Sketching Landscapes of Page Farms

Bin Zhou
Simon Fraser University, Canada
bzhou@cs.sfu.ca

Jian Pei
Simon Fraser University, Canada
jpei@cs.sfu.ca

Abstract

The Web is a very large social network. It is important and interesting to understand the "ecology" of the Web: the general relations of Web pages to their environment. The understanding of such relations has a few important applications, including Web community identification and analysis, and Web spam detection.

In this paper, we propose the notion of *page farm*, which is the set of pages contributing to (a major portion of) the PageRank score of a target page. We try to understand the "landscapes" of page farms in general: how are farms of Web pages similar to or different from each other? In order to sketch the landscapes of page farms, we need to extract page farms extensively. We show that computing page farms is NP-hard, and develop a simple greedy algorithm. Then, we analyze the farms of a large number of (over 3 million) pages randomly sampled from the Web, and report some interesting findings. Most importantly, the landscapes of page farms tend to also follow the power law distribution. Moreover, the landscapes of page farms strongly reflect the importance of the Web pages.

1 Introduction

The Web is a very large social network. Extensive work has studied a wide spectrum of Web technologies, such as searching and ranking Web pages, mining Web communities, etc.

In this paper, we investigate an important aspect of the Web – its "ecology". It is interesting to analyze the general relations of Web pages to their environment. For example, as rankings of pages have been well accepted as an important and reliable measure for the utility of Web pages, we want to understand generally how Web pages collect their ranking scores from their neighbor pages.

We argue that the "ecological" information about the Web is not only interesting but also important for a few Web applications. For example, we may detect Web spam pages effectively if we can understand the "normal" ways that Web pages collect their ranking scores. A Web page is a suspect of spam if its environment is substantially different from those normal models. Moreover, the ecological information can also help us to identify communities on the Web, analyze their structures, and understand their evolvement.

In this paper, we try to model the environment of Web pages and analyze the general distribution of such environment. We make two contributions.

First, *we propose the notion of* page farm, which is the set of pages contributing to (a major portion of) the PageRank score of a target page. We study the computational complexity of finding page farms, and show that it is NP-hard. We develop a simple greedy method to extract approximate page farms.

Second, *we empirically analyze the page farms of a large number of (over* 3 *million) Web pages randomly sampled from the Web, and report some interesting findings.* Most importantly, the landscapes of page farms tend to also follow the power law distribution. Moreover, the landscapes of page farms strongly reflect the importance of the Web pages, and their locations in their Web sites. To the best of our knowledge, this is the first empirical study on extracting and analyzing page farms. Our study and findings highly suggest that sketching the landscapes of page farms provides a novel approach to a few important applications.

The remainder of the paper is organized as follows. The notion of page farm is proposed in Section 2. We give a simple greedy method to extract page farms in Section 3, and report an empirical analysis on the page farms of a large number of Web pages in Section 4. In Section 5, we review the related work. The paper is concluded in Section 6.

2 Page Farms

The Web can be modeled as a directed *Web graph* $G = (V, E)$, where V is the set of Web pages, and E is the set of hyperlinks. A link from page p to page q is denoted by edge $p \rightarrow q$. An edge $p \rightarrow q$ can also be written as a tuple (p, q). Hereafter, by default our discussion is about a directed Web graph $G = (V, E)$.

PageRank [13] measures the importance of a page p by considering how collectively other Web pages point to p directly or indirectly. Formally, for a Web page p, the PageRank score is defined as

$$PR(p, G) = d \sum_{p_i \in M(p)} \frac{PR(p_i, G)}{OutDeg(p_i)} + (1 - d), \quad (2.1)$$

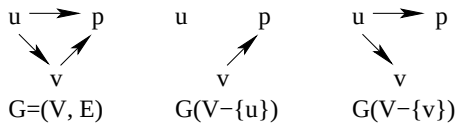

Figure 1: Page contributions.

where $M(p) = \{q|q \to p \in E\}$ is the set of pages having a hyperlink pointing to p, $OutDeg(p_i)$ is the out-degree of p_i (i.e., the number of hyperlinks from p_i pointing to some pages other than p_i), and d is a damping factor which models the random transitions on the Web.

To calculate the PageRank scores for all pages in a graph, one can assign a random PageRank score value to each node in the graph, and then apply Equation 2.1 iteratively until the PageRank scores in the graph converge.

