A Large-Scale Study of Link Spam Detection by Graph Algorithms

Hiroo Saito[1,2] Masashi Toyoda[2] Masaru Kitsuregawa[2] Kazuyuki Aihara[1,2]

[1]Aihara Complexity Modelling Project, ERATO, JST
Komaba 4-6-1, Tokyo, 153-8505, Japan

[2]Institute of Industrial Science, University of Tokyo
Komaba 4-6-1, Tokyo, 153-8505, Japan

saito_h@aihara.jst.go.jp {toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp aihara@sat.t.u-tokyo.ac.jp

ABSTRACT

Link spam refers to attempts to promote the ranking of spammers' web sites by deceiving link-based ranking algorithms in search engines. Spammers often create densely connected link structure of sites so called "link farm". In this paper, we study the overall structure and distribution of link farms in a large-scale graph of the Japanese Web with 5.8 million sites and 283 million links. To examine the spam structure, we apply three graph algorithms to the web graph. First, the web graph is decomposed into strongly connected components (SCC). Beside the largest SCC (core) in the center of the web, we have observed that most of large components consist of link farms. Next, to extract spam sites in the core, we enumerate maximal cliques as seeds of link farms. Finally, we expand these link farms as a reliable spam seed set by a minimum cut technique that separates links among spam and non-spam sites. We found about 0.6 million spam sites in SCCs around the core, and extracted additional 8 thousand and 49 thousand sites as spams with high precision in the core by the maximal clique enumeration and by the minimum cut technique, respectively.

1. INTRODUCTION

Link spam refers to attempts to promote the ranking of spammers' web sites by deceiving link-based ranking algorithms in search engines. Spammers often create densely connected link structure of sites so called "link farm" using various techniques as summarized in the paper [8]. Since these activities could deteriorate quality of search engines, it is an important issue to detect such spam sites among the World Wide Web.

There has been little work on the link structure of link farms. In this paper, we study the overall structure and distribution of link farms in a large-scale graph of the Japanese Web with 5.8 million sites and 283 million links.

To examine the spam structure, we apply three graph algorithms to the web graph. First, the web graph is decomposed into strongly connected components (SCC). We found the largest SCC, which we call *core*, which contains about 30% of whole sites as observed in [5]. Surprisingly, we have observed that most of large components around the core consist of link farms. We found about 0.6 million spam sites in these components. Since the core still contains many spam sites, we need more stronger criterion to extract link farms in it. Concretely, we enumerate maximal cliques in the core as seeds of link farms. This enumeration found about 8,000 spam sites that form tightly connected link farms. We can use those link farms (large SCCs and maximal cliques) as a reliable spam seed set. Then, we expand those link farms by a minimum cut technique that separates links among spam and non-spam sites. This technique extracted additional 49 thousand sites as spams with high precision. Hence, we obtained about 57 thousand spams in the core.

The rest of this paper is organized as follows. In Section 2, we briefly overview previous works related to our results. In Section 3, we describe our data set. In Section 4, we show the results of the SCC decomposition to our dataset. In Section 5, we apply a maximal clique enumeration algorithm to the decomposed components. Furthermore, in Section 6, we propose a clustering by minimum cuts in a network obtained by web graph. Finally, we present concluding remarks in Section 7.

2. RELATED WORK

Link spamming mainly attacks link based ranking algorithms such as PageRank [16] and HITS [10] that consider a link to a page as an endorsement of that page. To increase these ranking scores, spammers often add outgoing links to popular sites and gather many links to their target sites in various way, for example, by connecting their sites each other [8]. Influences of link spamming techniques on PageRank are examined in [4, 7].

Various techniques have been proposed for combating link spam. Some of them use machine learning for classifying spam and non-spam pages based on features related to link structure, such as indegree, outdegree, and link based rank-

AIRWEB '07, May 8, 2007, Banff, Alberta, CANADA.

ing scores of pages [1, 2, 3]. Another approach is to improve link-based ranking algorithms to be robust against link spamming. Various ranking algorithms are proposed and tested on real web data in [4, 9, 11]. There have also been some link-based approaches [15, 17, 13] that extract link spam using some graph theoretical notions such as biconnected components and commonality of neighbors.

As far as we know, there has been no study investigating the distribution of link spam sites and their connections in the global web graph structure, that are important for locating a large amount of spam sites.

Broder et al. [5] also applied SCC decomposition to a web graph, and observed a global property of the graph. In addition to their observations, we have found that almost all sites in the large SCCs around the core are spams and that they are connected to by the core.

Flake et al. proposed an efficient identification of web communities, which are sets of sites related to a topic, by maximum flow/ minimum cut framework [6]. These approaches ignore directions of links among sites. However, in link spam detection, the direction of links is significantly important because spam sites often point to good sites and good sites seldom point to spam sites as observed in [9]. Hence, we propose a maximum flow framework taking into account directions of links and applied it to the core.

