Structure of the Thai Web Graph

Kulwadee Somboonviwat
Dept. of Info. and Comm. Engineering
The University of Tokyo
kulwadee@tkl.iis.u-tokyo.ac.jp

Shinji Suzuki
Institute of Industrial Science
The University of Tokyo
suzuki@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
Institute of Industrial Science
The University of Tokyo
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract

This paper presents structural properties of the Thai Web graph. We conduct an empirical study on the Web graphs induced from two Thai web snapshots crawled during January 2007 (5.7M nodes and 12M directed edges) and May 2007 (18.8M nodes and 70M directed edges). From each Thai web snapshot, we extract statistical and structural properties of the associated graph including degree distribution, weakly and strongly connected components, and macroscopic structure. Our main findings are: (1) Although differing in scale and sampled time both samples of the Thai Web exhibit similar graphical properties, and (2) The macroscopic structure of the Thai Web differs from other Web subgraphs studied earlier.

1. Introduction

The Web can be viewed as a very large, dynamic graph whose nodes correspond to web pages and whose edges correspond to hyperlinks. Insights into structural properties of the Web graph is beneficial to the development of efficient algorithms and Web applications. For example, link-based web documents ranking algorithms such as PageRank [9] and HITS [17] are based on the results from the study of the Web as a graph.

The objective of this paper is to study structural properties of the Thai Web graph. We define the Thai Web as a set of web pages related to Thailand, and the Thai Web graph is a graph induced from this set of Thai web pages. In recent years, the characteristics of the national Webs of many different countries have been investigated e.g. [3, 8, 16, 21, 23]. The results of these studies reveal both the characteristics peculiar to the national Webs and the characteristics that are consistent with the Web in general. Examples of applications and algorithms that can take advantages from the understanding of the Web graph of a country are: (1) language-specific web resource discovery [24, 25], (2) topic-specific web resource discovery [11, 12, 22], (3) study of the Web communities of a country, and (4) development of tools for sociological and marketing researches.

The first challenge in the study of a national Web is how to decide whether a web page belongs to a specific country or not. Most previous studies on the national Webs use country-code top-level domains (ccTLDs) and/or IP addresses to check the nationality and the physical location of a web page. Based on these two criteria, a web page will belong to a country if (1) the domain part of its URL matches the country-code of that country e.g. '.th' for Thailand, or (2) its IP address is located inside that country. Nevertheless, according to the observations found in [24, 25], many web pages written in the Thai language are residing outside the '.th' top-level domain of Thailand. Therefore, for the purpose of studying the Web of Thailand, deciding whether a web page belongs to Thailand using these two criteria is inadequate because it would result in low coverage of the Thai Web.

To alleviate the above problem, we need an efficient method for crawling Thai web pages outside '.th' domain. Because the Thai language is unique to Thailand, one possible solution is to employ a language-specific web crawling method proposed in [24, 25]. Therefore, in our work, we will decide whether a web page is related to Thailand and should be included in the sample of the Thai Web using three properties of the web page i.e. (1) top-level domain name, (2) geographical location corresponding to the IP address, and (3) language.

In this paper we investigate the structural properties of the Thai Web using two Thai web snapshots, crawled on January 2007 (550K web pages) and May 2007 (1.4M web pages) respectively. Based on the Webgraphs induced from these two datasets, we extract graphical properties of the Thai Web including degree distribution, weakly and strongly connected components, and large-scale link structure. The results confirm the existence of power-law degree distributions in the Thai-related portion of the Web and reveal asymmetric bow-tie structure of the Thai Web.

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 reviews related literatures on the study of the global and national Web graphs. Section 4 explains the method used in collecting our Thai web snapshots. Then, in Section 5, we present the derived graphical and structural properties of the Thai Webgraph. Finally, Section 6 concludes the paper.

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 *Webgraph* is defined as a directed graph induced from a set of hyperlinked web pages. 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 *path* from node u to node v is a sequence of nodes such that from each of its node there exists an edge to the next node in the sequence.

A *connected component* of an undirected graph is a set of nodes such that for any pair of nodes u and v, there exists a path from u to v. A *strongly connected component (SCC)* of a directed graph G is a set of nodes S such that for every pair of nodes $u, v \in S$, there exists a *path* from u to v and from v to u. A *weakly connected component (WCC)* of a directed graph G is a set of nodes W where W is a connected component of the undirected graph obtained by ignoring the directions of all edges in G.

