Connectivity of the Thai Web Graph

Kulwadee Somboonviwat[1], Shinji Suzuki[2], and Masaru Kitsuregawa[2]

[1] Graduate School of Information Science and Technology,
The University of Tokyo
7–3–1 Hongo, Bunkyo–ku, Tokyo, 113–0033, Japan
[2] Institute of Industrial Science, The University of Tokyo,
4–6–1 Komaba, Meguro–ku, Tokyo 153–8505, Japan
{kulwadee, suzuki, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract. The study of a national Web graph is challenging and can provide insight into social phenomena specific to a country. However, because there is no country border in the Web, deciding whether a web page belongs to that country or not is difficult. In this paper we aim at studying the characteristics of the Thai Web graph. We first address the challenge of gathering Thailand-related web pages from the borderless Web by proposing a set of criteria for defining Thailand-related web pages. Three Thai web snapshots have been collected during July 2004 (18M web pages), January 2007 (550K web pages), and May 2007 (1.4M web pages) respectively. We then analyze and report various statistical properties related to connectivity of the associated Thai Web graphs.

1 Introduction

The World Wide Web consists of huge amount of interconnected web pages. 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 [11, 12], web searching [9, 17], and web community discovery [18].

This paper studies graphical properties of the Thai Web. We define the Thai Web as a set of web pages related to a country, Thailand. As a subgraph of the global Web, the Thai Web graph is a graph induced from these Thailand related web pages. Study of the Web of a country (or the national Web) has been conducted at many diffrent scales by various countries e.g. African [8], China [19], Korea [16], Spain [3], and Thailand [21]. While confirming many phenomena already observed in the global Web graph (e.g. 'small world, power-law degree distribution, and bow-tie structure phenomena as reported in [1, 10]), statistics of the national Web graphs also reveal many characteristics that are specific to each national Web graph. The characteristics of these local national Web graphs have emerged as a result of many individual factors peculiar to each country such as penetration of the internet, internet usage behavior, social values, education levels, language, and culture etc. Thus, the study of a national Web

graph of a country poses not only intriguing mathematical problems but can also potentially provides insight into society, culture, and economic of a country. In addition, many link-based algorithms and web applications can also be improved by exploiting link information specific to a country. Examples of such applications include language-specific resource discovery [22, 23], topic-specific resource discovery [11, 12, 20], and localized web search.

The first challenge in the study of the Web of a country is how to decide whether a web page belongs to that country or not. The most commonly used criteria for deciding the nationality of a web page are country-code top-level domains (ccTLDs) and physical location of the web server containing that web page as determined by an IP address of the server [3, 8, 21]. Based on these two criteria, a web page will be assigned as belonging to a country if (1) the domain part of its URL matches the country-code of that country e.g. '.th' for Thailand and '.jp' for Japan, or (2) the IP address of the web server containing the web page is physically assigned to a geographical location in that country.

However, according to the statistics of the Thai Web observed in [22, 23], there are many Thai web servers registered under the international domain names especailly '.com' and '.net' domains, and/or physically located outside Thailand. As a result, deciding if a web page belongs to Thailand based only on the two aforementioned criteria will result in low coverage of the Thai Web and thus is inappropriate for the purpose of studying the Web of Thailand. In order to obtain a higher coverage of the Thai Web, we need to address the problem of how to effectively collect Thai web pages outside the '.th' domain name. One way to do this is to make use of a characteristics unique to Thai web pages i.e. the Thai language, an official language of Thailand (Thai language is used in the schools, the media, and government affairs in Thailand). A web page will be assigned as belonging to the Thai Web if it contains some information written in the Thai language. We propose that in order to obtain a more complete snapshot of the Thai Web, it is necessary to gather not only web pages inside '.th' domain but also those web pages written in Thai language outside '.th' domain.

