# A Study of Link Farm Distribution and Evolution using a Time Series of Web Snapshots

Young-joo Chung
chung@tkl.iis.u-tokyo.ac.jp

Masashi Toyoda
toyoda@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
kitsure@tkl.iis.u-tokyo.ac.jp

Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, JAPAN

## ABSTRACT

In this paper, we study the overall link-based spam structure and its evolution which would be helpful for the development of robust analysis tools and research for Web spamming as a social activity in the cyber space. First, we use strongly connected component (SCC) decomposition to separate many link farms from the largest SCC, so called the core. We show that denser link farms in the core can be extracted by node filtering and recursive application of SCC decomposition to the core. Surprisingly, we can find new large link farms during each iteration and this trend continues until at least 10 iterations. In addition, we measure the spamicity of such link farms. Next, the evolution of link farms is examined over two years. Results show that almost all large link farms do not grow anymore while some of them shrink, and many large link farms are created in one year.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Experimentation, Measurements

## Keywords

Link analysis, Web spam, Information retrieval

## 1. INTRODUCTION

Addressing web spam is critical not only for search engines but also for web analysis applications based on Web archives, such as topic tracking, time-frequency analysis of blog postings, and web community extraction. Although the main purpose of web spammers is boosting the ranking of their pages in search results, their spamming techniques also confuse various methods for web analysis. For example, when we use link-based community extraction methods such as HITS [1] and trawling [6], the results would include many link farms, densely connected Web pages created intentionally for boosting PageRank [2]. Term spam which stuffs numerous keywords artificially in pages can easily contaminate the result of time-frequency analysis of terms in the Web.

We have developed a Socio-Sense system [3] to analyze social activities and behaviors from our web archive with Japanese-centric Web contents crawled for 9 years (10 billion pages in total). The main users of the system are sociologists and marketing people who are interested in how the Web evolves according to activities in the real and cyber worlds. The system makes it possible to observe and track trends on topics, by providing Web structural analysis tools, such as a relation map of link-based web community [4], temporal analysis of community evolution [5]. For eliminating spam from those results, and for developing robust analysis tools, it is important to understand the overall structure of spam sites and their evolution. On the other hand, the evolution of spam itself is a fascinating social activity in the cyber space that might be researched by sociologists.

In this paper, we study the overall distribution and evolution of link farms in a large host graph of the Japanese Web crawled in 2004, 2005, and 2006. In our previous work [7], using a single Web snapshot of 2004, we applied strongly connected component(SCC) decomposition algorithm to the Web graph, and showed that almost all of large SCCs, except for the largest SCC so called the core, are link farms.

We expanded out previous work to examine the distribution of denser link farms in the core. That is, we prune nodes with small degrees from the core, and apply SCC decomposition to the pruned core recursively with increasing degree threshold. After extracting link farms, we evaluated the spamicity of them. Next, the evolution of these link farms is examined over two years. We show that almost all large link farms do not grow anymore, and most of them are created in one year.

The rest of this paper is organized as follows. In Section 2, we review previous work related with our study. Section 3 describes datasets. In Section 4, the experimental results are presented. Finally we summarize and conclude our work in Section 5.

## 2. PREVIOUS WORK

Link spamming is one of the web spamming techniques that try to mislead link-based ranking algorithms such as


*AIRWeb* '09, April 21, 2009 Madrid, Spain.

PageRank [2] and HITS [1]. Since these algorithms consider a link to pages as an endorsement for that page, spammers create numerous false links and construct an artificially interlinked link structure, so called a spam farm, to centralize link-based importance to their own spam pages [11].

To understand the web spamming, Gyöngyi et al. described various web spamming techniques in [10]. Optimal link structures to boost PageRank scores are also studied to grasp the behavior of web spammers [11]. Fetterly et al. found out that outliers in statistical distributions are very likely to be spam by analyzing statistical properties of linkage, URL, host resolutions and contents of pages [8].

To demote link spam, Gyöngyi et al. introduced TrustRank [12] that is a biased PageRank where rank scores start to be propagated from a seed set of good pages through outgoing links. By this, we can expect spam pages to get low rank. Optimizing the link structure is another approach to demote link spam. Carvalho et al. proposed the idea of noisy links, a link structure that has a negative impact on link-based ranking algorithms [15]. Qi et al. also estimated the quality of links by similarity of two pages [16].

To detect link spam, Benczúr et al. introduced SpamRank [13]. SpamRank checks PageRank score distributions of all in-neighbors of a target page. If this distribution is abnormal, SpamRank regards a target page as a spam and penalizes it. Becchetti et al. employed link-based features for the link spam detection. They built a link spam classifier with several features of the link structure like degrees, link-based ranking scores, and characteristics of out-neighbors. [14]

Saito et al. employed a graph algorithm [7] to detect link spam. They decomposed the Web graph into strongly connected components and discovered that large components are spam with high probability. Link farms in the core were extracted by maximal clique enumeration. This work is similar to ours in the respect that both apply SCC decomposition algorithm on the Web, but we introduced a recursive SCC decomposition to extract spam structures in the core instead of clique enumeration. Moreover, we observed the change in spam components extracted from the time series of the Web snapshots which has never been explored as far as we know.