2.2 Power-law Distribution

A discrete power-law distribution is a distribution of the form $Pr(X = k) = Ck^{-\gamma}$ for $k = 1, 2, ...$ 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 the World Wide Web, social networks, transportation networks, biological networks, and so on [4]. 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 e.g. exponential and Gaussian, in a power-law distribution the average behavior is not the most typical.

3. 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, 5, 10]. [1, 5] analyze the Web graph of the University of NotreDame (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.

[10] studies various graphical properties of the Web using two large datasets from AltaVista crawls (with more than 200 million nodes, and 1.5 billion links). [10] 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. [10] 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, [10] 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.

[7] studies the Web graph at a higher abstract level i.e. at a website level. [7] analyzes linkage between web hosts using the hostgraphs, and finds that the number of connections per node in the hostgraphs also obeys the power-law distribution. [13] studies the bow-tie structure at a more fine-grained level and describes the organization and connection between topic-specific bow-tie structures in the Web graph.

Recently, [14] studies properties of the Web graph using a WebBase project crawl of 200M pages and about 1.4 billion edges [Stanford WebBase project http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/]. [14] observes 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 study of a Web graph of a country has been done by several countries such as [3, 8, 16, 21]. These studies reveal

many interesting characteristics of the Web subgraphs pertaining to specific countries. Here, we will give a brief summary of some selected studies on the national Web graphs.

African.[8] crawled the African Web and analyzed its content, link, and interconnection between domains of countries in the African Web. The reported power-law exponent of in-degree distribution is 1.92. The macroscopic structure of the African Web consists of a single giant SCC pointing to several smaller SCCs, which is different from the bow-tie structure of the global Web.

Spain. [3] studied the characteristics of the Web of Spain in terms of content, link, and technology usage at three levels i.e. web pages, sites, and domains. The reported power-law exponent for the in-degree and out-degree distributions are 2.11 and 2.84 respectively. [3] depicted link structure among web sites in the Web of Spain using the extended version of the bow-tie structure defined in [2].

China. [21] measured many properties and evolution of the China Web graph. The reported power-law exponent for the in-degree and out-degree distributions of the China Web graph are 2.05 and 2.62 respectively. The graph structure of the China Web differs from the global Web reported in [10]. While the bow-tie of the global Web has a MAIN SCC component containing approximately 1/4 of total web pages, the bow-tie of the China Web has a very large MAIN SCC component containing roughly 4/5 of total web pages.

Korea. [16] reported the power-law distribution of the number of connectivities per node for the Korea Web, the power-law exponents for in-degree and out-degree distributions are 2.2 and 2.8 respectively. Like the China Web, they also observed a bow-tie structure with a large MAIN SCC.

Regarding the study of the Thai Web, [23] presents quantitative measurements and analyses of various properties of web servers and web pages of Thailand. Their dataset consists of 700K web pages downloaded from over 8,000 web servers registered under '.th' domain on March 20, 2000. Their analyses focuses on the content and technology usage of the Thai Web, there is no link analysis result given in [23]. On the other hand, in this paper we report link analysis results using two recent snapshots of the Thai Web.

4. The Thai Web Datasets

As mentioned earlier in Section 1, the criteria used to decide whether a web page is Thai and should be included in a Thai Web dataset are as follows.

(1) Top-level domain of a web page is '.th'.

(2) IP address of its web server is assigned in 'Thailand'.

(3) Language of a web page is 'Thai'.

The first criterion can be implemented by adding a predicate function to check the top-level domain name 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. We implemented the second criterion by using Geolite country API (Geolite country API and data available from http://www.maxmind.com/) to check the country assigned to the IP address.

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 using a language-specific web crawling method proposed in [24, 25]. In [24], we analyzed characteristics of links between Thai web pages and proposed a set of page-level language specific crawling strategies, while [25] extended the study of link characteristics to the level of web servers and proposed a set of server-level language specific crawling strategies. We implemented the third criterion by using a combination of aforementioned page-level and server-level language specific web crawling strategies.

Incorporating the implementations of the three criteria as discussed above into our crawler, we downloaded static html pages from the Thai Web by starting from a set of popular web portals in Thailand on January 2007 and May 2007. For Jan2007 dataset, there are 551,233 crawled pages (261,235 pages are under '.th' domain). For May2007 dataset, there are 1,402,206 crawled pages (609,028 pages are under '.th' domain).

