper.

To compare the characteristics of the mobile web with that of the fixed web, we have used two different crawls for the fixed web:

(i) WebBase 2001 [3]: a publicly available crawl done in 2001 by the WebBase project at Stanford and was used to study the structure of the fixed web [8]. This corpus is old and might not accurately capture the characteristic of today's web. However, we still use it as one of the references to compare the properties of mobile web graphs with because one hypothesis in the folklore is that today's mobile web can be regarded as a primitive version of the fixed web. Hence, the properties of a snapshot of the fixed web from the past will serve as an interesting reference point. Webbase 2001 is the earliest publicly available web corpus we could find. This corpus has been analyzed in detail in [8].

(ii) Fixed Web 2007: This corpus is based on the Google's web index of June 2007.

We analyze these web graphs at two different levels of granularity: page-level and domain-level. In the *page-level graph*, each node represents a web page, and an edge represents a hyperlink. In the *domain-level graph*, each node represents an entire web domain.

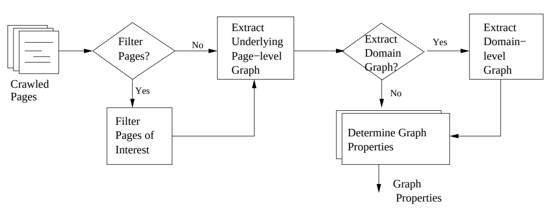


Fig. 2. Processing pipeline for extracting graph properties.

B. Processing Steps

Fig 2 shows the schematic diagram of the processing pipeline. We describe each of the component in this pipeline below.

- *Filter Pages of interest*: This step applies filters to the data set if any are defined, for example, when we restrict our study to pages in a specific language.
- *Extract Underlying Page-level Graph*: This involves parsing the web pages, extracting hyperlinks, and creating an adjacency matrix to represent the graph. Note that we retain only those edges which point to a page within the set of pages being considered. Any link to a page outside this set is simply dropped.
- *Extract Domain-level graph*: This step collapses all the nodes within a domain to a *super-node*. The super-node inherits all the links originating from and terminating at any nodes within the domain. Any duplicate link between a pair of super-nodes is removed.
- *Determine Graph Properties*: Determine average in- and out- degrees of nodes, in- and out-degree distributions, and the relative sizes of the different components in the bow-tie structure.

We have developed tools based on the large-scale distributed processing tool MapReduce to accomplish the first three steps: filtering pages, extracting the underlying page-level graph, and extracting the domain-level graph. We also used these to determine the average in- and out- degree of nodes in the graph. We used techniques described in [4] to compute the coefficient of the power-law distributions for in- and out-degree. We used the COSIN tools developed by Donato *et al.* [6] to obtain the relative sizes of the different components in the bow-tie structure.

Note that we exclude the Fixed Web 2007 corpus when comparing page-level graph properties because the COSIN tools [6] used to derive the relative sizes of the components of the bow-tie structure cannot handle the scale of this graph. Scaling and parallelizing COSIN tools to handle web graph with billions of nodes is beyond the scope of this paper.

V. Page-level graph properties

In this section, we present the characteristics of page-level web graph for mobile. We compare these properties with that for the fixed web.

A. Degree Distributions

Figure V-A shows the in-degree and out-degree distributions for both the XHTML+WML and CHTML corpora. Table I compares the average node degree and the coefficient of the power law distribution for both in-degree and out-degree distributions for both the corpora with the corresponding statistics for WebBase 2001 (the statistics for WebBase 2001 are taken from Donato *et al.* [7]). We make the following observations:

– The average node degree of XHTML/WML as well as CHTML corpus is considerably smaller than that for Webbase 2001. This implies that these two mobile-web graphs are significantly sparser. In general, we do expect the mobile web graphs to be significantly sparser than fixed web graph as mobile web pages tend to have fewer hyperlinks. However, it is interesting to see that the mobile web graph of 2007 is sparser than the fixed web of 2001.

– The in-degree and out-degree distributions for both the mobile-web graphs fall according to a power-law. This is similar to the behavior seen in the fixed web, and we believe that generally speaking similar underlying causes operate for the mobile web. However, note that the out-degree distributions fall off much more rapidly for both the mobile subgraphs than for the fixed web. This implies that there are fewer mobile pages with a large number of hyperlinks compared to pages in Webbase 2001, and is to be expected.

– The CHTML corpus has significantly higher average node degree compared to the XHML/WML corpus (5.06 versus 3.75), implying that the CHTML web graph has significantly more edges than the XHTML/WML web graph. Since CHTML pages are almost exclusively consumed in Japan, we hypothesize that this is due to the advanced state of the

Corpus	Avg node degree	Coefficient of power-law distribution	
		In-degree distribution	Out-degree distribution
XHTML/WML	3.75	2.00	3.49
CHTML	5.06	1.99	4.06
WebBase 2001	7.0	2.1	2.7

TABLE I
INDEGREE AND OUTDEGREE DISTRIBUTION PROPERTIES FOR BOTH THE MOBILE WEB CORPORA AND THE WEBBASE 2001 CORPORA.