For a Web page p, can we analyze which other pages contribute to the PageRank score of p? An intuitive way to answer the above question is to extract the Web pages that contribute to the PageRank score of the target page p. This idea leads to the notion of page farms.

Generally, for a page p, the *page farm* of p is the set of pages on which the PageRank score of p depends. Page p is called the *target page*. According to Equation 2.1, the PageRank score of p directly depends on the PageRank scores of pages having hyperlinks pointing to p. The dependency is transitive. Therefore, a page q is in the page farm of p if and only if there exists a directed path from q to p in the Web graph.

As indicated in the previous studies [1, 3], the major part of the Web is strongly connected. Albert et al. [1] indicated that the average distance of the Web is 19. In other words, it is highly possible to get from any page to another in a small number of clicks. A strongly connected component of over 56 million pages is reported in [3].

Therefore, the page farm of a Web page can be very large. It is difficult to analyze large page farms of a large number of Web pages. Instead, *can we capture a subset of pages that contribute to a large portion of the PageRank score of a target page?*

According to Equation 2.1, PageRank contributions are only made by the out-edges. Thus, a vertex in the Web graph is *voided* for PageRank score calculation if all edges leaving the vertex are removed. Please note that we cannot simply remove the vertex. Consider Graph G in Figure 1. Suppose we want to void page v in the graph for PageRank calculation. Removing v from the graph also reduces the out-degree of u, and thus change the PageRank contribution from u to p. Instead, we should retain v but remove the out-link $v \to p$.

For a set of vertices U, the *induced subgraph* of U (with respect to PageRank score calculation) is given by $G(U) = (V, E')$, where $E' = \{p \to q|p \to q \in E \wedge p \in U\}$. In other words, in $G(U)$, we void all vertices that are not in U. Figure 1 shows two examples.

To evaluate the contribution of a set of pages U to the PageRank score of a page p, we can calculate the PageRank score of p in the induced subgraph of U. Then, the PageRank contribution is given by

$$Cont(U, p) = \frac{PR(p, G(U))}{PR(p, G)} \times 100\%$$

PageRank contribution has the following property. The proof can be found in [16].

COROLLARY 2.1. (PAGERANK CONTRIBUTION) *Let p be a page and U, W be two sets of pages. If $U \subseteq W$, then $0 \leq Cont(U, p) \leq Cont(W, p) \leq 1$.* ∎

We can capture the smallest subset of Web pages that contribute to at least a θ portion of the PageRank score of a target page p as the θ-(page) farm of p.

DEFINITION 1. (θ-FARM) Let θ be a parameter such that $0 \leq \theta \leq 1$. A set of pages U is a θ-farm of page p if $Cont(U, p) \geq \theta$ and $|U|$ is minimized. ∎

However, finding a θ-farm of a page is computationally costly on large networks.

THEOREM 2.1. (θ-FARM) *The following decision problem is NP-hard: for a Web page p, a parameter θ, and a positive integer n, determine whether there exists a θ-farm of p which has no more than n pages.*

Proof sketch. The proof is constructed by reducing the NP-complete knapsack problem [11] to the θ-farm problem. Please see [16] for the complete proof. ∎

Searching many pages on the Web can be costly. Heuristically, the near neighbors of a Web page often have strong contributions to the importance of the page. Therefore, we propose the notion of (θ, k)-farm.

In a directed graph G, let p, q be two nodes. The *distance* from p to q, denoted by $dist(p, q)$, is the length (in number of edges) of the shortest directed path from p to q. If there is no directed path from p to q, then $dist(p, q) = \infty$.

DEFINITION 2. ((θ, k)-FARM) Let $G = (V, E)$ be a directed graph. Let θ and k be two parameters such that $0 \leq \theta \leq 1$ and $k > 0$. k is called the **distance threshold**. A subset of vertices $U \subseteq V$ is a **(θ, k)-farm** of a page p if $Cont(U, p) \geq \theta$, $dist(u, p) \leq k$ for each vertex $u \in U$, and $|U|$ is minimized. ∎

We notice that finding the exact (θ, k)-farms is also NP-hard. The details can be found in [16] as well.

3 Extracting Page Farms

Extracting the exact θ-farm and (θ, k)-farm of a Web page is computationally challenging on large networks. In this section, we give a simple greedy method to extract approximate page farms.

