Spam, Damn Spam, and Statistics

Using statistical analysis to locate spam web pages

Dennis Fetterly
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
fetterly@microsoft.com

Mark Manasse
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
manasse@microsoft.com

Marc Najork
Microsoft Research
1065 La Avenida
Mountain View, CA 94043, USA
najork@microsoft.com

ABSTRACT

The increasing importance of search engines to commercial web sites has given rise to a phenomenon we call "web spam", that is, web pages that exist only to mislead search engines into (mis)leading users to certain web sites. Web spam is a nuisance to users as well as search engines: users have a harder time finding the information they need, and search engines have to cope with an inflated corpus, which in turn causes their cost per query to increase. Therefore, search engines have a strong incentive to weed out spam web pages from their index.

We propose that some spam web pages can be identified through statistical analysis: Certain classes of spam pages, in particular those that are machine-generated, diverge in some of their properties from the properties of web pages at large. We have examined a variety of such properties, including linkage structure, page content, and page evolution, and have found that outliers in the statistical distribution of these properties are highly likely to be caused by web spam.

This paper describes the properties we have examined, gives the statistical distributions we have observed, and shows which kinds of outliers are highly correlated with web spam.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia; K.4.m [**Computers and Society**]: Miscellaneous; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Measurement, Experimentation, Algorithms

Keywords

Web characterization, web spam, statistical properties of web pages

1. INTRODUCTION

Search engines have taken pivotal roles in web surfers' lives: Most users have stopped maintaining lists of bookmarks, and are instead relying on search engines such as Google, Yahoo! or MSN Search to locate the content they seek. Consequently, commercial web sites are more dependant than ever on being placed prominently within the result pages returned by a search engine. In fact, high placement in a search engine is one of the strongest contributors to a commercial web site's success.

For these reasons, a new industry of "search engine optimizers" (SEOs) has sprung up. Search engine optimizers promise to help commercial web sites achieve a high ranking in the result pages to queries relevant to a site's business, and thus experience higher traffic by web surfers.

In the best case, search engine optimizers help web site designers generate content that is well-structured, topical, and rich in relevant keywords or query terms. Unfortunately, some search engine optimizers go well beyond producing relevant pages: they try to boost the ratings of a web site by loading pages with a wide variety of popular query terms, whether relevant or not. However, such behavior is relatively easily detected by a search engine, since pages loaded with disjoint, unrelated keywords lack topical focus, and this lack of focus can be detected through term vector analysis. Therefore, some SEOs go one step further: Instead of including many unrelated but popular query terms into the pages they want to boost, they synthesize many pages, each of which contains some tightly-focused popular keywords, and all of which redirect to the page intended to receive traffic. Another reason for SEOs to synthesize pages is to boost the PageRank [11] of the target page: each of the dynamically-created pages receives a minimum guaranteed PageRank value, and this rank can be used to endorse the target page. Many small endorsements from these dynamically-generated pages result in a sizable PageRank for the target page. Search engines can try to counteract such behavior by limiting the number of pages crawled and indexed from any particular web site. In a further escalation of this arms race, SEOs have responded by setting up DNS servers that will resolve any host name within their domain (and typically map it to a single IP address).

Most if not all of the SEO-generated pages exist *solely* to (mis)lead a search engine into directing traffic towards the "optimized" site; in other words, the SEO-generated pages are intended only for the search engine, and are completely useless to human visitors. In the following, we will refer to such web pages as "spam pages". Search engines have an incentive to weed out spam pages, so as to improve the search experience of their customers. This paper describes a variety of techniques that can be used by search engines to detect a portion of the spam pages.

In the course of two earlier studies, we collected statistics on a large sample of web pages. As part of the first

 Seventh International Workshop on the Web and Databases (WebDB 2004), June 17-18, 2004, Paris, France.