# 3. DATASET

We used two different datasets for our experiments. The first set is a set of large scale snapshots of Japanese Web archive. These snapshots are built by crawling that conducted from 2004 to 2006. Basically, our crawler is based on the breadth first crawling, except that it focuses on pages written in Japanese. We collected pages outside the .jp domain if they were written in Japanese. 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, our snapshot contains pages written in various languages such as English. Our crawler does not have an explicit spam filter while it detects mirror servers and tries to crawl only representative servers. Therefore, our archive includes spam hosts without mirroring.

In this paper, we will use a host graph, where each node is a host and each edge between nodes is a hyperlink between pages in different hosts. Host graphs for 2004, 2005 and 2006 were built. In each graph, we included only hosts that existed in the 2006 archive, and did not consider hosts disappeared from 2004 to 2005. This is because we could not distinguish whether those hosts really disappeared or they were just not reached by our crawler. As a result, we focus on the growth rate of link farms that had existed for two or three years. The properties of our Web snapshot are shown in Table 1.

The second set is WEBSPAM-UK datasets. These are public datasets achieved by crawling .uk in May 2006 and May 2007 [18]. Both labeled and unlabeled hosts are included in them.[1] Due to the several differences between 2006 and 2007 datasets, we did not examine the evolution of link farms.

**Table 1: The Properties of the Japanese host graph**

| Year | 2004 | 2005 | 2006 |
|---|---|---|---|
| Number of nodes(hosts) | 2.98M | 3.70M | 4.02M |
| Number of edges | 67.96M | 83.07M | 82.08M |

**Table 2: The Properties of the WEBSPAM-UK host graph**

| Year | 2006 | 2007 |
|---|---|---|
| Number of nodes(hosts) | 11,402 | 114,529 |
| Number of nodes(hosts) | 11,402 | 114,529 |
| Number of edges | 730,774 | 1,836,441 |
| Number of labeled hosts | 10,662 | 6,479 |

# 4. EXPERIMENTS

In this section, we introduce our approach to extract links farms from the large-scale Web graph, and describe the details of link farms in two different datasets (See Section 3). We evaluate the spamicity of such link farms and observe their changes through time.

## 4.1 Strongly Connected Component Decomposition with Node Filtering

In order to extract the link spam structure, we decomposed the host graph into strongly connected components (SCCs), where every pair of nodes has a directed path between them. SCCs of a directed graph are maximal strongly connected subgraphs. The result of SCC decomposition of the Web graph is known to include the largest SCC (so called the core) with about 30% of all nodes, and many smaller SCCs [9]. Since spam sites construct a densely connected link structure [11], and links between spam and normal site seldom exist, it can be expected that spam sites form a SCC. Our previous work [7] confirmed that 95% of SCCs around the core whose size is over 100 are link farms, but we could not efficiently find denser link farms left in the core.

We expand the previous work by introducing a recursive SCC decomposition with node filtering. That is, we prune nodes with small degrees from the core, and apply SCC decomposition to the pruned core recursively with increasing degree threshold. That is, after we decompose the whole host graph into SCCs, we filter out nodes in the core whose in-degree and out-degree are smaller than 2, and decompose the remaining hosts in the core again. As a result, we can extract denser SCCs in the core. Next, we consider the largest among newly obtained SCCs, and discard nodes

[1]As for UKSPAM-2006, we used Version 2.0. data. 93.5% of total hosts was labeled. 8,123 hosts were normal, 2,113 were spam and 426 were undecided.

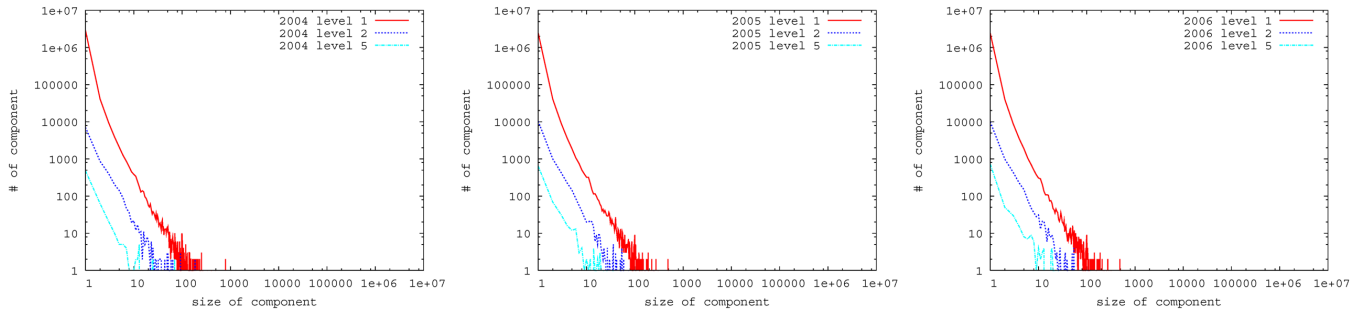


**Figure 1: SCC size distribution of each year. Each graph shows the size distribution of different level SCCs.**

of which in and out degrees are smaller than 3, and apply the decomposition algorithm to the remaining hosts. This process is performed recursively with incrementing degree threshold, and continued while we have large SCCs in the results. Here is terminology we will use in this paper.