Intuitively, if we can measure the contribution from any single page v towards the PageRank score of a target page p, then we can greedily search for pages of big contributions and add them into the page farm of p.

Definition 3. (Page contribution) For a target page $p \in V$, the *page contribution* of page $v \in V$ to the PageRank score of p is $PCont(v, p) = PR(p, G) - PR(p, G(V - \{v\}))$ when $v \neq p$, and $PCont(p, p) = 1 - d$ where d is the damping factor. ∎

Example 1. (Page contributions) Consider a simple Web graph G in Figure 1. The induced subgraphs $G(V - \{u\})$ and $G(V - \{v\})$ are also shown in the figure. As specified in Section 2, all vertices are retained in an induced subgraph.

Let us consider page p as the target page, and calculate the page contributions of other pages to the PageRank of p. According to Equation 2.1, the PageRank score of p in G is given by $PR(p, G) = -\frac{1}{2}d^3 - d^2 + \frac{1}{2}d + 1$. Moreover, the PageRank score of p in $G(V - \{u\})$ is $PR(p, G(V - \{u\})) = -d^2 + 1$, and the PageRank score of p in $G(V - \{v\})$ is $PR(p, G(V - \{v\})) = -\frac{1}{2}d^2 - \frac{1}{2}d + 1$.

Thus, the page contributions are calculated as $PCont(u, p) = PR(p, G) - PR(p, G(V - \{u\})) = -\frac{1}{2}d^3 + \frac{1}{2}d$, and $PCont(v, p) = PR(p, G) - PR(p, G(V - \{v\})) = -\frac{1}{2}d^3 - \frac{1}{2}d^2 + d$. ∎

Using the page contributions, we can greedily search a set of pages that contribute to a θ portion of the PageRank score of a target page p. That is, we calculate the page contribution of every page (except for p itself) to the PageRank score of p, and sort the pages in the contribution descending order. Suppose the list is $u_1, u_2, \cdots$. Then, we select the top-l pages $u_1, \cdots, u_l$ as an approximation of the θ-farm of p such that $\frac{PR(p, G(V - \{u_1, \cdots, u_l\}))}{PR(p, G)} \geq \theta$ and $\frac{PR(p, G(V - \{u_1, \cdots, u_{l-1}\}))}{PR(p, G)} \leq \theta$. To extract (θ, k)-farms, we only need to consider those pages at a distance at most k to the target page p.

The above greedy method is simple. However, it may be still quite costly for large Web graphs. In order to extract the page farm for a target page p, we have to compute the PageRank score of p in induced subgraph $G(V - \{q\})$ for every page q other than p. The computation is costly since the PageRank calculation is an iterative procedure and often involves a huge amount of Web pages and hyperlinks. On our current PC, extracting 5000 page farms in a Web graph containing about $3M$ pages needs more than 3000 seconds. A more efficient greedy algorithm can be found in [16].

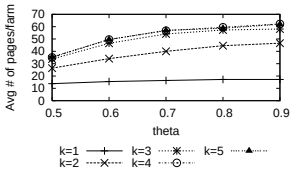

Figure 2: The effects of parameters k and θ.

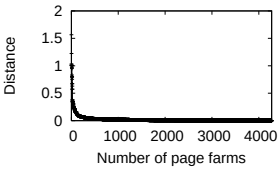

Figure 3: The distribution of distance to the mean of the data set.

4 Empirical Analysis of Page Farms

In this section, we report an empirical analysis of page farms of a large sample from the Web. The data set we used was generated by the Web crawler from the Stanford WebBase project (http://www-diglib.stanford.edu/~testbed/doc2/WebBase). Some prior studies [8, 9, 10] used the same data set in their experiments. The Web crawler, WebVac, randomly crawls up to a depth of 10 levels and fetches a maximum of 10 thousand pages per site. The whole directed Web graph file for May, 2006 is about 499 GB and contains about 93 million pages.

Limited by the computational resource available to us, in our experiments, we only used a random sample subgraph of the whole Web graph. The sample we used is about 16 GB and contains $3,295,807$ pages. Each page in our data set has a viable URL string.

