# Identifying Web Spam by Densely Connected Sites and its Statistics in a Japanese Web Snapshot

Hiroshi Ono, Masashi Toyoda, Masaru Kitsuregawa
Institute of Industrial Science
The University of Tokyo
Komaba 4-6-1, Meguro-ku, Tokyo, 153–8505 Japan
{h-ono,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

## Abstract

*Web spamming refers to actions intended to mislead search engines into ranking certain pages higher than they deserve. Recently, the amount of web spam has increased dramatically, leading to a degradation of search results. One of the most effective spamming techniques is link spamming. This is done by setting up an interconnected structure of pages for deceiving link-based ranking methods, such as PageRank. In this paper, we analyze distributions of link spam in our archive of Japanese web pages using link analysis techniques.*

## 1 Introduction

Search engines have become indispensible when accessing information on the web. It has become crucial for businesses to have their web pages shown on the top of query results lists. Methods known as SEO (Search Engine Optimization) are employed to improve rankings. These methods include optimizing page contents, and site structure.
There are some cases in which SEO methods are misused to mislead search engines and to acquire higher rankings than appropriate. Such activity is called web spam, and entities involved in web spam are called spammers. Web spam causes unrelated pages to be displayed in search results, and results in bias in search results, and degrading information quality.
Web spam can be classified largely in two techniques. The first is to adjust text on the page to match queries in search engines. The other is to manipulate link structure in the vicinity of one's site, in order to raise the ranking results. This is called link spam and is used against search engines which rank web sites by analyzing link structure. This is done by creating many sites which have a dense link structure. Currently, many search engines use link structure analysis as an important factor in ranking web pages. As a result, link spam is used frequently, and appropriate measures to deal with it are needed. Our goals are to examine how widely link spam is used, and to develop anti-spam measures.
Two major link analysis methods which which are targeted by link spam are Pagerank[2], and HITS[3].
In PageRank, the score of a page is calculated by the sum of all scores which have links to that page. A type of link spam which targets the PageRank algorithm is one where many sites cooperate and link to each other, or a group of sites owned by the same party with dense links between them. This is called a link farm. We examine link farms by extracting maximum cliques, and propose a method for extracting large link farms by extracting approximate maximum cliques.
HITS assigns a hub score and an authority score to each web page. Under the HITS algorithm, an important hub page is one that links to important authority pages. An important authority page is one that is linked to by many hub pages. A search engine using the HITS algorithm returns both high ranking hub pages and authority pages as a result. In order to obtain a high authority score, many links must be obtained from highly ranking hub sites. Link spam which targets HITS algorithm is described as follows.
Some web directory services permit free link registration. By submitting links to these web directory services, spammers can create links to target pages. Web directories usually have a high PageRank score and hub score, so this method results in a boosted PageRank score and authority score. We examine spam which use link farms and also target the HITS algorithm.
The rest of the paper is organized as follows. Methods used in this work for extracting web spam is explained in Section 2. Section 3 describes the web archive using in this work, and how the sitegraph used in the experiments are constructed. Section 4 shows the results of the experiment

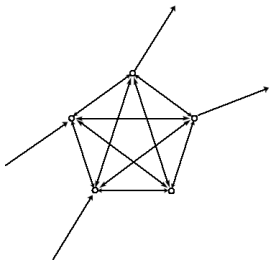

**Figure 1. A clique of size $5$.**

for extracting maximum cliques from the web archive. Section 5 shows the results of the experiment for extracting approximate maximum cliques. Section 6 shows the results for the experiment of extracting web spam which uses both link-farm type spam structure and hub-authority type spam structure. Finally, we conclude in Section 7.

## 2 Web Spam Extraction Methods

In this work, we examine web spam structure focusing on link farms and cliques. These cliques appear on the web graph when link spam is performed by registering links in web directories. The web graph is a directed graph with each website as a node, and links between websites as edges. This is called a site-graph. In a page sized graph, link exchanges between multiple pages are difficult to extract, we use a site as a unit. Following are the spam extraction methods used in our work.

### 2.1 Extraction of Link-farm Type Web Spam

Sites incorporating link-farms are connected densely on the site-graph. If a non-directed graph is extracted from the site-graph such that all edges exist only when there are mutual links between 2 sites, almost all link farms would contain cliques. A clique is a subgraph in which all nodes are mutually connected to each other by edges. As an example, a clique of size $5$ is shown in Figure 1. We call web spam using link farms "link-farm type spam".

#### 2.1.1 Web Spam Extraction Using Maximum Clique Enumeration

Cliques that are not contained in other cliques are called maximum cliques. By extracting maximum cliques, the central structure of a link-farm can be captured. For extracting maximum cliques, we used an algorithm proposed by Makino and Uno [1]. For a graph with $n$ nodes, $m$ edges, and a maximum degree of $\Delta$, using this algorithm, maximum cliques can be enumerated with $O(\Delta^4)$ computation time, and $O(n+m)$ memory. With this algorithm, it was difficult to calculate when the maximum degree became larger than $80$, so we performed the experiments with a site-graph of degrees lower than $80$. In section 4 we extract link farms by extracting maximum cliques from the site-graph, and examine different types of link spam.