**Core** Core is the largest strongly connected component obtained by SCC decomposition of the graph.

**Level 1 graph** Level 1 graph is the whole host graph.

**Level n graph** Level $n$ graph consists of nodes in the core of level $n-1$ graph whose in and out degree are more than $n$.

**Level n SCC** Level $n$ SCC is the strongly connected components obtained by decomposing level $n$ graph.

## 4.2 Strongly Connected Components of Japanese Web Archive

### 4.2.1 *Size Distribution of Strongly Connected Components*

The results of the decomposition of the whole hosts, level 2 and 5 graphs are described in Table 3, 4 and 5. The fraction of the core size to whole nodes in each level graph increases drastically between level 1 and level 2, in archives of all years. From level 5 to 10, the fraction is relatively stable, so we verified SCCs obtained until 10 iterations.

Figure 1 shows the size distributions of different level SCCs of each year. As the Figure indicates, the size distribution of SCCs follows the power law, which agrees with Broder et al. [9]. Moreover, size distributions of SCCs from different level graphs are also similar in the power-law exponents. (See Table 7.) Note that an abnormal distribution appears at the tail of each distribution graph. Such phenomenon is clear particularly in SCCs with over 100 hosts. We measured their spamicity and discovered large SCCs with over 100 hosts are very likely to be spam. Details of measurement will be explained in Section 4.2.2.

Figure 2, 3, 4 illustrate the overall structure of level 1 and level 2 SCCs for each year. The left hand side depicts level 1 SCCs and the right one is for level 2 SCCs. A big gray node represents a core, black nodes represent SCCs with over 100 nodes, and white nodes represent smaller SCCs that connect large SCCs. The size of a node represents the number of hosts included in the SCC. Two SCCs are connected by a directed edge when hyperlinks exist between hosts in SCCs at both ends. Each edge starts from the thick end and goes to the thin end.

When comparing left and right sides of Figures, we can see the similar structure appears in the decomposition result of both the level 1 and level 2 graph. In addition, most large SCCs are directly connected to the core. Some large SCCs

**Table 3: The result of level 1 SCCs**

| Year | 2004 | 2005 | 2006 |
|---|---|---|---|
| # of nodes | 2,978,223 | 3,702,029 | 4,017,250 |
| # SCCs | 1,888,550 | 2,188,035 | 2,483,446 |
| Size of the core | 749,166 | 1,271,253 | 1,245,152 |
| (%) | 25.15 | 34.34 | 31.0 |

**Table 4: The result of level 2 SCCs**

| Year | 2004 | 2005 | 2006 |
|---|---|---|---|
| # of nodes | 556,190 | 949,742 | 918,826 |
| # SCCs | 9,055 | 12,633 | 12,182 |
| Size of the core | 520,554 | 890,703 | 872,269 |
| (%) | 93.60 | 93.78 | 95.00 |

**Table 5: The result of level 5 SCCs**

| Year | 2004 | 2005 | 2006 |
|---|---|---|---|
| # of nodes | 302,613 | 517,057 | 499,031 |
| # of SCCs | 612 | 830 | 899 |
| Size of the core | 301,120 | 512,370 | 495,451 |
| (%) | 99.51 | 99.10 | 99.28 |

**Table 6: The result of level 10 SCCs**

| Year | 2004 | 2005 | 2006 |
|---|---|---|---|
| # of nodes | 196,218 | 329,990 | 315, 644 |
| # SCCs | 127 | 135 | 215 |
| Size of the core | 195,926 | 329, 290 | 314,950 |
| (%) | 99.85 | 99.79 | 99.78 |

**Table 7: Exponent of SCC size distributions**

| Year/Level | 1 | 2 | 5 |
|---|---|---|---|
| 2004 | -2.50 | -2.50 | -2.67 |
| 2005 | -2.44 | -2.60 | -2.52 |
| 2006 | -2.45 | -2.54 | -2.29 |

form larger link farms by connecting to other large SCCs. We also checked how level 1 SCCs are connected to level 2 SCCs. Surprisingly, we found that most of level 1 SCCs are connected directly to the core of the level 2 graph. This means that most link farms are created independently.