In this paper, we focus our study on the connectivity of the Thai Web. Statistics about degree distribution and connected components were extracted from three Thai web snapshots. The three Thai web snapshots were crawled using the language specific web crawling method proposed in [22, 23], coupled with the checking of ccTLDs and geographical locations of the IP addresses. Thai web crawls were conducted during July 2004 (18M web pages crawled), January 2007 (550K web pages crawled), and May 2007 (1.4M web pages crawled) respectively. For each dataset, we conducted experiments to measure the following properties of the Thai Web graph:(1) in-degree and out-degree distributions, (2) weakly and strongly connected components (SCC and WCC),(3) the macroscopic structure, and (4) connectivity between domains under the Thailand national domain. From the experimental results, we observed the ubiquitous of power-law in the Thai Web graph. Furthermore, anomalies (or outliers) in the log-log plots of the degree distributions were also observed and a major root of these anomalies has also been uncovered. The analysis of the connected components in the Thai Web

graph reveals an asymmetric bow-tie large-scale structure of the Thai Web. And, the statistics of connectivity between domain names indicate that linkage in the Web is, to a certain extent, reflecting the relationship between organizations in the real world.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 gives the definition of the Thai Web and describes characteristics of the Thai Web datasets. Section 4 presents our experimental results. Section 5 concludes the paper with directions for further work.

2 Related Work

The information available on the Internet is associated together by hyperlinks between pairs of web pages, forming a large graph of hyperlinks or the Web graph. The Web graph shares various graphical properties with other kinds of complex networks e.g. a citation network ,a power grid network, etc. 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 $\gamma > 1$. These results have remarkable impact in graph theory, and the design of efficient algorithms for applications such as web crawling, web search, and web mining. A power-law connectivity were observed in various scales of the Web graphs [1, 5, 10]. The study results in [1] show that the distribution of links on the World Wide Web follows the power-law, with power-law exponent of 2.1 and 2.45 for the in-degree and the out-degree distribution respectively. [10] reported on the power-law connectivity of the Web graph having exponent of 2.09 and 2.72 for the in-degree and the out-degree distribution respectively.

Another emerging property of the Web graph is its bow-tie structure. [10] conducted a series of experiments including analyses on WCC, SCC, and random-start BFS using large-scale Web graphs induced from two AltaVista crawls. Based on the results of these experiments, [10] inferred and depicted macroscopic structure of the Web graph as a "bow-tie" in which more than 90% of nodes reside in the largest weakly connected component. Because there is a disconnected component in the bow-tie structure, it is clear that there are a sizable portions of the Web that cannot be reached from other portions. Interestingly, the sizes of the WCC and SCC in this study also follow a power-law distribtuion. [14] studied the bow-tie structure in subgraphs of the Web graph associated with a particular topic.

The study of a Web graph of a country has been done by several countries such as [3, 8, 16, 19, 21].

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 many small SCCs.

China. [19] 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 bow-tie structure of the China Web has a very large MAIN SCC component which consists of approximately 4/5 of the total number of web pages in the China Web graph.

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 graph, The bow-tie structure of the Korea Web has a very large MAIN SCC component.

Spain. [3] comprehensively analyzed 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 notion of the bow-tie structure, proposed in [2].

Thailand. [21] conducted a study of the Thai Web (i.e. the Web of Thailand). [21] presented 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 2000. The study in [21] presents several statistics regarding to the content of the Thai Web. There is no statistics about the structure and characteristics of the Thai Web graph given in [21].

3 Definition of the Thai Web and the Thai Web Datasets

As stated earlier in the first section, 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 Jul2004 dataset which was obtained by breadth-first-search crawling in July 2004, we found that more than half of the web pages written in the Thai language are web pages of Web sites registered outside '.th' top-level domain of Thailand (see Table 1).

Consequently, 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 easily implemented by adding a predicate function to check the value of the top-level domain of each URL before adding it into the

Table 1. Language identification result of the Jul2004 dataset categorized by domain (in number of pages). More than half of the Thai-language web pages belong to Web sites registered outside Thailand's country top-level domain (i.e. '.th' domain).

