Inferring Link Behavior from the Connectivity Distributions of Web Pages

Kulwadee SOMBOONVIWAT † Masaru KITSUREGAWA ‡

† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

‡ Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan

E-mail: † kulwadee@tkl.iis.u-tokyo.ac.jp, ‡ kitsure@tkl.iis.u-tokyo.ac.jp

Abstract Since link-based measures are one of the most important factors in ranking mechanisms of today's web search engines, the intentions of creating hyperlinks need to be carefully investigated. This study analyzes the connectivity distributions of web pages and draws conclusions on the link behavior of web page authors. As a whole, the link behavior follows a ubiquitous power-law distribution. However, with different kinds of linkages (e.g. cross-site links vs. intra-site links) and within a specific domain, the connectivity distributions varies greatly. From a snapshot of the Thai Web, we extract connectivity distributions of different kinds of linkages and different subsets of web pages and try to infer some meaningful patterns of web links.

Keyword Web measurement, Web link behavior

1. Introduction

In recent years, link-based measures have become one of the most important factors in search results ranking mechanisms of today's web search engines. As a result, the intentions of creating hyperlinks need to be carefully investigated. This paper analyzes the connectivity distributions of web pages and draws conclusions on the link behavior of web page authors.

Mathematically, the Web can be represented as a graph whose nodes correspond to web pages and whose edges correspond to hyperlinks. Study of graphical properties and characteristics of the Web graph is not only theoretically challenging but also practically useful in the development of efficient algorithms for Web applications such as web crawling, web searching, and identification of web community.

The Webgraph shares various graphical properties with other kinds of complex networks e.g. a citation network, a power grid network, and so on. There are several previous works on the empirical studies of the Web graph. These studies consistently reported emerging properties of the Web graph at diffrent scales. One of the most notable emerging property is a *power-law connectivity* in which the number web pages having k number of connections decays polynomially as $k^{-\gamma}$, with $\gamma > 1$.

As a whole, the Web link behavior follows a ubiquitous power-law distribution. However, with different kinds of linkages (e.g. cross-site links vs. intra-site links) and within a specific domain, the connectivity distributions varies greatly. In this paper, we conduct statistical analysis on the Webgraph (and also the Hostgraph. See Section 2.1 for the formal definitions of a Webgraph and a Hostgraph) to extract connectivity distributions of different kinds of linkages and different subsets of web pages and try to infer meaningful patterns of web links.

Our dataset is a snapshot of the Thai Web collected by a web crawl on January 2007 which consists of around 550K crawled web pages. Based on this dataset, we create link databases for the corresponding Webgraph (5.7M nodes and 12M edges) and Hostgraph (1.2M nodes and 2.9M edges). Then, we measure the distributions of the connectivity per node for various kinds of nodes and edges. Based on the obtained statistical results, we will discuss and infer characteristics of Web link behaviors.

The rest of this paper is organized as follows. In the next section, we describe some basic graph terminologies and a power-law distribution. Section 3 describes the method used to create the Thai Web snapshot and some characteristics of the Thai Web snapshot, the Thai Webgraph, and the Thai Hostgraph. Section 4 presents the results of our web graph measurements. Section 5 reviews related works on the study of the Web as a graph. Finally, Section 6 concludes the paper and discuss about the directions for our future works.

2. Preliminaries

2.1. Graph Terminologies

A *directed graph* consists of a set of nodes and a set of ordered pairs of nodes, called edges. The *in-degree* of a node is the number of incoming edges incident to it. The out-degree of a node is the number of outgoing edges incident to it. A *Webgraph* [4] is a directed graph induced from a set of web pages where each node represents a web page and each directed edge represents a hyperlink from a source web page to a destination web page. A *Hostgraph* [13] is a weighted directed graph where each node represents a web host and each directed edge represents the hyperlinks from web pages on the source host to web pages on the destination host. The weight of the edge is equal to the number of such hyperlinks.

2.2. Power-law Distribution

A discrete power-law distribution is a distribution of the form $\Pr(X=k) = Ck^{-\gamma}$ for $k = 1,2,\ldots$, where γ is a coefficient (or a power-law exponent), X is a random variable and C is a constant. A power-law distribution can be checked by plotting the data in a log-log plot. The signature of the power-law distribution in a log-log plot is a line with slope determined by the coefficient γ .