#### 4.2.2 Spamicity of Strongly Connected Components

After extracting SCCs, we evaluated whether they are likely to be a link farms. For spamicity measurement, we used hostname properties based on the study of Fetterly et al. [8] and Becchetti et al. [14]. We used two metrics; hostname length and spam keywords in a hostname. Spammers tend to generate long URLs like `"sample-job-reference-letters.974.us"` and stuff terms like `porn,casino,cheap,download` in URLs. Since these metrics do not guarantee perfect spam detection as manual classification, we performed the manual classification on large SCCs when they have low spamicity. Average hostname length of SCC members and the percentage of members with a hostname containing spam keywords were computed. Spam keywords were obtained as follows. First, we extracted hostnames from SCCs in the 2004 archive, of which cardinality is over 1,000. These hostnames are split into words by non-alphabetic characters, such as periods, dashes and digits. Then, we made a frequency list of these words and chose manually 114 words from 1,000 words with high frequency. Our spam keyword list contains words in English, Spanish, Italian, French and Japanese so that it could cover most spam hostnames in various languages. We counted hostnames that contain more than one spam keyword. Hostnames whose first field contains only non-alphabetic words such as dashes and digits are also counted. The percentage of spam members was obtained, by dividing the number of spam hostnames with the total number of hostnames in a SCC. For all nodes in the dataset, the average hostname length was 24.25, and the percentage of hostnames that contain spam keywords are 8.97%. Figure 5 and 6 show the results of the measurement. We used log-scale only on x axis, which represents the size of a SCC. The spamicity of SCCs except the core was examined from different level graphs.

We can observe that as the size of a SCC increases, the hostname length and spam keyword ratio also increase. This indicates that most SCCs with relatively large size (especially, over 100) have very high spamicity. This agrees with the result of [7]. As for SCCs in deep level graphs, although overall spamicity decreased, large SCCs still have high spamicity. Some large SCCs with low spamicity appeared, so we assessed them manually and found out they are spam. For example, in the right graphs in Figure 5 and 6, large SCCs in 2004 show very low spamicity. However, after the investigation, we found out hostnames in these SCCs are also spam, which are very short and consist of a series of spam keywords without any non-alphabetic characters(e.g. `"www.dvdporno.net"`). As for data of 2006 with short hostname length and relatively low spamicity, members with short host names either including meaningless digits and characters like `"www.ib5.x1024.com"`, or containing only spam keywords appeared. Table 8 shows the number of SCCs with over 100 hosts and the number of hosts in them.

To confirm whether the tendency that large SCC are very likely to be a spam structure continues in the depth of the core, we investigated SCCs whose size over 100 in from level 5 to level 10 graphs. Details are described in Table 9. We can see that such a trend remains even when we perform SCC decomposition on nodes in deeper levels.

**Table 8: Number of SCCs (size over 100)and hosts in them**

| Year/Level | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 2004 | # SCCs | 228 | 24 | 7 | 9 | 2 |
| | # hosts | 182285 | 18650 | 9306 | 5032 | 242 |
| 2005 | # SCCs | 167 | 32 | 18 | 13 | 7 |
| | # hosts | 95347 | 38111 | 8236 | 15566 | 2789 |
| 2006 | # SCCs | 180 | 26 | 21 | 6 | 8 |
| | # hosts | 146015 | 26127 | 11092 | 9084 | 1499 |

**Table 9: Number of link farms among SCCs (size over 100), in deep level graphs**

| Year/Level | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| 2004 Spam / Total | 2/2 | 1/2 | 1/2 | 1/1 | 2/2 | 0/0 |
| 2005 Spam / Total | 6/7 | 3/3 | 3/3 | 1/1 | 1/1 | 1/1 |
| 2006 Spam / Total | 8/8 | 2/2 | 3/3 | 1/1 | 1/1 | 0/0 |

**Table 10: The result of SCC decomposition of WEBSPAM-UK**

| Year | 2006 | | 2007 | |
|---|---|---|---|---|
| Level | 1 | 2 | 1 | 2 |
| # of nodes | 11,402 | 7,266 | 114,529 | 45,565 |
| # SCCs | 2,935 | 574 | 54,822 | 969 |
| Size of the core | 7,945 | 6,683 | 59,160 | 44,564 |
| (%) | 69.68 | 91.98 | 51.66 | 97.80 |
| Size of 2nd largest SCC | 73 | 6 | 8 | 3 |

### 4.3 Strongly Connected Components of WEBSPAM-UK Dataset

We applied SCC decomposition on WEBSPAM-UK dataset (see Section 3) , and results are very different from those of Japanese dataset. In Table 10, we can see the fraction of nodes in the largest SCCs are larger than those of Japanese datasets, and the size of other SCCs are far smaller than 100.

In WEBSPAM-UK2006, we found out 21 SCCs whose cardinality is 10 or more are suspicious. Hosts in those SCCs are classified with existing labels. Manual classification is also performed using Wayback machine [19] data of a correspondent period, if a host is labeled as "undecided" or is not labeled. As a result, we verified that 230 among 293 hosts in level 1 SCCs with size 10 or more are spam. We found a two link farms where most members are labeled as normal. One of them contained 14 hosts of a online shopping mall which uses different domains for each category. The other contains 38 hosts and all of them referred to used-car shopping and have similar hostnames like `"www.used-fordcars.co.uk","www.used-suzukicars.co.uk","www.used-daewoo-cars.co.uk"`. If we regard these two link farms as a spam, total 282 host among 293 hosts are spam. In addition to this, both level 1 and level 2, the largest SCCs except the core are composed of spam hosts.

Figure 7 shows the percentage of hosts labeled as spam in each SCC of a different size. As the Figure shows, some large SCCs have relatively low spamicity. These SCCs are the link farms that we explained.