Corpus	SCC	IN	OUT	Tendrils	Disconnected
XHTML/WML	10.5%	18%	10.4%	18.3%	42.7%
CHTML	22%	25.9%	14.2%	22%	15.8%
Webbase 2001	33%	11%	39%	13%	4%

TABLE II
RELATIVE SIZES OF THE VARIOUS COMPONENTS IN THE BOW-TIE STRUCTURE FOR BOTH THE MOBILE WEB CORPORA AND THE WEBBASE 2001 CORPORA.

mobile web in Japan; sophisticated mobile phones and faster cellular networks contribute to the creation of richer content by publishers.

B. Bow-tie structural properties

Table II presents the relative sizes of the various components of the bow-tie structure for XHTML/WML, CHTML, and WebBase 2001 corpus. In general, the mobile web graphs have larger Disconnected (and Tendril) components, and smaller SCC and OUT components as compared to the fixed web graph. This data indicates that the mobile web graphs are more disconnected than the fixed graph. They also have a larger IN component than the OUT component which is the reverse of what is observed for the fixed web corpus. This implies that the mobile subgraphs have more pages pointing to the central core (SCC) than are being pointed from the core.

In addition, we again observe that the characteristics of CHTML corpus are significantly different from that for the XHTML/WML corpus. CHTML corpus has larger SCC, larger Tendrils, and smaller Disconnected component. This is in line with our observation earlier that CHTML web graph is considerably more connected than XHTML/WML web graph.

VI. LANGUAGE PROPERTIES OF THE MOBILE WEB GRAPH

Dill *et al.* [5] have shown that the fixed web graph has the property of *self-similarity* — subsets of pages in the web graph that all share a common trait often retain the structural properties of the entire graph. To study if this property holds for the mobile web, we split the mobile web corpus based on each page's language, and examine the structural properties of the per-language corpus.

The distribution of the most common languages for both the corpora are shown in Table III. Unexpectedly, the top language in the XHTML/WML corpus is Chinese, and not English. In fact the number of Chinese web pages is almost twice that of English web pages. We speculate that this is due to the relatively high penetration of WAP in China caused by different factors like the cheap flat-rate WAP data plans, and the relative unavailability of access to desktop PCs [2]. In contrast, as expected, the top language for the CHTML subgraph is Japanese, with English being a distant second.

Corpus	Language	Fraction of nodes
XHTML/WML	Chinese	42.6%
	English	22.3%
	Russian	13.4%
	French	3.4%
	German	2.3%
CHTML	Japanese	92.3%
	English	5.9%

TABLE III
DISTRIBUTION OF TOP LANGUAGES.

We next present the relative sizes of the components of the bow-tie structure for the top three languages in the XHTML/WML corpus in Table IV. For comparison, we also present the relative sizes for the XHTML/WML corpus as a whole. We observe that the characteristics of language-specific graphs are very different from the aggregate characteristics of the entire XHTML/WML corpus. Of the three language considered, surprisingly the English language subgraph is most disconnected while the Russian language subgraph is the most connected even though it has fewer nodes than either English or Chinese. At present we do not have an clear explanation for this finding.

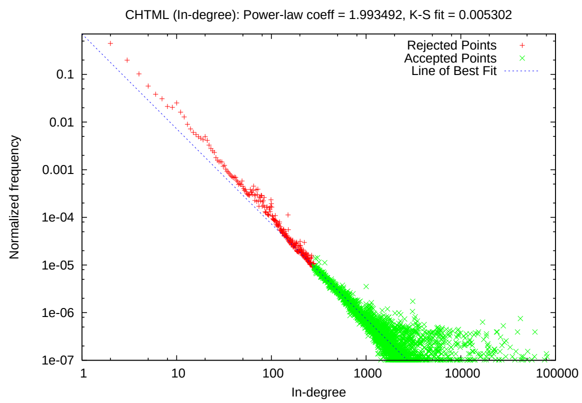
We did not perform a similar analysis for the CHTML Japanese subgraph since the CHTML graph is comprised almost entirely of Japanese pages.