The power-law distribution is ubiquitous, it has been observed in many complex networks such as social networks, transportation networks, biological networks, etc. [9]. Nevertheless, a notable characteristics of the power-law distribution is the fact that the power-law distribution decays polynomially for large values of independent variable x. As a result, in contrast to other standard distributions such as exponential and Gaussian, in a power-law distribution the average behavior is not the most typical.

3. Dataset

3.1. Thai Web Snapshot

Most studies on the properties of the Web of a country usually define the Web of a country as a set of web pages of all Web sites that are registered under the country top-level domain or that are hosted at an IP associated with that country. We argue that this definition is not appropriate for defining the Web of Thailand. Based on the language identification result of web pages in our Thai web dataset crawled on July 2004 by using a naïve breadth-first-search crawling in July 2004, we found that more than half of the web pages written in Thai language are web pages of Web sites registered outside ".th" top-level domain of Thailand (see Table 1).

Table 1: Language identification result classified by domain name (in number of pages). More than half of Thai-language web pages are belonging to Web sites registered outside Thailand's ccTLD (".th" domain).

Languages	Domains		Total
	".th"	Other	
Thai	591,683	1,131,088	1,722,771
Non-Thai	263,777	16,357,579	16,621,356
Total	855,460	17,488,667	18,344,127

Accordingly, if we crawl only web pages with the corresponding Thai top-level domain and/or the physically assigned location of the IP address of the Web sites then we will fail to collect a large portion of Thai-language web pages. Therefore, to increase the completeness of the Thai Web dataset, it is necessary to add to the definition of the Thai Web a criterion which is based on the language of a web page. Formally, we propose to use the following criteria to decide whether a web page is Thai.

(1) Top-level domain of the web page is ".th".
(2) IP address of its web server is physically assigned in "Thailand".
(3) Language of the web page is "Thai".

The first criterion can be implemented by adding a predicate function to check the value of the top-level domain of each URL before adding it into the URL queue of a crawler. For the second criterion, we need to check a geographical location of an IP address of each web server. The third criterion states that a web page should be included into the dataset if it is written in Thai regardless of its top-level domain. We achieved this by applying a language-specific web crawling method as proposed in [11,12]. In this study, we conducted experiments on a snapshot of Thai web crawled January 2007 (Jan2007 dataset). The start seed sets for the crawl consists of a number of popular websites and web portals in Thailand. The number of crawled web pages is 551,233 pages.

3.2. Thai Webgraph

A *Webgraph* [4] is a directed graph where each node represents a web page and each directed edge represents a hyperlink from a source web page to a destination web page. For each Thai web dataset, we have constructed a link database which provides access to inlink and outlink information of a web page corresponding to an input URL address. The numbers of vertices and directed edges of the Thai Webgraph induced from Jan2007 dataset are as shown in Table 2. It can be seen clearly from the Table that the resulting Thai Webgraph is sparse.

3.3. Thai Hostgraph

We define a *web host* as a set of pages sharing the server part of a canonical form of the URL For example, a server part of a URL http://www.asite.co.th:80/index.html is www.asite.co.th:80. Based on this definition, there are 1,214,457 hosts in the hostgraph generated from the Jan2007 dataset described in Section 3.1 (including both crawled and uncrawled nodes). The number of hosts under ".th" domain is 27,668. The average numbers of pages per host are 5 pages per host (all hosts) and 9 pages per host (only host under ".th" domain) respectively.

Figure 1 shows a log-log plot of the distribution of the number of pages per host. From the Figure, it can be seen that the distribution of host sizes is very skewed with the middle parts (number of pages = [20, 1000]) of the graph adheres to the power-law distribution. When the number of pages per host is between 199 and 204, we can see some anomalous outliers (i.e. those points deviating from the power-law) in the log-log plot. After manual examination, we found that those outliers are mostly corresponding to spam pages generated automatically by machine because they have very similar patterns of URL addresses and also very similar pattern of content.

A *Hostgraph* [13] is a weighted directed graph where each node represents a **web host** and each directed edge represents the hyperlinks from web pages on the source host to web pages on the destination host. The weight of the edge is equal to the number of such hyperlinks.