Languages	'.th' domain	other domains	Total
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

Table 2. Number of vertices and directed edges of the Thai Web Graphs

	Jul2004	Jan2007	May2007
number of vertices	39,078,795	5,785,349	18,864,382
number of directed edges	123M	12M	70M

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 using a language-specific web crawling method proposed in [22, 23].

In this study, we conducted experiments on three snapshots of Thai web crawls. The first snapshot (Jul2004 dataset) was crawled in July 2004. For the first snapshot, we used a naive BFS web crawling strategy to get a sample of the Thai Web. The second and the third snapshots (Jan2007 and May2007) were crawled in January, and May 2007. For these two datasets, we have implemented the three criteria as described earlier by applying a language-specific web crawling method [22, 23] in our Thai web crawling. The start seed sets for all datasets consists of a number of popular websites and web portals in Thailand. The number of crawled web pages for Jul2004, Jan2007 and May2007 datasets are 18,344,127 pages, 551,233 pages and 1,402,206 pages respectively.

4 Connectivity of the Thai Web graph

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 number of vertices and directed edges of the Thai Web graphs induced from Thai web datasets are as shown in Table 2. In the following subsections, we will present various statistical results and discuss about the characteristics of the connectivity of the Thai Web graph.

4.1 Indegree and Outdegree distributions

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. The power-law distribution has been described by the "rich

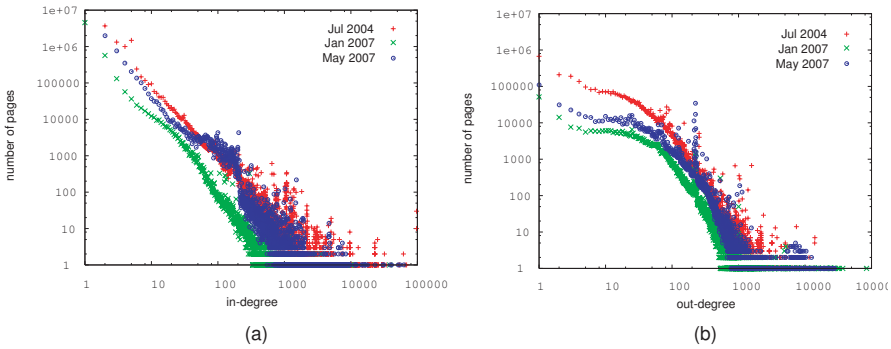
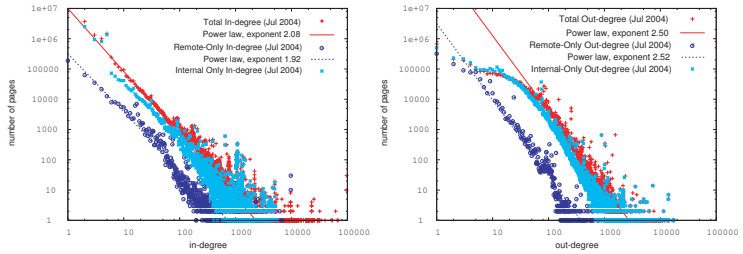


Fig. 1. Indegree and Outdegree distribution of the Thai Webgraphs

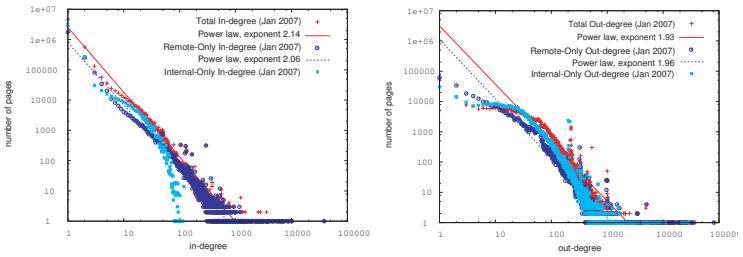
gets richer" phenomenon (or preferential attachment) where new links are more likely to point to web pages that already have many links pointing to them. We have plotted the degree distribution of the Thai Webgraphs. The in-degree distribution is shown in Fig. 1(a), and the out-degree distribution is shown in Fig. 2(b).