VII. DOMAIN-LEVEL GRAPH PROPERTIES

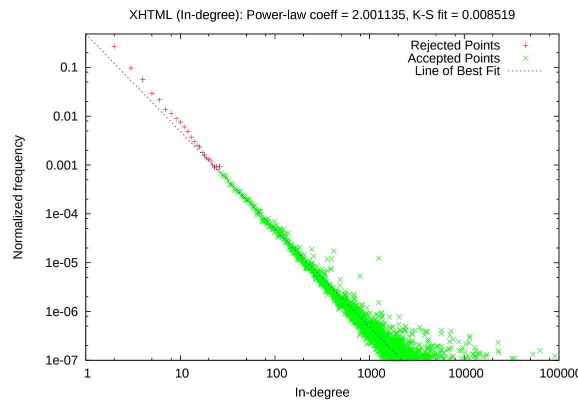
As described in section IV, we obtained domain-level graphs by collapsing all nodes for a domain into a single super-node. This allows us to understand the web-graph at a coarser granularity and extract domain-level features. We present the bow-tie structure properties of the domain level graphs for three corpora: XHTML/WML, CHTML, and Fixed Web 2007.

Table V summarizes the average node degree and the relative sizes of the various components of the bow-tie structure for the three domain level graphs. We make the following observations:

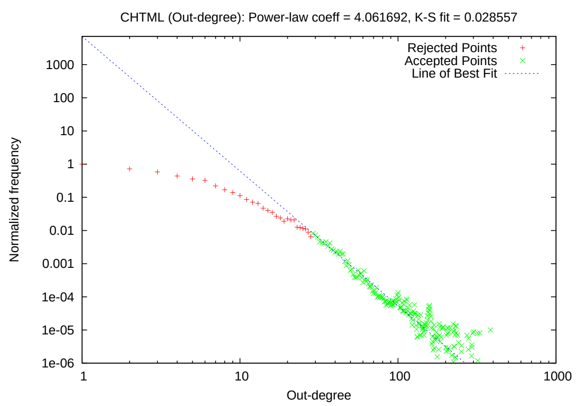
(i) The average node degree and the size of the SCC for the domain-level graphs is higher as compared to the corresponding page-level graphs. This implies that the domain-level graphs are better connected. This is



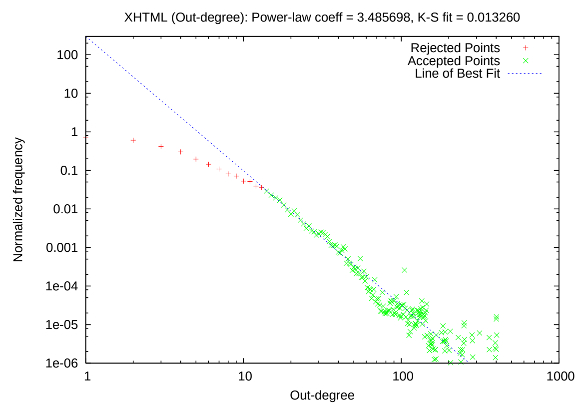
(a)



(b)



(c)



(d)

Fig. 3. Log-log plots of degree distributions. We also use the Kolmogorov-Smirnov goodness of fit test [4] to demonstrate how well the power law distribution approximates the *tail* of each degree distribution. The KS values lie in the interval $[0, 1]$; the larger the KS value, the worse the fit.

Corpus	SCC	IN	OUT	Tendrils	Disconnected
XHTML/WML	10.5%	18%	10.4%	18.3%	42.7%
Chinese	13%	22%	9%	14%	42%
English	2%	3%	7%	25%	63%
Russian	22%	40%	8%	11%	19%

TABLE IV
RELATIVE SIZES OF THE VARIOUS COMPONENTS IN THE BOW-TIE STRUCTURE OF LANGUAGE SUBGRAPHS.

Domain-level graph	Avg node degree	SCC	IN	OUT	Tendrils+Disconnected
XHTML+WML (domain-graph)	3.91	40.6%	40.7%	2.73%	15.9%
CHTML (domain-graph)	5.56	83%	16.4%	0.22%	0.36%
Fixed Web (Google 2007) (domain-graph)	35.75	93.9%	5.62%	0.4%	0.03%

TABLE V
AVERAGE NODE DEGREE AND THE RELATIVE SIZES OF THE COMPONENTS OF THE BOW-TIE STRUCTURE OF THE DOMAIN-LEVEL GRAPHS.

expected because super-node for a domain inherits all the in- and out- links of nodes in the domain.

(ii) As we observed with page-level graphs, the average node degree of the domain-level graph for the fixed web is much higher than that for both the XHTML/WML and CHTML corpus. Between the two, domain-level graph for CHTML has higher average node degree than that for XHTML/WML.

(iii) The XHTML/WML domain level graph has a much larger IN component as compared to the CHTML and the fixed web domain-level graphs. Recall from the definition that all domains that fall in the IN component do not have any in-links from any domain in the SCC. In addition we observe that the XHTML/WML domain-level graph has a significantly large Disconnected component, indicating that it is fairly disconnected even at the domain level.