Table 1. Number of nodes and directed edges of the Thai Web Graphs

	Jan2007	May2007
number of nodes (crawled+uncrawled)	5,785,349	18,864,382
number of nodes (crawled)	551,233	1,402,206
number of edges	12M	70M

5. The Thai Web Graph

We will first describe the characteristics of the Thai Web graph derived from web pages in the Jan2007 and May2007 datasets. Then, in the following sections, we will present the study results on graphical properties and macroscopic structure of the Thai Web graph.

For Jan2007 dataset, the graph consists of 5,785,349 nodes, 5,234,116 of which correspond to uncrawled pages. The number of directed edges is approximately 12 million links, excluding duplicates and loops. For May2007 dataset, the graph consists of 18,864,382 nodes, 17,462,176 of which correspond to uncrawled pages. The number of directed edges is approximately 70 million links, excluding duplicates and loops.

Table 1 summarizes basic statistics of these two Thai Web graphs.

5.1. Power-law Connectivity

Many studies have consistently reported that the distribution of the number of connectivities per node (i.e. in-degree and out-degree distribution) of the Web graph, in several different scales, follows a power-law distribution [3, 10, 16, 21]. In [10], the exponent of the in-degree and out-degree distributions of the global Web graphs derived from two AltaVista crawls in 1999 were reported to be 2.09 and 2.72 respectively.

Figure 1 (a) and (b) shows the in-degree and out-degree distribution of Jan2007 and May2007 datasets respectively. In each plot, we show the power-law degree distributions in three cases: all links, remote-only links, and internal-only links.

The in-degree distributions for both datasets exhibit power-law distributions with approximate power-law exponents of 2.14 (Jan2007) and 2.27 (May2007) respectively. In the case of out-degree distributions, the Thai Web graph also exhibit power-law distributions with approximate power-law exponents of 1.93 (Jan2007) and 2.27 (May2007). However, the first portion (for $x < 100$) of the out-degree distributions deviate from the signature line of the power-law distribution. This deviation was also observed in [10]. This suggests that the initial portion of the out-degree distribution another distribution.

For all plots in Figure 1, we can observe a number of anomalous bumps, which are groups of pages sharing the same number of inlinks or outlinks. For our Thai Web graph, these anomalous bumps appear after the number of in/out-links is greater than 100. These anomalous bumps have also been observed in [3, 10]. One of the possible explanations for these bumps is the effect of spam pages [15, 26]. We have tried to manually observe some web pages corresponding to these anomalies. According to our observation, we found that most of them have very similar patterns of URLs and contents. Therefore, these anomalies may be one of the signatures of spam pages generated automatically by machines.

For the remote-only and internal-only cases in each plot, it can be seen clearly that the internal links correspond to the anomalies discussed earlier. Comparing the remote-only and internal-only cases, we also observe that (1) most links in web pages of our datasets are internal links i.e. links between web pages within the same web server. (2) the distribution in the remote-only case is better fit with the power-law. These different characteristics between internal and remote links are significant. This suggests that an algorithm which relies on properties of hyperlinks between web pages should manipulate internal and remote links differently.

5.2. Bow-tie Structure

To understand the topology of the Thai Web graph, we conducted a series of following experiments on the Web graph induced from crawled nodes of Jan2007 and May2007 Thai web snapshots. The description of the experiments can be found in [10].

(1) Weakly connected component (WCC).

(2) Strongly connected component (SCC).

(3) Random-start BFS.