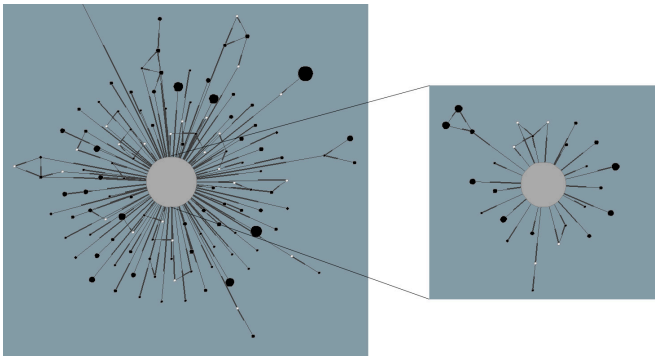

**Figure 2: Connectivity of level 1 and level 2 SCCs in 2004**

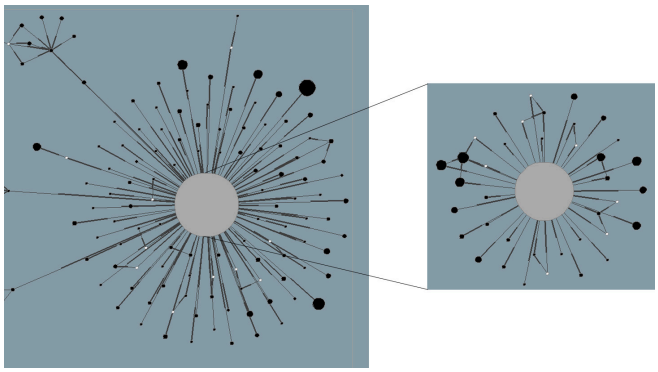

**Figure 3: Connectivity of level 1 and level 2 SCCs in 2005**

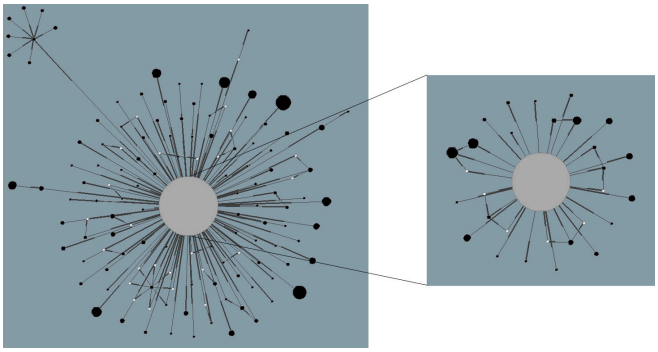

**Figure 4: Connectivity of level 1 and level 2 SCCs in 2006**

WEBSPAM-UK2007 data is very different from 2006 dataset in size and connectivity. Although the size of 2007 dataset is about ten times larger than that of 2006, we found that 2007 dataset consists of many smaller SCCs. The size of the second largest SCC was 8, which is much smaller than that of 2006.

## 4.4 Evolution of Link Farms

After we confirmed that a large SCC is likely to be a spam farm, the evolution of SCCs of the entire host graph over time was examined.

We observed changes in the size of a SCC and computed the growth rate of it. We follow the evolution metrics of web communities from [5], but we use SCCs instead of web communities.

In this paper, we consider the growth and shrinkage of strongly connected components. Some notations are introduced for this.

$t_1, t_2, ..., t_n$ : Time when each archive crawled. Time unit of our archives is a year.

$C(t_k)$ : SCC at time $t_k$.

$N(C(t_k))$ : Size or cardinality of a SCC. The number of hosts in a SCC is used.

In order to understand how a single SCC, $C(t_k)$, has evolved, we find out a SCC corresponding to $C(t_k)$ at time $t_{k-1}$. This *corresponding SCC* $C(t_{k-1})$ is a SCC that shares the most members with SCC, $C(t_k)$. In case multiple SCCs exist at $t_{k-1}$ which share the same number of members with $C(t_k)$, we select the largest SCC as the corresponding SCC. The pair of $(C(t_k), C(t_{k-1}))$ is called a *mainline*.

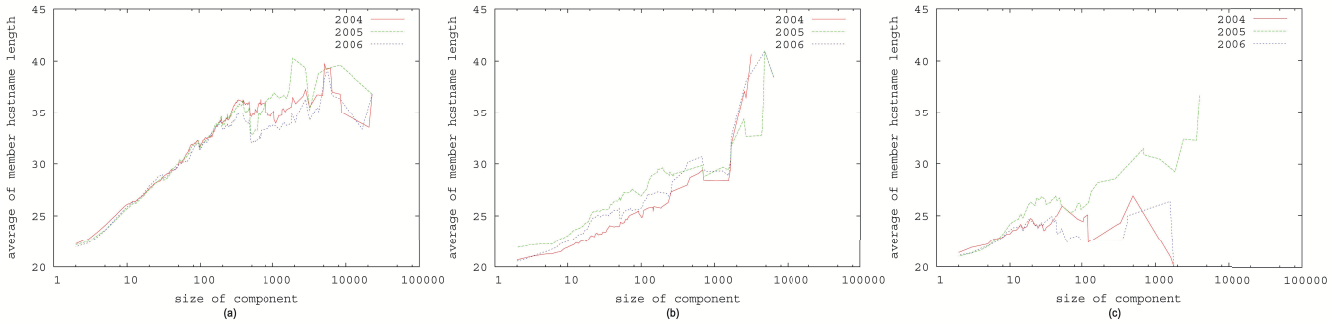


**Figure 5: Average member hostname length. The average hostname length of (a) level 1 SCCs (b) level 2 SCCs, (c) level 4 SCCs. All graph contains the results of year 2004, 2005 and 2006.**