3. DATASET

,

For the experiments, we used a large-scale snapshot of our Japanese web archive built by a crawl conducted in May 2004. Basically, our crawler is based on breadth-first crawling [14]; except that it focuses on pages written in Japanese. We collected pages outside the .jp domain if they were written in Japanese. We used a web site as a unit when filtering non-Japanese pages. The crawler stopped collecting pages from a site, if it could not find any Japanese pages on the site within the first few pages. Hence, this dataset contains fairly amount of English or other language pages. The amount of Japanese pages is estimated to be 60%. This snapshot is composed of 96 million pages and 4.5 billion links.

We use a site level graph of the Web, in which nodes are web sites and edges represent the existence of links between pages in different sites. In the site graph, we can easily find dense connections between spam sites, that cannot be found in the page level graph.

To build the site graph, we first chose the representative page of each site that has 3 or more in-links from other sites, and whose URL is within 3 tiers (i.e. http://A/B/C/). Then, pages below each representative page are contracted to one site. Finally, edges between two sites are created when there exist links between pages in these sites. The site graph built from our snapshot includes 5.8 million sites and 283 million links.

We call this dataset web graph in this paper. Certain properties and its statistics of domains of our web graph is shown in Table 1 and 2, respectively. Our computation was performed on a workstation with four AMD Opteron 2.8 GHz processors and 16 GB RAM.

4. STRONGLY CONNECTED COMPONENTS DECOMPOSITION

A digraph is called *strongly connected* if for each pair of nodes there exists a directed path between them. *Strongly connected component* (SCC) of a digraph is its maximal strongly connected subgraph. Note that the SCCs uniquely decompose the node set into a disjoint union of node subsets and that such decomposition can be found in linear time. Since spam sites densely connect each other by links and good sites seldom link to spam ones, we can expect that spams form SCCs without good sites.

Table 1: Properties of the web graph

Number of nodes	5,869,430
Number of arcs	283,599,786
Maximum of indegree (outdegree)	61,006 (70,294)
Average of indegree (outdegree)	48 (48)

Table 2: Domains in the web data

Domains	Numbers	Ratio (%)
.com	2,711,588	46.2
.jp	1,353,842	23.1
.net	436,645	7.4
.org	211,983	3.6
.de	169,279	2.9
.info	144,483	2.5
.nl, .kr, .us, etc.	841,610	14.3

In the following, we present some observations of our benchmark data. The web graph is decomposed into distinct 3,424,385 components. The size distribution of components, as observed in [5], follows the power law with the exponent 2.32. There is the largest component (core) whose cardinality is 1,784,322. This core contains many prestigious and can be viewed as a center of our data. The number of components with a single page is 3,398,316.

Following the paper [5], we present a bow tie structure in Figure 1 which shows connection among the core and other parts with their ratio of sites contained. We will see how link farms are distributed in this structure in the rest of this section.

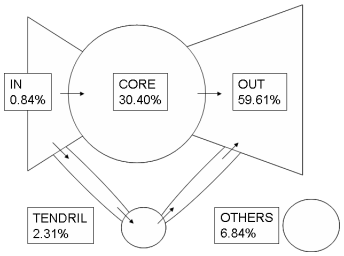

Figure 1: The bow tie structure of our dataset

Next, we consider relatively large components with more than 100 nodes except for the core. The number of such components is 551 and about 571 thousand sites are contained therein. The frequent domains are .com (70.2%), .info (16.4%), and .net (8.5%). We mention that there is no .jp domain site.

Figure 2 shows the density of each component. The density of an induced subgraph by $S \subseteq V$ is defined by the ratio of the number of its arcs to the largest possible number of arcs, i.e., $A(S)/(|S|(|S|-1))$. We can see that more than half of them forms dense link structures with density greater than 0.5.

We also examined contents of randomly sampled 550 sites

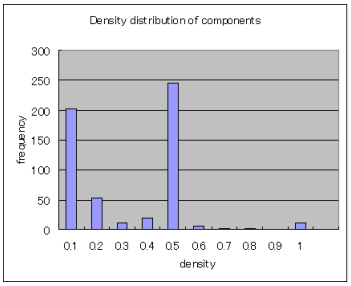

Figure 2: Density of components

by choosing 110 components and 5 sites from those components and classified them into three categories, i.e., spam, suspicious, and non-spam. Table 3 shows that most of the samples are spams.