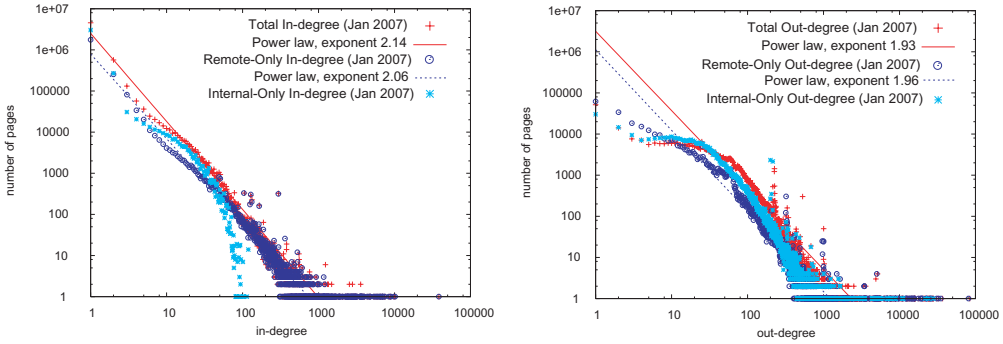
According to [10], the bow-tie structure consists of five components. (1) SCC: consists of web pages in the largest strongly connected component in the graph. (2) IN: consists of web pages that can reach the SCC but cannot be reached from the SCC. (3) OUT: consists of web pages that can be reached from the SCC but cannot reach any pages in SCC. (4) TENDRILS: consists of web pages that can be reached from IN and those that can only reach to OUT. (5) DISCONNECTED: consists of web pages outside the largest weakly connected component in the graph.

The size of each component in the bow-tie structure of Jan2007 and May2007 Thai Web snapshots is as shown in Table 2. We depict the bow-tie of each dataset in Figure 2. From these results, It is clear that the bow-tie structure of the Thai Web is asymmetric, with a large OUT component. In Table 2, we also show the size of the components in the bow-ties of other Web subgraphs. From the table, it can be seen the different characteristics of the bow-tie from each sub-region of the Web graph. For example, while the China and the Korea bow-tie webs have a very large central SCC and small augmented IN and OUT, the SCC of the Thai Web bow-tie is rather small and the size of OUT is quite larger than the IN region. It is interesting to investigate into each region of the bow-tie to find out the answers for the questions such as (1) What kinds of web pages belong to each region? (2) Is there any relationships between the topology of bow-tie structure and the ecology and/or sociology of Internet usage in each country? We plan to conduct these analyses in our future work.

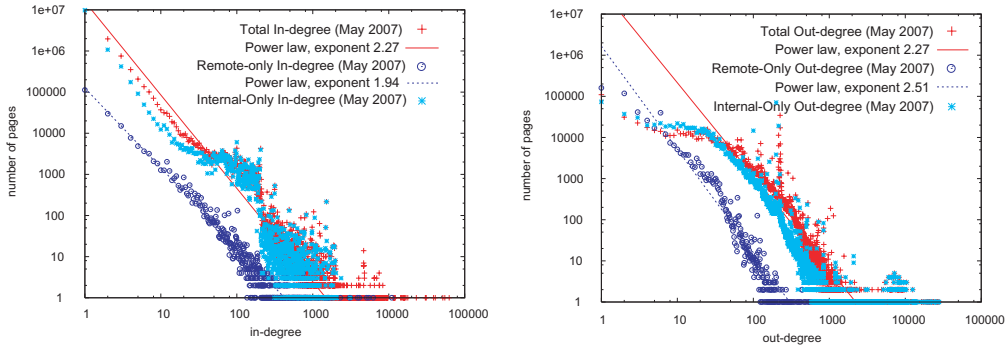
6. Conclusion and Future Works

In this paper, we have analyzed graphical properties and linkage structure of two Thai web snapshots crawled in January 2007 and May 2007. Based on the Webgraphs constructed from each dataset, we have obtained in- and out-degree distributions, and the large-scale link structure of the Thai Web.

Our analyses results confirm the emergence of the power-law distribution in the Thai-related region of the Web. By plotting the degree distributions of the internal and the external links separately, we have revealed and confirmed statistical differences between these two kinds of



(a) Jan2007 dataset



(b) May2007 dataset

Figure 1. Power-law degree distribution of the Thai Webgraphs

links. We also revealed the large-scale link structure of the Thai Webgraph. We found that the structure of the Thai Web can be depicted as the bow-tie. Nevertheless, unlike other subgraphs of the Web, the bow-tie of the Thai Web is asymmetric i.e. it has a large OUT component.

For our future works, we plan to investigate more about the degree distributions of different kinds of links such as links between specific domain names. We are also interested in studying of the evolution of the large-scale link structure of the Thai Web.

References

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

[2] R. Baeza-Yates and C. Castillo. Relating web characteristics with link based web page ranking. In *Proc. of the 8th Int'l Symposium on String Processing and Information Retrieval (SPIRE'01)*, pages 21–32, 2001.