#### 2.1.2 Proposed Method for Extracting Approximate Maximum Cliques

Next we clustered sites according to the number of common nodes. This is done by the following steps. First, all edges of the undirected graph are sorted and loaded. Next, for both nodes connected by each edge, nodes that are linked to by both nodes are read, and the number of nodes that are commonly linked to are counted. This can be computed in the order of the maximum degree. If the number of commonly linked nodes is above a certain threshold N, the 2 nodes are clustered in the same set. We used the Union-find algorithm for clustering sets. This is an algorithm to merge disjoint sets. If the merged sets are larger than size N, it is extracted from the site-graph. With this method, we can extract clusters with which are larger than N, and have more than N common nodes between each of the nodes. The total amount of computation is in the order of the product of the number of edges and the maximum degree. In Section 5, using this method, we extract link farms which are have similar structure to maximum cliques.

### 2.2 Extraction of Web Spam Which Use Both Link-farm Type Web Spam and Hub-authority Type Web Spam

In link spam using automatic link registration, dense one sided edges will be formed from registration sites to target sites. This structure can be grasped as a bipartite clique. A graph with 2 node sets with all edges connecting them is called a bipartite graph. Of these subgraphs, such a graph that has all nodes in the 2 node sets connected is called a bipartite clique (Figure 2). Link spam using automatic link registration results in a bipartite clique. This form of spam is called "hub-authority type spam". In Figure 2, each of the nodes in vertex set 1 is called a complete hub of vertex set 2. In Section 6 we examine hub-authority type spam by extracting complete hubs regarding the maximum cliques obtained in Section 4.

## 3 Data Set

### 3.1 Japanese Web Archive

The dataset we used for experiments is based on a large-scale crawl of Japanese web pages done in May 2004. The crawler stops collecting pages in a site, if it could not find

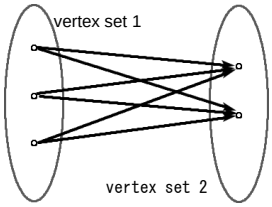


**Figure 2. A complete bipartite graph.**

any Japanese pages in the site after getting a few pages. The archive data consists of 96 million pages, and 4.5 billion links.

### 3.2 Construction of the Site-graph

First, we construct a site-graph from the dataset. As the representative page of each site, pages with in-links of over 3, and whose URL is within 3 tiers (i.e. http://A/B/C/)were chosen. This was done to group together web pages of a single party. This graph had 6.8 million nodes, and 280 million edges. From this site-graph, we extracted an undirected graph in which edges exist only when 2 sites are mutually linked. This undirected graph had 1.6 million nodes, and 39 million edges.

## 4 Extraction of Maximum Cliques

When extracting maximum cliques, many cliques are extracted which have duplicate nodes. Therefore, instead of counting the number of maximum cliques, we calculated the largest size of the maximum cliques in which each site is included. The distribution of number of sites per size of maximum cliques is shown in Figure 3.

Since the seed pages of the original site-graph were selected only if they had more than 3 inlinks, data is shown only for sizes over 3. It can be said that distribution of maximum cliques follow Zipf's law. In the case where maximum degree was set to 80, the total number of nodes constructing the maximum cliques were 600 thousand, which was 37.5% of the nodes in the original site-graph. Next, we randomly chose samples from the maximum cliques, and manually inspected their contents. Results are shown in Table 1. "Link directory sites" are sites whose main content is link information. "Sales promotion sites" are sites which main purpose is to advertise a certain product. Online casinos sites and prize promotion sites have been put in this category. "Non-spam sites" are those which do not correspond to neither link directory sites, sales promotion sites nor sexually explicit sites. This category includes personal sites, corporation sites, and sites of public institutions.

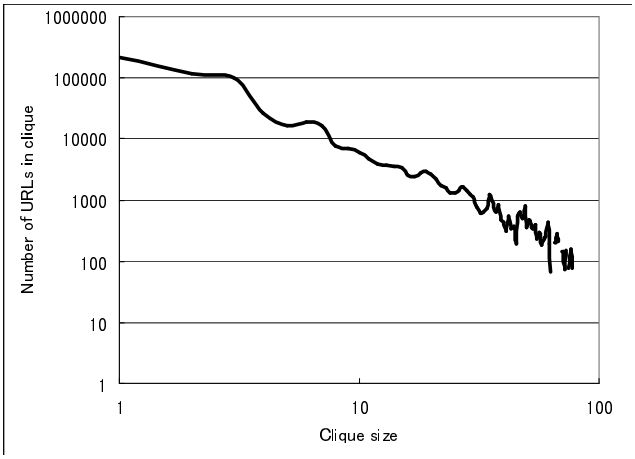


**Figure 3. The distribution of number of sites per size of maximum cliques.**

**Table 1. Spam classification of maximum cliques.**

| Classification | Number of sites | Percentage |
|---|---|---|
| Non-spam | 60 | 17% |
| Sales Promotion | 132 | 37% |
| Link Directory | 55 | 15% |
| Sexually Explicit | 113 | 31% |
| Total | 360 | 100% |

## 5 Extraction of Approximate Maximum Cliques

We call results by clustering sites according to the number of common nodes "approximate maximum cliques". The size distribution of approximate maximum cliques are shown in comparison with those of maximum cliques in Figure 4. The sizes of approximate maximum cliques are larger than those of maximum cliques. This is because the extraction conditions are relaxed in the case of approximate maximum cliques, compared to maximum cliques.