All the experiments were conducted on a PC computer running the Microsoft Windows XP SP2 Professional Edition operating system, with a 3.0 GHz Pentium 4 CPU, 1.0 GB main memory, and a 160 GB hard disk. The program was implemented in C/C++ using Microsoft Visual Studio. NET 2003.

4.1 Extracting Page Farms To understand the effects of the two parameters θ and k on the page farms extracted, we extracted the (θ, k)-farms using different values of θ and k, and measured the average size of the extracted farms. Figure 2 shows the results on a sample of $4,274$ Web pages from site "`http://www.fedex.com`".

When θ increases, more pages are needed to make up the contribution ratio. However, the increase of the average page farm size is sublinear. The reason is that when a new page is added to the farm, the contributions of some pages already in the farm may increase. Therefore, a new page often boosts the contributions from multiple pages in the farm. The larger and denser the farm, the more contribution can be made by adding a new page. On average, when

Site-id	Site	# pages crawled
Site-1	http://www.fedex.com	4274
Site-2	http://www.siia.net	2722
Site-3	http://www.indiana.edu	2591
Site-4	http://www.worldbank.org	2430
Site-5	http://www.fema.gov	4838
Site-6	http://www.liverpoolfc.tv	1854
Site-7	http://www.eca.eu.int	4629
Site-8	http://www.onr.navy.mil	4586
Site-9	http://www.dpi.state.wi.us	5118
Site-10	http://www.pku.edu.cn	6972
Site-11	http://www.cnrs.fr	2503
Site-12	http://www.jpf.go.jp	5685
Site-13	http://www.usc.es	2138

Table 1: List of sites with different domains.

$\theta \geq 0.8$, page farms are quite stable and capture the major contribution to PageRank scores of target pages.

When k is small, even selecting all pages of distance up to k may not be able to achieve the contribution threshold θ. Therefore, when k increases, the average page farm size increases. However, when k is 3 or larger, the page farm size is stable. This verifies our assumption that the near neighbor pages contribute more than the remote ones.

We also compared the page farms extracted using different settings of the two parameters. The farms are quite robust. That is, for the same target page, the page farms extracted using different parameters overlap largely. We also conducted the same experiments on other sites. The results are consistent. Thus, in the rest of this section, we report results on $(0.8, 3)$-farms of Web pages.

4.2 Page Farm Analysis on Individual Sites To analyze a large collection of page farms, we conducted the clustering analysis on the page farms extracted. Our analysis was in two steps. First, we analyzed the page farms in individual sites. Then, we analyzed the page farms in the whole data set (Section 4.3).

In the data set, there are about 50 thousand different sites and about 30 different domains[1]. In order to analyze the page farms of individual sites, we randomly selected 13 sites with different domains, as listed in Table 1. These sites include some popular domains, such as .com, .net, .edu, .org and .gov, as well as some unpopular ones, such as .tv, .int and .mil. Moreover, some domains from different countries and different languages are also involved, such as .us(USA), .cn(China), .fr(France), .jp(Japan) and .es(Spain).

[1]Details can be found at http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/crawl_lists/crawled_hosts.05-2006.f

# clusters	C_1	C_2	C_3	C_4	C_5
2	22	4252			
3	19	103	4152		
4	19	89	543	3623	
5	19	87	230	1280	2658

Table 2: The number of pages in each cluster when the number of clusters varies from 2 to 5.

We first generated the complete Web graph from the data set containing nearly 3.3 million Web pages. A normal power method [2] was used to calculate the PageRank scores. For the pages in each site, we then extracted the $(0.8, 3)$-farm.

Based on Definition 2, a page farm U is a set of pages. We can easily obtain the induced graph $G(U)$ by adding the links between pages in the farm. To analyze the page farms, we extracted the following features of each farm and its corresponding induced graph: (1) the number of pages in the farm; (2) the total number of intra-links in the induced graph; and (3) the total number of inter-links in the induced graph. Here, intra-links are edges connecting pages in the same farm, and inter-links are edges coming from or leaving a farm. We also considered some other features, such as the average in- and out-degrees, average PageRank score, and diameter of the induced graph. The clustering results are consistent. Thus, we only used the above three features as representatives to report the results here.

The above 3 attributes are independent with each other and each one is an important factor to reveal the characteristics of the page farms. Each attribute has the same importance in our analysis. Thus, we normalized all attribute values into the range $[0, 1]$ in the clustering analysis. These 3 normalized attribute values form the vector space for each page farm. We applied the conventional k-means clustering, where the Euclidian distance was adopted to measure the distance between two page farm vectors.