As can be seen from Fig. 1(a), in-degree distributions of all Thai Webgraphs in the log-log plots can be approximated by a straight line, 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 out-degree distributions in Fig. 1(b), 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 Fig. 1. Manual inspection reveals that most of the web pages corresponding to these anomalies are spam pages.

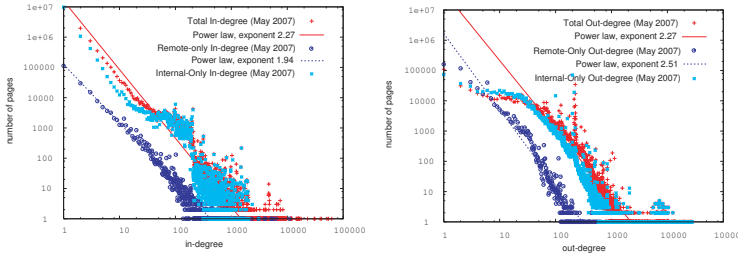
In Fig. 2, we have separately plotted the degree distributions of total, internal-only, and remote-only inlinks and outlinks. An internal link is a hyperlink between web pages within the same website. Conversely, a remote link is a hyperlink between web pages residing in different websites. According to Fig. 2, 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 the internal 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 internal linkage. Consequently, while the process of hyperlinking between web sites can be described by the "rich get richer" model, another different model or a modified



(a) Jul2004 dataset



(b) Jan2007 dataset



(c) May2007 dataset

Fig. 2. internal-only vs. remote-only degree distribution

version of "rich get richer" model is needed for explaining the phenomenon found in linkage between web pages within the same web sites.

Table 3 shows average number of inlinks and outlinks for all datasets. According to Table 3, the average number of connectivity per node is 2–4 connections per node. So, the Thai Web graph is a sparse graph with some densely connected regions which may corresponding to homepages of some popular web sites, spam pages, or web pages with pornographic content.

4.2 Weakly and Strongly Connected Components

In graph theory, a weakly connected component (WCC) is a set of nodes such that all nodes can be reached by all other nodes in the set by traversing a set

Table 3. Average number of in-degree and out-degree per page

Type of linkage	Average value for	Jul2004	Jan2007	May2007
Total	in-degree	3.2	2.2	3.7
	out-degree	3.8	3.8	4.3
Internal-only	in-degree	1.8	1.1	2.5
	out-degree	2.3	2.3	2.8
Remote-only	in-degree	1.4	1.1	1.2
	out-degree	1.5	1.5	1.5

Table 4. Weakly and strongly connected components in the Thai Web graph

	Jul2004	Jan2007	May2007
power-law exponent of WCC size distribution	1.73	1.60	1.48
total number of WCC	404,086	75,215	251,638
number of nodes in the largest WCC	38,055,065 (98.5%)	5,665,462 (98.0%)	18,314,704 (97.0%)
power-law exponent of SCC size distribution	2.00	2.12	1.34
total number of SCC	35,801,812	5,624,313	18,134,695
number of nodes in the largest SCC	2,193,836 (5.7%)	121,422 (2.1%)	539,195 (3.0%)

of undirected links. A strongly connected component (SCC) is a set of nodes in a directed graph such that all nodes can be reached by all other nodes in the set. The connectivity of a graph is an important measure of its robustness as a network. In the context of the Web graph, [10] analyzed connected components of the Web graphs induced from two AltaVista crawls. It was reported that the distribution of the size of connected components also obeys a power-law distribution. In this subsection, we will describe the statistical results obtained from the analysis of weakly connected components and strongly connected components in the Thai Web graph. Statistics of the weakly and strongly connected components of the Thai Web graphs are shown in Table 4.