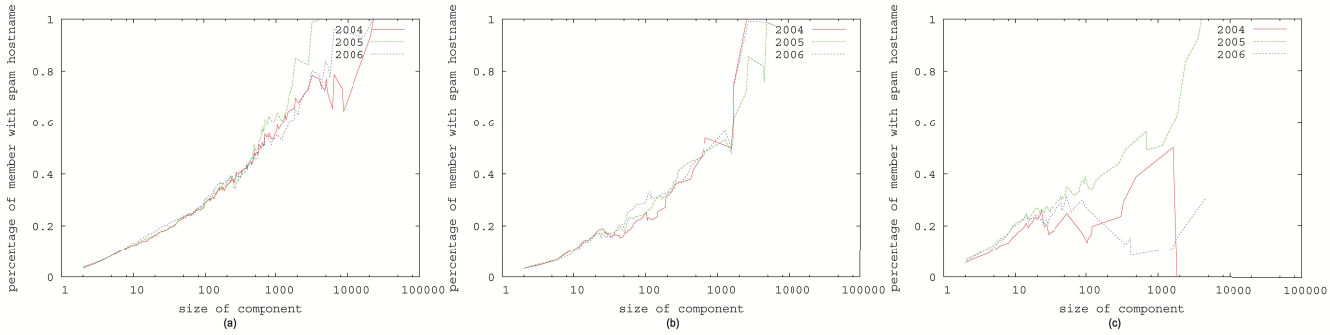


**Figure 6: Rate of members with a spam hostname. The spam hostname rate of (a) level 1 SCCs, (b) level 2 SCCs, (c) level 4 SCCs. All graph contains the results of year 2004, 2005 and 2006.**

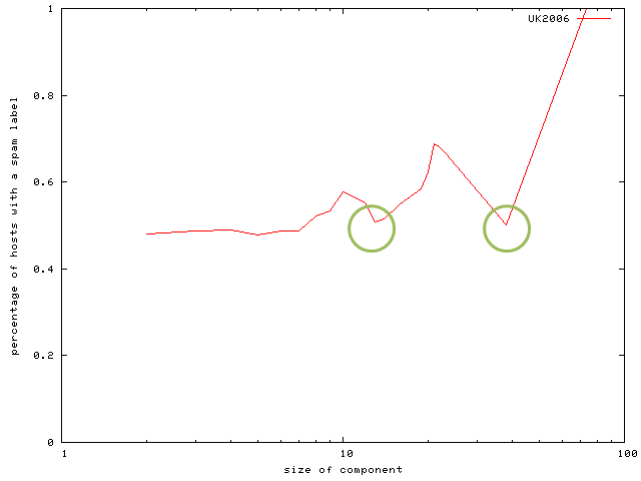


**Figure 7: Rate of members with a spam label**

We observed the change in size and the *growth rate* of mainlines from 2004 to 2005, and from 2005 to 2006. The growth rate of $C(t_k)$ is defined as $N(C(t_k))/N(C(t_{k-1}))$.

Figure 8 and 11 show the change in the SCC size during a year, and Figure 9 and 12 show the growth rate of the SCC size. In all Figures, we can notice the size of most SCCs is stable. Size stability of the SCC becomes stronger as the size of a SCC increases. Considering that most large SCCs are a spam structure, we can expect that a spam farm hardly expands. Note that a few large SCCs shrink significantly, which can be observed in the right-bottom side of Figure 8 and 9. Such decrease in the size could occur when spammers abandon their link farm and consequently the densely connected SCCs split into small ones. More link farms would shrink in reality, since we ignored hosts that disappeared from our host graphs. If we consider disappeared hosts during one year, the shrinkage trend becomes clearer.

Interestingly, we confirmed that the growth rate of relatively small SCCs (from 10 to 100 nodes) follows Gibrat's law. That is, the growth rate of a SCC is independent of its previous size. Gibrat's law have been observed from firm-size growth in economics, and recently some relationships between the power-law distribution of firm size and Gibrat's law is confirmed in [20].

For further understanding of the evolution of spam structures, we investigated the previous size ratio of large SCCs, $N(C(t_{k-1}))/N(C(t_k))$ for $N(C(t_k))$ is over 100. Results are illustrated in Figure 10 and 13 where x axis represents the previous size ratio and y axis represents the number of SCC. Peaks are observed at size ratio 0 and 1, and ratio 0 means all members of $C(t_k)$ were newly appeared at $t_k$, or $C(t_k)$ was emerged from a very small SCC. This suggests that most of large link farms existed in the last year, or emerged in one year.