We varied the number of clusters, and compare the clusters obtained. Interestingly, if we sort all clusters according to the size (i.e., the number of pages in the clusters), those small clusters are robust when the number of clusters increases. Setting the number of clusters larger tends to split the largest cluster to generate new clusters.

For example, Table 2 shows the number of pages in each cluster when the number of clusters varies from 2 to 5. A set of $4,274$ Web pages sampled from Web site "`http://www.fedex.com`" was used. By comparing the pages in the clusters, we found that the pages in C_1 are largely the same no matter how the number of clusters

Cluster	URLs
C_1	http://www.fedex.com/
	http://www.fedex.com/us/customer/
	http://www.fedex.com/us/
	http://www.fedex.com/us/careers/
	http://www.fedex.com/us/services/
C_2	http://www.fedex.com/legal/?link=5
	http://www.fedex.com/us/search/
	http://www.fedex.com/us/privacypolicy.html?link=5
	http://www.fedex.com/us/investorrelations/?link=5
	http://www.fedex.com/us/about/?link=5
C_3	http://www.fedex.com/legal/copyright/?link=2
	http://www.fedex.com/us?link=4
	http://www.fedex.com/us/about/today/?link=4
	http://www.fedex.com/us/investorrelations/financialinfo/2005annualreport/?link=4
	http://www.fedex.com/us/dropoff/?link=4
C_4	http://www.fedex.com/ca_english/rates/?link=1
	http://www.fedex.com/legal/
	http://www.fedex.com/us/about/news/speeches?link=2
	http://www.fedex.com/us/customer/openaccount/?link=4
	http://www.fedex.com/us/careers/companies?link=4
C_5	http://www.fedex.com/?location=home&link=5
	http://www.fedex.com/ca_french/rates/?link=1
	http://www.fedex.com/ca_french/?link=1
	http://www.fedex.com/ca_english/?link=1
	http://www.fedex.com/us/careers/diversity?link=4

Table 3: The top-5 URLs with the highest PageRank scores in each cluster.

is set. When the number of clusters varies from 3 to 5, the clusters C_2 of different runs also largely overlap with each other.

The above observation strongly indicates that the distance from Web pages to the center of the whole data set may follow a power law distribution. To verify, we analyzed the distances between the page farms in the site to the mean of the sample set of the site. The results are shown in Figure 3. The distance follows the power law distribution as expected. This clearly explains why the smaller clusters are robust and the new clusters are often splitting from the largest cluster.

As the clusters are robust, how are the pages in different clusters different from each other? In Table 3, we list the top-5 URL's in each cluster that have the highest PageRank scores. Interestingly, most pages in the first cluster are the portal pages. The later clusters often have more and more specific pages of lower

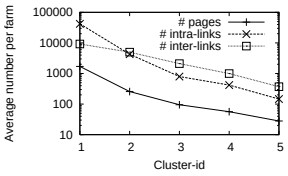

Figure 4: The features of clusters.

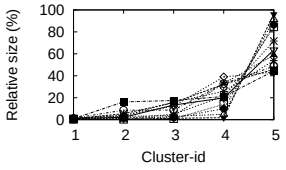

Figure 5: The distribution of cluster size.

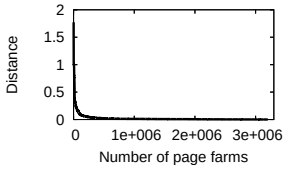

Figure 6: The distribution of the distance to the center of the data set.

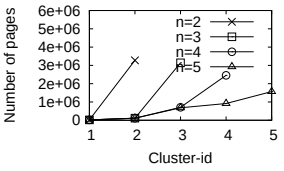

Figure 7: The size of clusters.

PageRanks. Correspondingly, In Figures 4, we show for each cluster the average size, the average number of intra-links, and the average number of inter-links. As can be seen, they follow the similar trend. The smaller the clusters, the larger the page farms and thus more intra- and inter-links in the farms.