Based on the aboved definition of the hostgraph, we have constructed the Thai hostgraph from Jan2007 dataset. Table 3 shows properties of the resulting Thai hostgraph. According to Table 2 and Table 3, of all 5.7 million hyperlinks in the Thai Webgraph, about 60% of those hyperlinks are intra-host hyperlinks (i.e. hyperlinks between web pages on the same host). According to Table 3, our Thai hostgraph is sparse, it consists of 1.2 million nodes and 2.9 million directed edges. There are 27K nodes in the hostgraph corresponding to hosts under the ".th" domain.

In the following section, we will report various connectivity distributions which are extracted from the Thai Webgraph and the Thai Hostgraph described earlier. To summarize, the distributions we are going to present include: (1) degree distributions of the Thai Webgraph: all links, intra-site links (local links), and inter-site links (or remote links), (2) degree distributions of Thai Hostgraph (all nodes vs. nodes corresponding to hosts under ".th" domain name).

Table 2: Properties of Thai Webgraph

Vertices (crawled + uncrawled web pages)	5,785,349
Vertices (crawled web pages)	551,233
Directed edges (millions)	12

Table 3: Properties of Thai Hostgraph

Vertices (hosts)	1,214,457
Hosts in ".th" domain	27,668
Directed edges	2,904,632
Inter-host hyperlinks (sum of edge weights)	8,878,268
Percentage of intra-host hyperlinks	60.4%

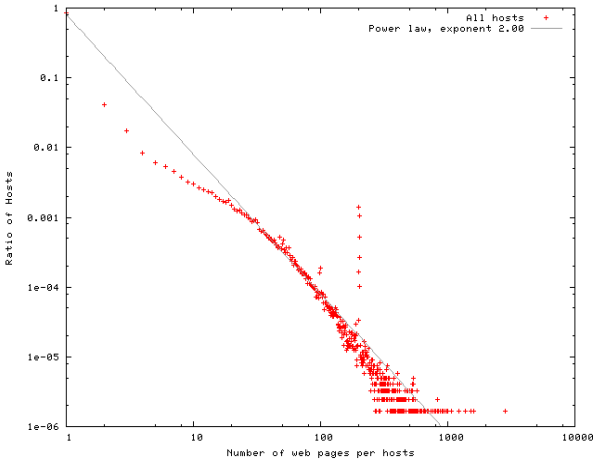

Figure 1: Distribution of the number of pages per host

4. Connectivity Distributions of Webgraph and Web Link Behavior

4.1. Webgraph connectivity

The distribution of the number of connectivity per node or degree distribution of many subgraphs of the Web has been consistently reported to follow a power-law distribution e.g. [4,5,6,7].

The power-law distribution can be described by "the Rich gets richer" phenomenon (or the preferential attachment) where new links are more likely to point to web pages that already have many links pointing to them. We have plotted the degree distribution of the Thai Webgraph. The in-degree distribution is shown in Figure 2, and the out-degree distribution is shown in Figure 3 respectively. Note that the values of the best-fit power-law exponents of all distribution plots are as shown in Table 4.

It can be seen from Figure 2 that the in-degree distributions of the Thai Webgraph in the log-log plots can be approximated by a straight line which is a signature of the power-law distribution. After examining

the web pages with large number of inlinks, we found that those web pages are homepages of popular Thai Web sites providing services such as free online-diary, blogs, and online communities. However, we also observed spam pages with very high number of inlinks.

In the case of the out-degree distribution in Figure 3, all log-log plots show approximately straight lines with concave in the first portion. By examining our crawled data, we found that most web pages with tremendously large number of outlinks are web pages from pornographic and spam sites. Note that, we observe anomalous bumps in both in-degree and out-degree distributions of the log-log plots in Figure 2 and Figure 3. Manual inspection reveals that most of the web pages corresponding to these anomalies are spam pages.

In Figure 2 and Figure 3, we have also separately plotted the degree distributions of “all links”, “local links only”, and “remote links only” of the inlinks and outlinks respectively. A local link means a hyperlink between web pages within the same website. Conversely, a remote link means a hyperlink between web pages residing in different websites.

According to Figure 2 and Figure 3, it can be seen clearly that the degree distributions of remote links are better fit with the power-law and contain a little number of anomalous bumps. This demonstrated that the anomalous bumps found in the degree distributions are largely caused by local links.