(iv) The structural differences between domain-level mobile web graph 2001 and fixed web graph 2007 are similar to the differences between page-level mobile web graph 2007 and fixed web graph (WebBase 2001).

VIII. Application: Impact of structure on crawling algorithms

The structural differences between the mobile webgraph and the fixed webgraph raise some interesting issues for crawling algorithms. Since crawling is resource-intensive, tuning the crawl operation to be efficient is important. A crawling algorithm that is well-suited for exploiting the structural properties of the fixed web may not perform nearly as well on the mobile web, and vice versa. In particular:

– The degree of disconnectedness inherent in the mobile webgraph can have significant implications for crawl. The importance of a numerous and diverse set of seed URLs increases for the mobile web. Discovering, characterizing, selecting and maintaining such a seed set is also more important and more involved.

– Covering the IN component, in particular, requires more attention, for example by extensive and judicious selection of seed URLs in the IN component. This is especially the case for the CHTML subgraph.

– A depth-first search strategy for crawl may risk spending a disproportionate amount of resources exploring Tendrils or disconnected components. This is especially problematic if pages in SCC or OUT are more likely to be of interest to most users.

– If the amount of content in a given language is an indicator of user interest in content in that language, this implies that a crawling algorithm could be optimized for crawling sites in that language (say, XHTML/WML sites in Chinese) and, further, perform a deeper crawl for those sites. Of course, the danger is that this becomes a self-reinforcing cycle: more efficient and deeper crawl of such language-specific sites will result in more content in those languages. Thus any optimization of this nature needs to be applied carefully, for example by comparing not only the relative sizes of corpora but their relative growth.

– In general more care is required to crawl pages in languages where the graph is more disconnected. Conversely if a language is not of particular interest but the graph happens to be well connected (say, Russian in the XHTML/WML subgraph), the crawl depth could be reduced for that language to save crawl resources without significant adverse impact.

– The domain-level sparseness of the mobile web compared to the fixed web indicates that a mobile domain should be crawled as extensively as possible so as to discover as many of the domain-level interconnections that exist.

However, there are certain advantages for crawling the mobile web that the fixed web does not share. Since the link structure of the mobile web is generally much sparser, all other things being equal, a crawler is less likely to encounter pages that it has already visited, potentially increasing the efficiency of the crawl. By this same token, the XHTML/WML webgraph will enjoy this advantage to a greater extent than the CHTML webgraph.

IX. Conclusions and Future Work

We study the structure of the mobile web graph and show that it significantly differs from the fixed web graph. Specifically, the mobile web graph is sparser, more disconnected and is less concentrated in the SCC and OUT components and more concentrated in the Disconnected and Tendrils components as compared to the fixed web. We also find that the properties of the CHTML mobile web, which is mostly consumed in Japan, generally lie between those of the XHTML/WML mobile web and the fixed web. We also examine the language characteristics of the XHTML/WML mobile graph and observe a surprising preponderance of Chinese content. Unexpectedly, we also find the English language subgraph to be extremely disconnected as compared to the other top languages in the XHTML/WML subgraph. The structural characteristics of the mobile web graph can have significant implications for operations such as crawl, for example indicating that the crawling algorithms should use numerous and diverse seed URLs and should avoid using a depth first search.

This work is only a first step. Our results motivate the need for a deeper and more extensive analysis of the structure of the mobile web. Future work might include performing a more extensive study of the mobile web, proposing an alternative to the bow-tie model for the mobile web graph, better understanding the properties of the language subgraphs, and quantitatively characterizing the impact of the mobile web graph structure on the performance of different search algorithms.

Acknowledgements. We thank our colleagues Justin Wang, Ning Hu, and Yevgeniy Miretskiy for their help in our processing of data, and Arup Mukherjee, Matthieu Devin, and Vida Ha for invaluable discussions.

References

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, 33:309–320, June 2000.

[2] L. Chao. China mobile push pinches rivals, August 2007.

[3] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan, and G. Wesley. Stanford webbase components and applications. *ACM Trans. Inter. Tech.*, 6(2):153–186, May 2006.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, June 2007.

[5] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Inter. Tech.*, 2(3):205–223, 2002.

[6] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. A library of software tools for performing measures on large networks. 2004.

[7] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. The web as a graph: How far we are. *ACM Trans. Inter. Tech.*, 7(1):4, 2007.

[8] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. In *WebDB'05*, pages 145–150, 2005.

[9] Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, and Upfal. Stochastic models for the web graph. In *FOCS: IEEE Symposium on Foundations of Computer Science (FOCS)*, 2000.

[10] L. Laura, S. Leonardi, G. Caldarelli, and P. Rios. A multi-layer model for the web graph, 2002.

[11] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. In *Proceedings of COCOON'02*, 2002.