4.3 Page Farms of Multiple Sites and in the Whole Data Set The findings in Section 4.2 are not specific for a particular Web site. Instead, we obtained consistent observations in other Web sites, too. For example, we clustered the page farms for the 13 Web sites listed in Table 1 by setting the number of clusters to 5. For each site, the clusters were sorted in ascending order of the number of pages, and the ratio of the number of pages in a cluster versus the total number of pages sampled from the site was used as the relative size of the cluster. Figure 5 shows the result. We can observe that the distributions of the relative cluster size follow the same trend in those sites.

In Section 4.2, we examined the page farms in individual Web sites. To test whether the properties observed were scale-free, we conducted the similar experiments on the large sample containing $3,295,807$ Web pages. The experimental results confirm that the properties are scale-free: we observed the similar phenomena on the large sample.

Figure 6 shows the distribution of distances of page farms to the mean of the whole data set. Clearly, it follows the power law distribution.

Moreover, we clustered the page farms by varying the number of clusters from 2 to 5, and sorted the clusters in size ascending order. The results are shown in Figure 7, where parameter n is the number of clusters.

The figure clearly shows that the smaller clusters are robust and the new clusters are splitting from the largest clusters when the number of clusters is increased.

4.4 Summary From the above empirical analysis of the page farms of a large sample of the Web, we can obtain the following two observations.

First, *the landscapes of page farms follow a power law distribution and the distribution is scale-free.* The phenomena observed from individual large Web sites is nicely repeated on the large sample containing many Web sites across many domains.

Second, *Web pages can be categorized into groups according to their page farms. Some interesting features are associated with the categorization based on clustering*, such as the relative importance of the pages and the relative positions in the Web sites. The distinguishing groups are robust with respect to the clustering parameter settings.

5 Related Work

Our study is highly related to the previous work on the following two areas: (1) link structured-based ranking and its applications in Web community identification and link spam detection; and (2) social network analysis. Social network analysis is a topic that has been studied extensively and deeply (see [15, 14] as textbooks). In this section, we only focus on some representative studies on the first area.

A few link structured-based ranking methods such as HITS [12] and PageRank [13] were proposed to assign scores to Web pages to reflect their importance. The details of PageRank are recalled in Section 2. Using the link structure-based analysis, previous studies have developed various methods to identify Web communities – collections of Web pages that share some common interest on a specific topic.

For example, Gibson et al. [4] developed a notion of hyper-linked communities on the Web through an analysis of the link topology. As another example, Kleinberg [12] showed that the HITS algorithm, which is strongly related to spectral graph partitioning, can identify "hub" and "authority" Web pages. A hub page links to many authority pages and an authority page is pointed by many hub pages. Hubs and authorities are especially useful for identifying key pages related to some community.

Most of the popular search engines currently adopt some link structure-based ranking algorithms, such as PageRank and HITS. Driven by the huge potential benefit of promoting rankings of pages, many attempts have been conducted to boost page rankings by making up some linkage structures, which is known as link spam [2, 7].

Because the PageRank score are determined based on the link structure of the Web, PageRank is a natural target to link spam. Gyöngyi et al. [7, 6] referred link spam to the cases where spammers set up structures of interconnected pages, called link spam farms, in order to boost the connectivity-based ranking.

6 Conclusions

To the best of our knowledge, this is the first empirical study on extracting and analyzing page farms from samples of the Web. We developed a simple yet effective model of page farms, and devised a simple greedy algorithm to extract page farms for a large Web graph with numerous pages.

As future work, we plan to develop more efficient algorithms for page farm extraction and analysis, and extend the applications of page farm analysis.

References

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

[2] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology*, 5(1), 2005.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *WWW'00*.

[4] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[5] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *VLDB'06*.

[6] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *VLDB'05*.

[7] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb'05*.

[8] T. Haveliwala. Topic-sensitive pagerank. In *WWW'02*.

[9] T. Haveliwala, A. Gionis. Evaluating strategies for similarity search of the web. In *WWW'02*.

[10] G. Jeh and J. Widom. Scaling personalized web search. In *WWW'03*.

[11] R. M. Karp. Reducibility Among Combinatorial Problems. Plenum Press, 1972.

[12] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA'98*.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[14] J. Scott. Social Network Analysis Handbook. Sage Publications Inc., 2000.

[15] S. Wasserman and K. Faust. Social Network Analysis. Cambridge University Press, 1994.

[16] B. Zhou. Mining page farms and its application in link spam detection. Master thesis, Simon Fraser University, 2007.