Another point that is worth mentioning is the absence of concavity part of the out-degree distribution in the plot of remote-only links. Obviously, the concavity in the out-degree distributions is caused by the characteristics of local linking. Consequently, while the process of hyperlinking between web sites is suitably described by “the Rich get richer'' model, another different model or a modified version of “the Rich get richer''' model is needed for explaining the phenomenon found in the link behavior corresponding to linking between web pages within the same web sites.

Table 5 shows average number of inlinks and outlinks. According to the table, the average number of the connectivity per node is 2 to 4 connections per node. As a result, the Thai Webgraph is a sparse graph with some densely connected regions which may be corresponding to homepages of some popular web sites, spam pages, or web pages with pornographic content.

Table 4: Values of the best-fit power-law exponents for the degree distributions of Thai Webgraph

Type of linkage	Direction	Power-law exponents
all links	in-degree	2.14
	out-degree	1.93
remote links	in-degree	2.06
	out-degree	1.96

Table 5: Average number of connections per page (for Jan2007 dataset)

Type of linkage	Direction	Average value
all links	in-degree	2.2
	out-degree	3.8
local links	in-degree	1.1
	out-degree	2.3
remote links	in-degree	1.1
	out-degree	1.5

4.2. Hostgraph connectivity

The distributions of the weighted in-degree and out-degree of web hosts in the Thai hostgraph exhibit the power-law distribution with the exponent of 1.85 and 1.44 respectively, as shown in Figure 4 and Figure 5. In Figure 6 and Figure 7, we plot the weighted in-degree and out-degree distributions of the hosts under “.th” domain respectively. While the weighted in-degree distribution of “.th” hosts indicates the power-law distribution with exponent 1.45, the weighted out-degree distribution of “.th” hosts does not fit with the power-law distribution (we obtained the power-law exponent whose value is less than 1 when we were trying to fit the log-log plot with the power-law). Among the web hosts with high weighted in-degree, we observe that in most cases they are Thai portals and news websites. The hosts with high weighted out-degree are in most cases the Web directories and blog service websites.

5. Related Work

Studies on the measurements of the statistical and topological properties of the Web graph have been conducted on various scales e.g. [1, 2, 4]. [1, 2] analyze the Web graph of the University of NotreDame (consisting of 325,729 nodes and 1,469,680 edges). The empirical results in [1] show that the distribution of links on the World Wide Web follows the power-law, with power-law exponent of 2.45 and 2.1 for the out-degree and the in-degree distribution respectively. [1] also predicts that an average distance between two randomly chosen web pages on the Web (i.e. the diameter of the Web graph) is equal to 19. [4] studies various graphical properties of the

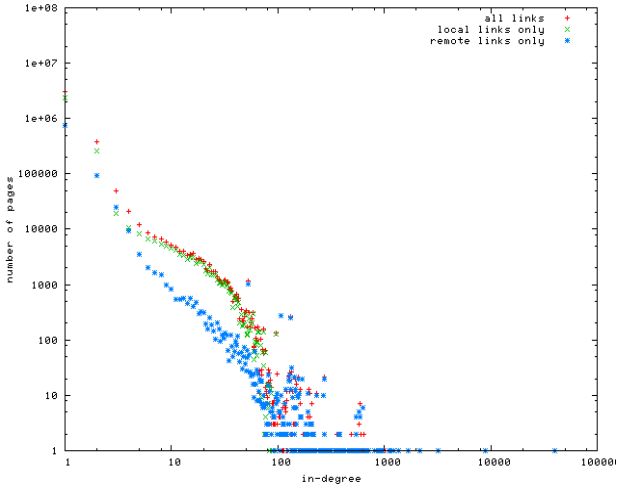

Figure 2: In-degree distributions of Thai Webgraph, plotted local and remote links separately

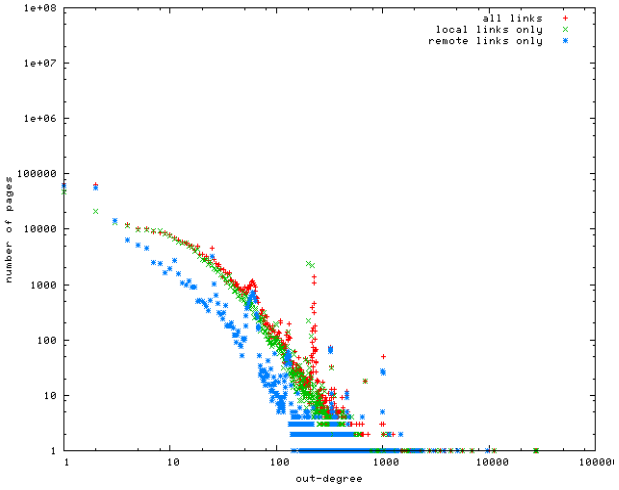

Figure 3: Out-degree distributions of Thai Webgraph, plotted local and remote links separately