## 5. SUMMARY

In this paper, we studied the overall link-based spam structure and its evolution from a time series of a large scale Web snapshot. These results could be useful for eliminating major link farms, and designing robust Web analysis methods. First, we proposed recursive SCC decomposition with node filtering for extracting denser and deeper link farms in the core. We showed that in each iteration, almost all large SCCs that contain more than 100 nodes turned out to be a link farm, and surprisingly we could find link farms after pruning large amount of small degree nodes. Using this method, we could extract about from 4.3% to 7.2% of hosts in all years as link farms.

We also examined the change in the size and the growth rate of SCCs over two years to understand the evolution of spam structures. Results show that most large link farms, which are very likely to be spam, did not grow anymore while some of them shrunk, and many large link farms are created within one year. This means that tracking emerging growth of small SCCs is more important than tracking large link farms.

In our experiments, we used rather small subsets of the entire Web, and the crawling interval is still quite long. We are planning to apply our methods to more global Web archives and to crawl the Web more frequently to observe finer-grained evolution.

## 6. REFERENCES


[1] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677, 1998.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th international conference on World Wide Web*, 1998.

[3] M. Kitsuregawa, T. Tamura, M. Toyoda and N. Kaji. Socio-Sense:A system for analysing the societal behavior from long term Web archive, In *Proceedings of 10th Asia-Pacific Web conference*, 2008.

[4] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *Proceedings of the 12th conference on Hypertext and Hypermedia*, 2001.

[5] M. Toyoda and M. Kitsuregawa. Extracting evolution of web communities from a series of Web archive. In *Proceedings of the 14th ACM conference on hypertext and hypermedia*, 2003.

[6] R. Kumar, P. Raghavan S. Rajagopalan and A. Tomkins. Trawling the Web for emerging cyber-Communities. *Proceedings of the 8th international conference on World Wide Web*, 1999.

[7] H. Saito, M. Toyoda, M. Kitsuregawa and K. Aihara. A large-scale study of link spam detection by graph algorithms In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the Web*, 2007.

[8] D. Fetterly, M. Manasse and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam Web pages. In *Proceedings of the 7th International Workshop on the Web and Databases*, 2004.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks, Volume 33, Number 1*, 2000, pp. 309-320.

[10] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the 1st international workshop on Adversarial information retrieval on the Web*, 2005.

[11] Z. Gyöngyi and H. Molina. Link Spam Alliance In *Proceedings of the 31st international conference on Very large Data Bases*, 2005.

[12] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. Combating Web spam with TrustRank. In *Proceedings of the 30th international conference on Very Large Data Bases*, 2004.

[13] A. A. Benczúr, K Csalogány, T Sarlós and M. Uher. SpamRank-fully automatic link spam detection. In *Proceedings of the 1st international workshop on Adversarial information retrieval on the Web*, 2005.

[14] L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates. Link-based characterization and detection of Web spam. In *Proceedings of the 2nd international workshop on Adversarial information retrieval on the Web*, 2006.

[15] A. Carvalho, P. Chirita, E. Moura and P. Calado. Site level noise removal for search engines. In *Proceedings of the 15th international conference on World Wide Web*. 2006.

[16] X. Qi, L. Nie and B. D. Davison. Measuring similarity to detect qualified links, In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the Web*, 2007.

[17] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, 2001.

[18] C. Castillo, D. Donato, L. Becchetti and P. Boldi. A reference collection for Web spam. SIGIR Forum, 40(2), 2006, pp 11-24.

[19] Internet Archive Wayback Machine. `http://www.archive.org`.

[20] Y. Fujiwara, C. Di Guilmi, H. Aoyama, M. Gallegati and W. Souma. Do Pareto-Zipf and Gibrat laws hold true? An analysis with European firms. Physica A(335) ,2004, pp. 197-216.