According to Table 4, the largest WCC components for Jul2004, Jan2007, and May2007 datasets comprise of 98.5%, 98%, and 97% of the total number of nodes respectively. Meanwhile, the total number of WCC components in each graph are 404,086 (Jul2004), 75,215 (Jan2007), and 251,638 (May2007). The topology of the Thai Web graph can be roughly seen as consisting of a single large giant connected component and several much smaller disconnected components. As a result, it can be implied that there are substantial portions of the Web that are not reachable from the other portions of the Web. The largest SCC components for Jul2004, Jan2007, and May2007 datasets comprise of 5.7%, 2.1%, and 3.0% of the total number of nodes respectively. Our largest SCCs are very small compared to the largest SCC observed in the global Web [10] ([10] reported the value of 28%). This is due to the effects of uncrawled nodes included in our link databases. We have tried running the WCC and SCC algorithms again on Jan2007 dataset, this time we consider only those nodes which are corresponding

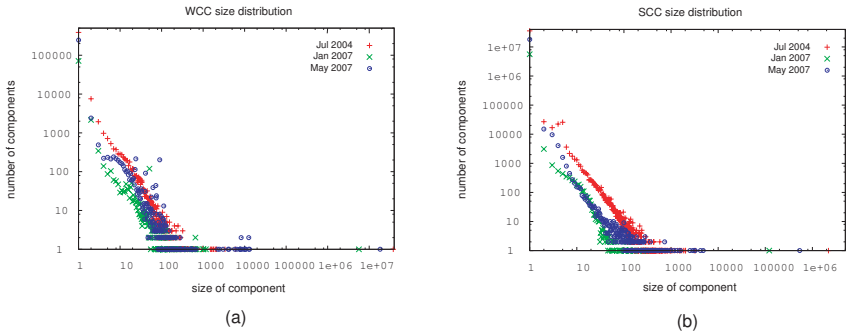


Fig. 3. distribution of the size of connected components

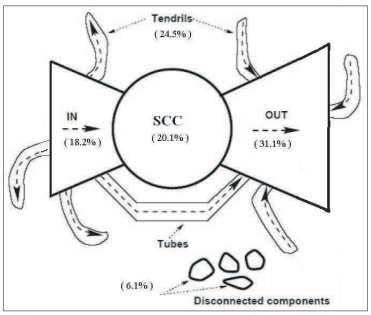


Fig. 4. asymmetric bow-tie structure of the Thai Web graph (Jan2007 dataset)

to crawled web pages. There are 551,233 nodes in Jan2007 dataset that are corresponing to crawled web pages. The Web graph induced from crawled web pages in Jan2007 dataset has the largest WCC containing (517,934 nodes [94.0% of total number of nodes]) and the largest SCC containing (110,787 nodes [20.1% of total number of nodes]). This result is more consistent with what has been observed in the global Web graph. The distributions of the size of WCC and SCC of the Thai Web graphs also adhere to the power-law. The values of the power-law exponents are given in Table 4.

4.3 Macroscopic structure of the Thai Web

In the following experiment, we would like to discover the macroscopic structure of the Thai Web by using the Web graph associated with crawled web pages of Jan2007 dataset. [10] conducted a series of SCC, WCC, and random-start BFS experiments on a large Web graph derived from AltaVista web crawls. Based on the interpretation of the experimental results, they depicted the structure of the Web as a bow-tie like structure consisting of five components: SCC, IN, OUT, TENDRILS, and DISCONNECTED. SCC is a large core component of the bow-tie. It consists of web pages in the largest strongly connected component

in the graph. IN is a component whose members are web pages that can reach the SCC but cannot be reached from the SCC. OUT is a component whose members are web pages that can be reached from the SCC but cannot reach any pages in the SCC. The TENDRILS component is a set of web pages that can be reached from IN and web pages that can only reach to OUT. Lastly, the DISCONNECTED component consists of all web pages outside the largest weakly connected component (WCC) in the Web graph.

According to our analysis result on the Jan2007 dataset, the macroscopic structure of the Thai Web is a bow-tie structure as shown in Fig. 4. However, the shape of our bow-tie is quite different from the bow-tie of the global Web. In contrast to the almost symmetric bow-tie structure of the global Web [10], the bow-tie of the Thai Web is asymmetric with large OUT component.