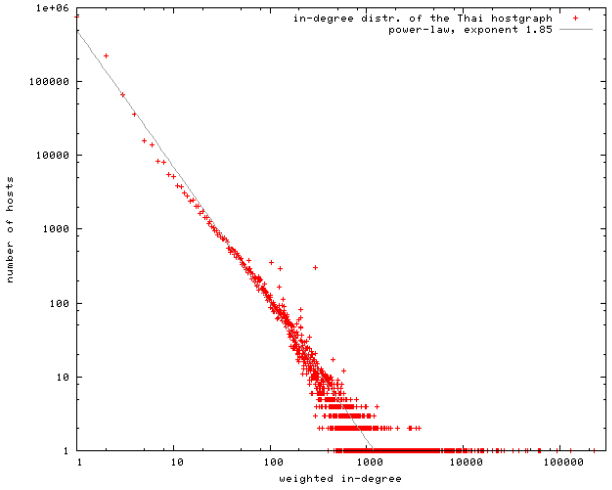

Figure 4: In-degree distribution of Thai Hostgraph

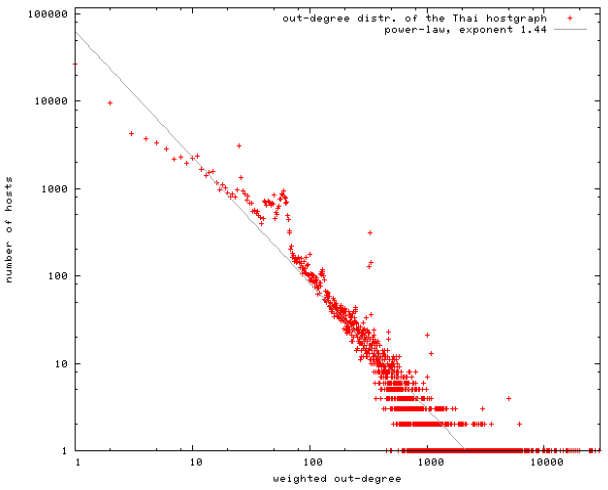

Figure 5: Out-degree distribution of Thai Hostgraph

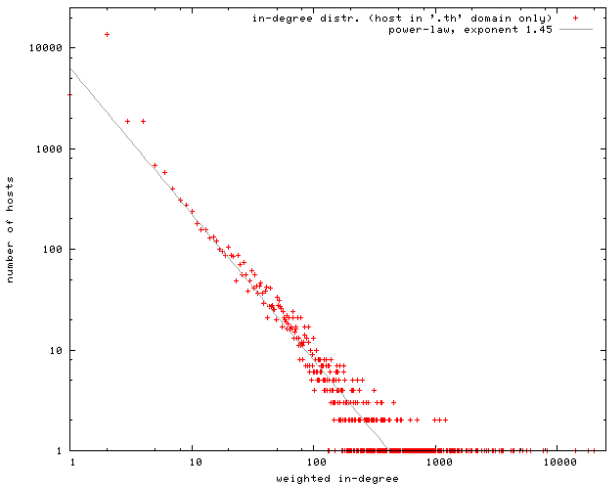

Figure 6: In-degree distribution of Thai Hostgraph, plotted only hosts in the ".th" domain name

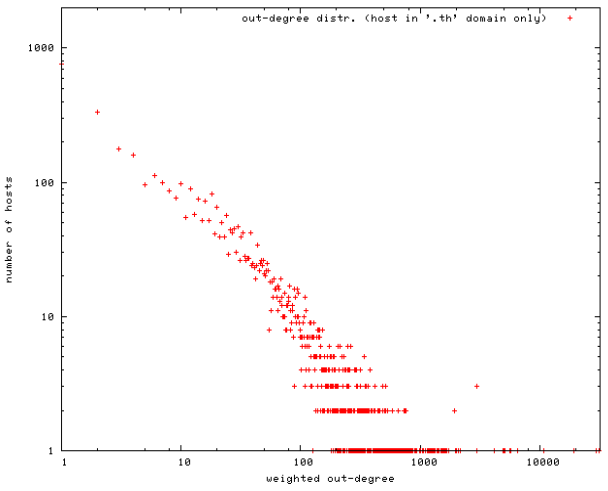

Figure 7: Out-degree distribution of Thai Hostgraph, plotted only hosts in the ".th" domain name

Web using two large datasets from AltaVista crawls (with more than 200 million nodes, and 1.5 billion links). [4] reports the power-law connectivity of the Web graph having exponent of 2.72 and 2.09 for the out-degree and the in-degree distribution respectively.