[3] R. Baeza-Yates, C. Castillo, and V. Lopez. Characteristics of the web of spain. *International Journal of Scientometrics, Informetrics and Bibliometrics*, 9(1), 2005.

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

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

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

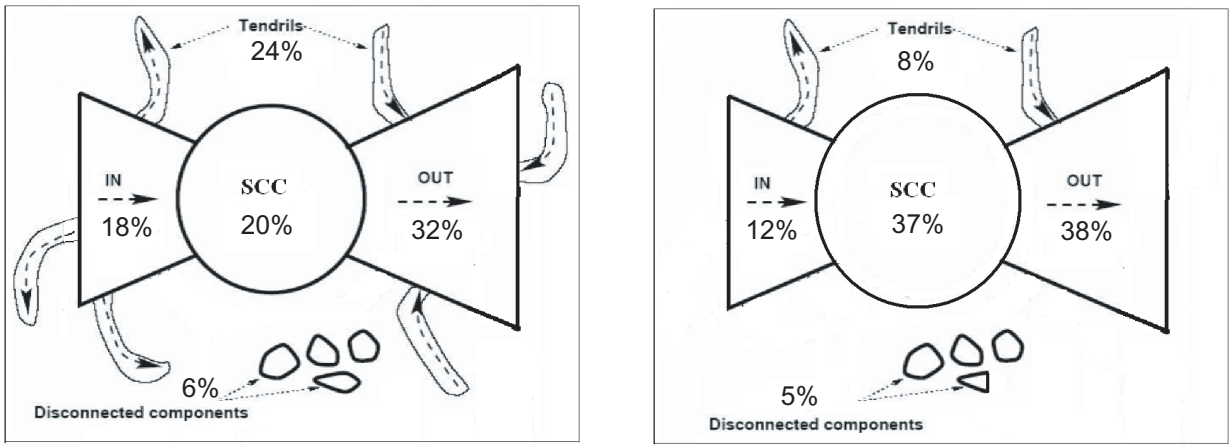
[7] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proc. of the 2001 IEEE Int'l Conf. on Data Mining (ICDM'01)*, pages 51–58, 2001.

[8] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the african web. In *Poster Proc. of the 11th Int'l Conf. on World Wide Web (WWW'02)*, 2002.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the 7th Int'l Conf. on World Wide Web (WWW'98)*, pages 107–117, 1998.

[10] 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.

[11] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. of the 8th Int'l Conf. on World Wide Web (WWW '99)*, pages 1623–1640, 1999.



(a) Jan2007 dataset

(b) May2007 dataset

Figure 2. bow-tie structure of the Thai Web

Table 2. Size of components in the bow-ties of different Web subgraphs

dataset	SCC	IN	OUT	TEN.	DIS.
Altavista [10] (1999)	28%	21%	21%	22%	9%
WebBase [14] (2001)	33%	11%	39%	13%	4%
China [21] (2005)	80%	12%	6%	1%	1%
Korea [16] (2007)	85%	8%	5%	1%	1%
Thailand (Jan2007)	20%	18%	32%	24%	6%
Thailand (May2007)	37%	12%	38%	8%	5%

[12] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *Proc. of the 7th Int'l Conf. on World Wide Web (WWW'98)*, pages 161–172, 1998.

[13] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. In *Proc. of 27th Int'l Conf. on Very Large Data Bases (VLDB'01)*, pages 69–78, 2001.

[14] 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.

[15] 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.

[16] I. K. Han, S. H. Lee, and S. Lee. Graph structure of the korea web. In *Proc. of the 12th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'07)*, pages 930–935, 2007.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[18] 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.

[19] 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.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proc. of the 8th Int'l Conf. on World Wide Web (WWW'99)*, pages 1481–1493, 1999.

[21] G. Liu, Y. Yu, J. Han, and G.-R. Xue. China web graph measurements and evolution. In *Proc. of the 7th Asia Pacific Web Conference (APWeb'05)*, pages 668–679, 2005.

[22] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Inter. Tech.*, 4(4):378–419, 2004.

[23] S. Sanguanpong, P. Piamsa-nga, Y. Poovarawan, and S. Warangrit. Measuring and analysis of the thai world wide web. In *Proc. of the Asia Pacific Advance Network conference*, pages 225–330, 2000.

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

[25] 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.

[26] M. Thelwall and D. Wilkinson. Graph structure in three national academic webs: power laws with anomalies. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):706–712, 2003.