Table 3: Sampling from the large components (without the core)

	Spam	Suspicious	Non-spam
Number	527	23	0
Ratio (%)	95.8	4.2	0

Finally, we consider the distribution of the obtained components in the bow tie structure. Table 4 shows the distribution which indicates that most of the link farms lie in OUT. The digraph in Figure 3 shows connection among the core

Table 4: Distribution of components with more than 100 sites (without the core) in the bow tie structure

	OUT	IN	TENDRIL	OTHERS
Sites	448,688	4,919	40,425	77,626
Components	450	10	32	59

and link farms. This graph is obtained by contracting each SCC in the core, IN, and OUT to a node and deleting every SCC whose size is less than 1,000. The largest component (the core) is denoted by the big black dot in Figure 3. There are many large SCCs around the core and, surprisingly, all of them are link farms. There are many arcs from the core to the large components in OUT. It means that those SCCs form optimal link farms [7] that collect inlinks and do not point to any sites outside. We remark that the components in Figure 3 are connected through smaller components ignored in illustration.

Since these SCCs contain spam sites with very high precision, we can use them as a seed set for further detection of link farms in Section 6.

Note that strongly connectivity is a very weak criterion for decomposition. For example, it is easy for spammers to counterattack this method simply by linking to good sites. Actually, core still contains many spam sites. Thus, we consider a stronger criterion for denseness of a subgraph in the next section.

5. ENUMERATION OF MAXIMAL CLIQUES

In this section, we enumerate maximal cliques in the web graph in order to find spam sites in the core. We consider that large cliques are probably made by spammers.

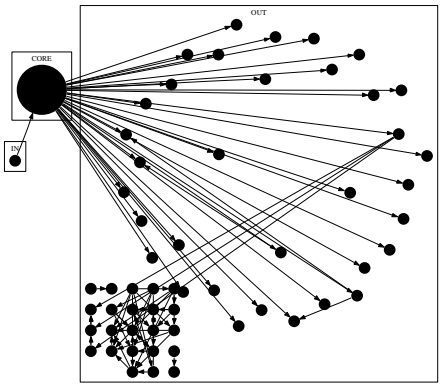

Figure 3: Connection of the components

For a digraph $G = (V, A)$, we call a node subset $K \subseteq V$ *clique* if there exist arcs $(u, v), (v, u) \in A$ for any pair of nodes $u, v \in K$. Let $G_C = (C, A(C))$ be a node induced subgraph of G for an SCC $C \subseteq V$. We define an undirected graph $G'_C = (V_C, E_C)$ from G_C by $E_C = \{\{u, v\} \mid (u, v), (v, u) \in A(C)\}$ and $V_C \subseteq C$ is the set of endpoints of each edge in E_C. Note that K is a clique in the directed graph G_C if and only if K is a clique (in the usual sense) in the undirected graph G'_C.

In the following, we show our computational results. In the rest of this paper, we concentrate the core C which is the largest component obtained in Section 4. By making the graph undirected, the number of nodes and edges becomes 1.09 million and 6.59 million, respectively. The maximum degree of the undirected graph G'_C is 3,094. Since the maximal clique enumeration algorithm by [12] takes $O(\Delta^4)$ to find a maximal clique, where Δ denotes the maximum degree of a graph, it takes too much time for G'_C because the maximum degree is 3,000. Hence, we truncate the nodes with degree greater than 80. This ad hoc procedure reduces the number of nodes and of edges to 1,064,922 and 3,222,189 and average degree to 6.5. Furthermore, we ignore maximal clique whose size is less than 40, because the number of such cliques is enormous and larger one tends to be a link farm.

The number of obtained cliques is 26,931 which contains 8,346 distinct sites. Note that most of these cliques overlaps each other. These 8,346 sites consist of .com (62.8%), .net (11.2%), and other domains. We sampled 165 sites from the

Table 5: Sampling from sites in cliques

	Spam	Suspicious	Non-spam
Number	157	8	0
Ratio (%)	95.2	4.8	0

8,346 sites and manually classified them in Table 5. It is observed that most of those cliques are composed of spam sites. Since these cliques also contain spam sites with very high precision, we can use them as a seed set for further detection of link farms in the next section.

6. SPAM DETECTION BY MINIMUM CUT

In this section, we propose a spam detection technique based on minimum s-t cut and its dual, i.e., maximum s-t flow. We utilize the SCCs and the cliques obtained in the previous sections as seed sets to be expanded with more

spams. Here, we again address to intuition that spam sites link to good ones whereas good sites seldom link to spam ones. This implies "local edge connectivity" from good sites to spam sites is small in the core. Hence, we can expect that a minimum cut separates links from good sites to spams when we consider a flow emanating from good sites towards spam sites.