[4] also depicts the macroscopic structure of the Web graph as a bow-tie. The interpretation of the bow-tie structure provides a more accurate view of the Web structure. Remarkably, because there is a disconnected component in the bow-tie structure, it follows that the average and maximal diameter of the Web are infinite (as opposed to the value of 19 predicted in [1]). Nevertheless, by considering only those pairs of nodes that can be reached each other, [4] estimates that the maximal minimal diameter of the central core of the bow-tie is about 16, and shows that over 75% of the time there is no directed path between two randomly selected nodes.

Recently, [5] studies properties of the Web graph using a WebBase project crawl of 200 million pages and about 1.4 billion edges. (Stanford WebBase project homepage: http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase). They observed that the graphical properties of the WebBase sample (crawled in year 2001) are slightly different from the older sample studied in the prior works.

The notion of the hostgraph has been firstly proposed by [13]. According to [13], the hostgraph is a directed graph where each node represents a web host and each directed edge represents the hyperlinks from web pages on the source host to web pages on the target host. The weight of the edge is equal to the number of such hyperlinks. [13] also raises many convincing reasons for studying the hostgraph, and demonstrates its practicality.

6. Conclusions

In this paper we have extracted and analyzed various connectivity distributions of the Thai Webgraph and Thai Hostgraph. According to our results, we found that although the global connectivity distributions of the Thai Web subgraph fit with the pure power-law distributions, when including only some kinds of links or vertices the connectivity distributions varies greatly, for example the differences between local vs. remote in-degree and out-degree distributions.

Based on the derived statistics, we further investigate the characteristics of some web pages or hosts which are corresponding to some interesting spots found in the log-log plots of the connectivity distributions e.g. the anomalous bumps found in the out-degree distributions.

Our analysis results reveal some interesting link behaviors of web page authors in different situations e.g. link creation within a web site and across web sites. The understanding of these link behaviors is invaluable not only for algorithmic aspect but also sociological and cultural aspects emerging in the World Wide Web.

For the future work, we would like to investigate the connectivity distributions of other kinds of subgraphs such as a subgraph corresponding to some categories in a Web directory (e.g. Google directory, or the ODP).

References

[1] R. Albert, H. Jeong, and A. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[2] A. Barabsi and R. Albert. Emergence of scaling in random networks.*Science*, 286(5439):509–512, 1999.

[3] A. Barabsi, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web. *Science*, 287(5461):2115, 2000.

[4] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1–6):309–320, 2000.

[5] 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.

[6] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: measurements, models and methods. In Proc. of the 5th Annual Int'l Computing and Combinatorics Conference (COCOON'99), pages 1-18, 1999.

[7] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The web as a graph. In *Proc. of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'00)*, pages 1-10, 2000.

[8] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In Proc. of the 7th Int'l Workshop on the Web and Databases (WebDB'04), pages 1–6, 2004.

[9] P. Baldi, P. Frasconi, and P. Smyth. Modeling the Internet and the Web: Probabilistic Methods and Algorithms. John Wiley & Sons, Ltd., 2003.

[10] Thomas Mandl. Web Link Behavior and Consequences for Connectivity Based Authority Measures. WWW (Posters), 2003.

[11] T. Tamura, K. Somboonviwat, and M. Kitsuregawa. A method for language-specific web crawling and its evaluation. *Systems and Computers in Japan*, 38(2):10-20, 2007.

[12] K. Somboonviwat, T. Tamura, and M. Kitsuregawa. Finding thai web pages in foreign web spaces. In *ICDE Workshops*, page 135, 2006.

[13] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In Proc. of the 2001 IEEE Int'l Conf. on Data *Mining (ICDM'01)*, pages 51-58, 2001.