4.4 Connectivity between Domains

[7] studied linkage affinity between several country domains. The result of this study shows interesting patterns of linkage between country domains. For example, (1) asymmetrc linkage between country domains (e.g. links between China, Hong Kong and Taiwan), (2) the preference of language over geographic location (e.g. a strong English language affinity among US, UK, Australia, and New Zealand). In this subsection, we will present the domain linkage statistics of the Thai Web by using the link information of Jan2007 dataset. First let us consider the linkage within '.th' domains (or intra '.th' domain links). A second-level domain of a registered domain name can be used to identify types of the organizations who own the domain name. The second level domains under the '.th' domain include 'ac.th' (academic), 'co.th' (commercial), 'go.th' (governmental), 'in.th' (individuals), 'mi.th' (military), 'net.th' (internet provider), and 'or.th' (non-profit organizations).

In Table 5, we list the number of links between all pair of source and destination second-level domains. A domain in each row represents the source domain, and a domain in each column in Table 5 represents the destination domain. For each row, the column with the largest number of links is written in boldface. According to the results in Table 5, It can be seen that in almost every case the number of links within the same domain is larger than links to different domains. And there is a strong relationship between governmental, military, and non-profit organizations. Note that, according to analysis results on Jan2007 dataset, most links going out of '.th' domain is pointing to '.com' and '.net' domains. Symmetrically, most links going into the '.th' domain is coming from '.com' and '.net' domains.

5 Conclusion

In this paper we have addressed the challenge of collecting Thailand-related web pages by defining a set of criteria for deciding whether a web page belongs to the Thai Web or not. Based on three Thai web snapshots, we have observed

Table 5. Links between second-level domains under '.th' domain (Jan2007 dataset)

	ac.th	co.th	go.th	in.th	mi.th	net.th	or.th
ac.th	**12366**	2513	3604	234	49	243	1977
	58.9%	12.0%	17.2%	1.1%	0.2%	1.2%	9.4%
co.th	208	**3887**	698	151	3	31	496
	3.8%	**71.0%**	12.8%	2.8%	0.1%	0.6%	9.1%
go.th	1689	2098	**10322**	725	97	122	1757
	10.0%	12.5%	**61.4%**	4.3%	0.6%	0.7%	10.5%
in.th	557	**15990**	1367	5642	18	39	514
	2.3%	**66.3%**	5.7%	23.4%	0.1%	0.2%	2.1%
mi.th	112	131	**547**	5	475	20	181
	7.6%	8.9%	**37.2%**	0.3%	32.3%	1.4%	12.3%
net.th	25	27	**92**	4	3	60	41
	9.9%	10.7%	**36.5%**	1.6%	1.2%	23.8%	16.3%
or.th	353	425	818	44	7	35	**1239**
	12.1%	14.5%	28.0%	1.5%	0.2%	1.2%	**42.4%**

many interesting characteristics of the Thai Web graph. First, although three datasets used in our experiments are all different in scale and acquisition time, statistics derived from each dataset show similar trends and phenomena. Second, we have identified the internal links as one of the possible causes of anomalies bumps frequently found in the log-log plot of in-degree and out-degree distributions. Third, our analysis result reveal the bow-tie structure of the Thai Web as an asymmetric bow-tie. Lastly, linkage between domains reflect relationships between organizations in the real-world. For the future work, we would like to (1) study the structure of the Thai Web graph in more detail, and (2) study the evolution of the Thai Web graph.

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.
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. B. D. Davison. Topical locality in the web. In *Proc. 23rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'00)*, pages 272–279, 2000.
14. S. Dill, S. 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.
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. 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.
19. 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.
20. F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Inter. Tech.*, 4(4):378–419, 2004.
21. 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.
22. K. Somboonviwat, T. Tamura, and M. Kitsuregawa. Finding thai web pages in foreign web spaces. In *ICDE Workshops*, page 135, 2006.
23. 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.
24. 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.