By randomly chosing samples from the approximate maximum cliques, we manually inspected their contents. Results for the same range as maximum cliques are shown in Table 2. The amount of web spam pages was around 68%.

Next we show results for larger sizes of approximate maximum cliques.Results are shown in Table 3. The amoung of spam sites is around 99%, and sexually explicit sites make up the majority.

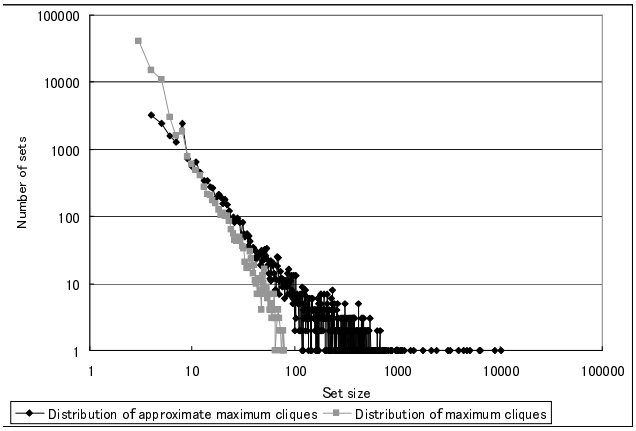


**Figure 4. Size distribution of approximate maximum cliques and maximum cliques.**

**Table 2. Spam classification of approximate maximum cliques.**

| Classification | Number of sites | Percentage |
|---|---|---|
| Non-spam | 100 | 32% |
| Sales Promotion | 83 | 27% |
| Link Directory | 11 | 3.5% |
| Sexually Explicit | 117 | 38% |
| Total | 311 | 100% |

**Table 3. Spam classification of approximate maximum cliques.**

| Classification | Number of sites | Percentage |
|---|---|---|
| Non-spam | 1 | 1% |
| Sales Promotion | 11 | 16% |
| Link Directory | 3 | 4% |
| Sexually Explicit | 54 | 78% |
| Total | 69 | 100% |

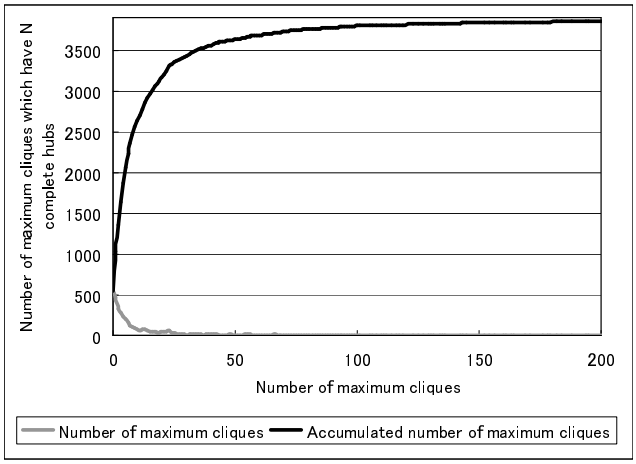


**Figure 5. Number of maximum cliques which have N complete hubs**

## 6 Extraction of Web Spam Which Use Both Link-farm Type Web Spam and Hub-authority Type Web Spam

Of the maximum cliques obtained in Section 4 we calculated complete hubs for those of which had sizes over 10. The distribution of numbers of complete hubs are shown in Figure 5. Maximum cliques which don't have complete hubs account for 14% of the total maximum cliques. More than half of the maximum cliques have over 4 complete hubs. Thus it can be said that most link-farm type spam sites simultaneously use hub-authority type spam structure.

## 7 Conclusion

We extracted spam sites from a large web archive, and examined the distribution of web spam. By extracting approximate maximum cliques, we were able to extract large scale web spam. By manually inspecting these sites, the majority were found to be web spam. Many of the extracted sites simultaneously use link-farm structure and hub-authority structure.

## References

[1] Makino, K., and Uno, T., "New Algorithms for Enumerating All Maximal Cliques," SWAT 2004, LNCS 3111, pp. 260-272, 2004.

[2] Page, L., Brin, S., Motwani, R., and Winograd, T., "The PageRank citation ranking: Bringing order to the web," Tech. rep., Stanford University, 1998.

[3] Kleinberg, J., "Authoritative sources in a hyperlinked environment," Journal of the ACM, 46(5),

[4] Gyongyi, Z. and Garcia-Molina, H. "Web Spam Taxonomy," Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Tokyo, Japan, May 2005.