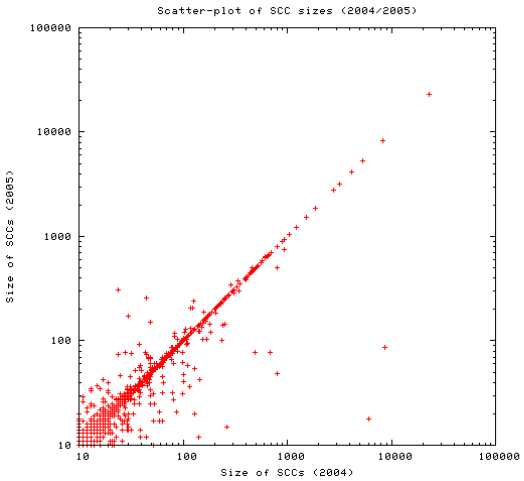


**Figure 8: Evolution of SCC size from 2004 to 2005**

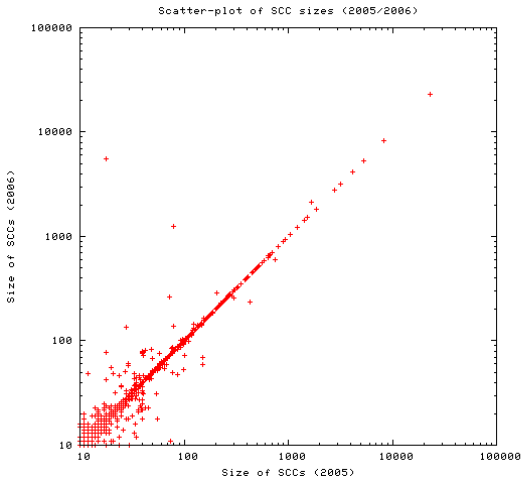


**Figure 11: Evolution of SCC size from 2005 to 2006**

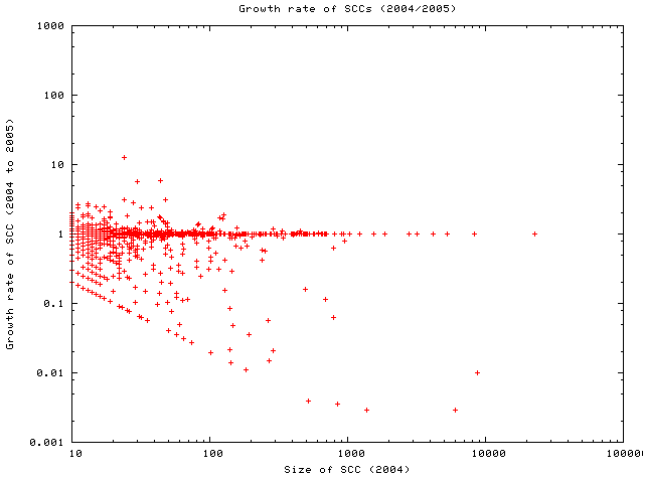


**Figure 9: Growth rate of SCCt size from 2004 to 2005**

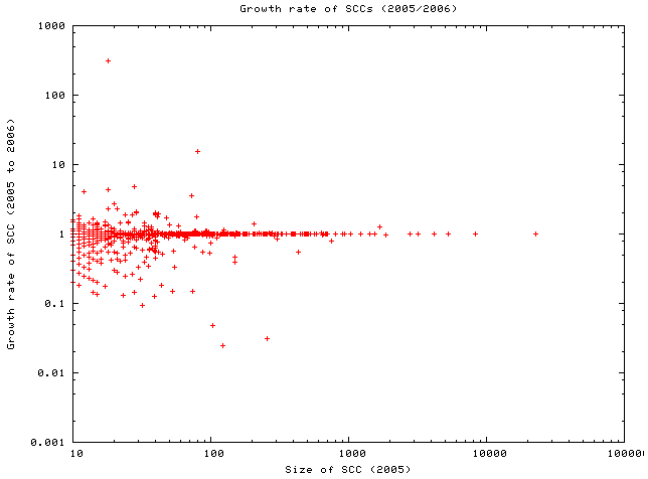


**Figure 12: Growth rate of SCC size from 2005 to 2006**

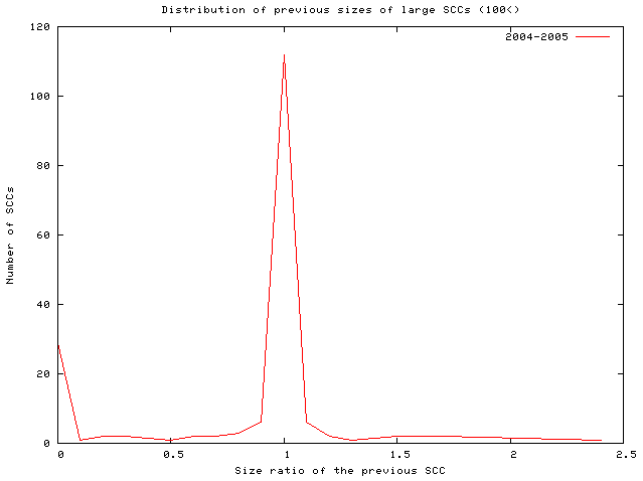


**Figure 10: Distribution of previous size of large SCCs in 2005**

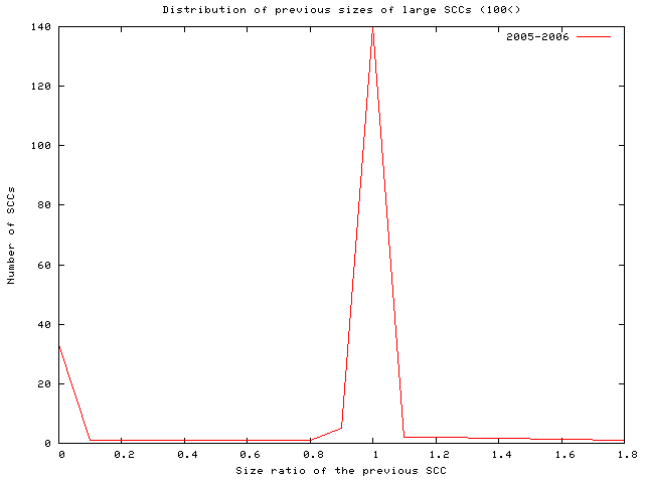


**Figure 13: Distribution of previous size of large SCCs in 2006**