Let $W \subseteq V$ and $B \subseteq V$ denote seed sets of good and spam sites, respectively. We introduce a network $\mathcal{N}_{W,B} = ((V \cup \{s,t\}, A \cup A_s \cup A_t), c, s, t)$ with a virtual source s, a virtual sink t, $A_s = \{(s,v) \mid v \in W\}$, $A_t = \{(u,t) \mid u \in B\}$, and arc capacity $c(a) = \infty$ for $a \in A_s \cup A_t$ and $c(a) = 1$ otherwise. In this network, we solve a minimum s-t cut problem

$$\min\{\sum_{a \in A(U:V\setminus U)} c(a) \mid U \cup \{s\} \text{ is an } s\text{-}t \text{ cut}\}, \qquad (1)$$

where $A(U : V \setminus U)$ is the cutset of U. We regard $V \setminus U^*$ as a link farm, where U^* is a node subset such that $U^* \cup \{s\}$ is a minimum s-t cut.

In actual computation, we solved a maximum s-t flow problem which is dual of (1) with the push-relabel algorithm by Goldberg and Tarjan, which took about 30 minutes. Since minimum s-t cut is not unique, we took a minimum s-t cut by depth first search from t in the (arc-reversed) residual graph of the obtained maximum flow.

Next, we present the computational result. A network is constructed with the core. We manually chose seed set of 210 good sites. Seed set of spams is the 8,346 sites in the cliques in Section 5 and 448,688 sites in the large SCCs (in OUT) connected from the core obtained in Section 4.

Our minimum cut technique expanded the seed set with 8,346 sites to 57,350 sites in the core. Hence, we obtained additional 49,004 sites in the core. The most dominant domain is .com (62.9%), and .net (19.8%) follows. It includes only 62 .jp domain pages. Table 6 is the result of sampling 487 sites from the obtained link farms, which shows a reasonable quality of our method.

Table 6: Sampling from sites in the minimum cut

	Spam	Suspicious	Non-spam
Number	459	27	1
Ratio (%)	94.25	5.54	0.21

7. CONCLUDING REMARKS

We have studied the overall structure and distribution of link farms in a large-scale web graph with graph algorithms. First, we applied the SCC decomposition to the web graph and found that most of large components around the core consist of link farms. Next, we enumerated maximal cliques in the core as seeds of link farms, and showed that the enumeration of large maximal cliques was effective for detecting seeds of link farms. Finally, we expanded these seed set by a minimum cut technique. We found about 0.6 million spam sites in SCCs around the core. In addition that, we extracted 8 thousand sites as spams by the maximal clique enumeration and obtained further 49 thousand sites in the core. Thus, we found about 57 thousand sites as spams with very high precision in the core. We observed that almost all of these link spam is made by sites written in English.

However, we can observe that there still remain many link farms in the core by simple sampling. Thus, it is an important issue to propose further methods that improve our link based approach.

With maximal clique enumeration, we can find only a small portion of hidden link farms in the core, because we ignore the nodes with high degree due to the computational time. Another reason is that cliques are rather strict criterion for denseness. Hence, it is an important issue to introduce more flexible definition for denseness of a subgraph.

There is still much room for improvement for our naive minimum cut approach because there are several alternatives for our choice of arc capacities and seed sets.

It is also an important issue to apply our method towards other benchmark data.

8. ACKNOWLEDGMENTS

We thank Takeaki Uno for providing us with an implementation of the maximal clique enumeration algorithm.

9. REFERENCES

[1] L. Becchetti, C. Castillo, and D. Donato. Link-based characterization and detection of web spam. In *Proc. of AIRWEB 2006, Seattle*, 2006.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Using rank propagation and probabilistic counting for link-based spam detetection. In *Proc. of KDD 2006, Philadelphia, Pennsylvania*, 2006.

[3] A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. In *Proc. of AIRWEB 2006, Seattle*, 2006.

[4] A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. Spamrank – fully automatic link spam detection. In *Proc. of AIRWEB 2005, Chiba*, 2005.

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

[6] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. of KDD 2000, Boston*, 2000.

[7] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proc. of VLDB 2005, Trondheim*, 2005.

[8] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. of AIRWEB 2005, Chiba*, 2005.

[9] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proc. of VLDB 2004, Toronto*, 2004.

[10] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46:119–130, 1997.

[11] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *Proc. of AIRWEB 2006, Seattle*, 2006.

[12] K. Makino and T. Uno. New algorithms for enumerating all maximal cliques. In *SWAT 2004, Humlebaek*, 2004.

[13] P. T. Metaxas and J. DeStefano. Web spam, propaganda and trust. In *Proc. of AIRWEB 2005, Chiba*, 2005.

[14] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proc. of WWW 2001, Hong Kong*, 2001.

[15] T. Ono, M. Toyoda, and M. Kitsuregawa. An examination of techniques for identifying web spam by link analysis. In *Proc. of DEWS 2006, Tokyo*, 2006.

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

[17] B. Wu and B. Davidson. Identifying link farm spam pages. In *Proc. of WWW 2005, Tokyo*, 2005.