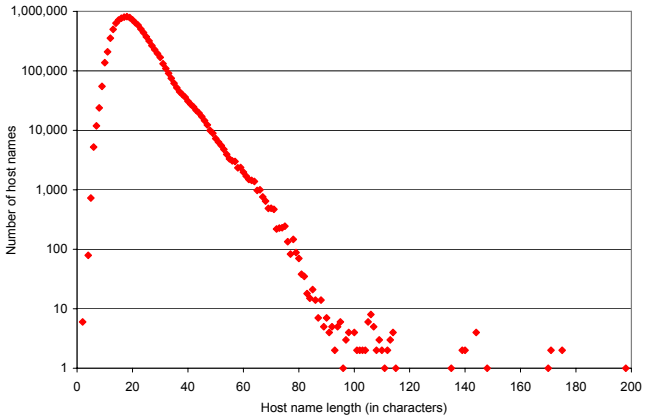

Figure 1: Distribution of lengths of symbolic host names

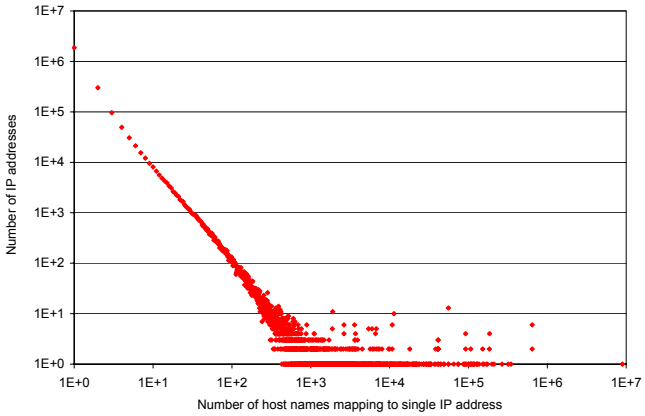

Figure 2: Distribution of number of different hostnames mapping to the same IP address

study [5], we crawled 429 million HTML pages and recorded the hyperlinks contained in each page. As part of the second study [8], we crawled 150 million HTML pages repeatedly, once a week for 11 weeks, and recorded a feature vector for each page allowing us to measure how much a given page changes week over week, as well as several other properties. In the study presented in this paper, we computed statistical distributions for a variety of properties in these data sets. We discovered that in a number of these distributions, outlier values are associated with web spam. Consequently, we hypothesize that statistical analysis is a good way to identify certain kinds of spam web pages (namely, various types of machine-generated pages). The ability to identify a large number of spam pages in a data collection is extremely valuable to search engines, not only because it allows the engine to exclude these pages from their corpus or to penalize them when ranking search results, but also because these pages can then be used to train other, more sophisticated machine-learning algorithms aimed at identifying additional spam pages.

The remainder of the paper is structured as follows: Section 2 describes the two data sets on which we based our experiments. Section 3 discusses how various properties of a URL are predictive of whether or not the page referenced by the URL is a spam page. Section 4 describes how domain name resolutions can be used to identify spam sites. Section 5 describes how the link structure between pages can be used to identify spam pages. Section 6 describes how even purely syntactic properties of the content of a page are predictive of spam. Section 7 describes how anomalies in the evolution of web pages can be used to spot spam. Section 8 discusses how excessive replication of the same (or nearly the same) content is indicative of spam. Section 9 discusses related work, and section 10 offers concluding remarks and outlines avenues for future work.

2. DESCRIPTION OF OUR DATA SETS

Our study is based on two data sets collected in the course of two separate previous experiments [5, 8].

The first data set ("DS1") represents 150 million URLs that were crawled repeatedly, once every week over a period of 11 weeks, from November 2002 to February 2003. For every downloaded page, we retained the HTTP status code, the time of download, the document length, the number of non-markup words in the document, a checksum of the entire page, and a "shingle" vector (a feature vector that allows us to measure how much the non-markup content of a page has changed between downloads). In addition, we retained the full text of 0.1% of all downloaded pages, chosed based on a hash of the URL. Manual inspection of 751 pages sampled from the set of retained pages discovered 61 spam pages, a prevalence of 8.1% spam in the data set, with a confidence interval of 1.95% at 95% confidence.

The second data set ("DS2") is the result of a single breadth-first-search crawl. This crawl was conducted between July and September 2002, started at the Yahoo! home page, and covered about 429 million HTML pages as well as 38 million HTTP redirects. For each downloaded HTML page, we retained the URL of the page and the URLs of all hyperlinks contained in the page; for each HTTP redirection, we retained the source as well as the target URL of the redirection. The average HTML page contained 62.55 links, the median number of links per page was 23. If we consider only distinct links on a given page, the average was 42.74 and the median was 17. Unfortunately, we did not retain the full-text of any downloaded pages when the crawl was performed. In order to estimate the prevalence of spam, we looked at current versions of a random sample of 1,000 URLs from DS2. Of these pages, 465 could not be downloaded or contained no text when downloaded. Of the remaining 535 pages, 37 (6.9%) were spam.

3. URL PROPERTIES

Link spam is a particular form of web spam, where the SEO attempts to boost the PageRank of a web page p by creating many pages referring to p. However, given that the PageRank of p is a function of both the number of pages endorsing p as well as their quality, and given that SEOs typically do not control many high-quality pages, they must resort to using a very large number of low-quality pages to endorse p. This is best done by generating these pages automatically; a technique commonly known as "link spam".

One might expect the URLs of automatically generated pages to be different from those of human-created pages, given that the URLs will be machine-generated as well. For example, one might expect machine-generated URLs to be

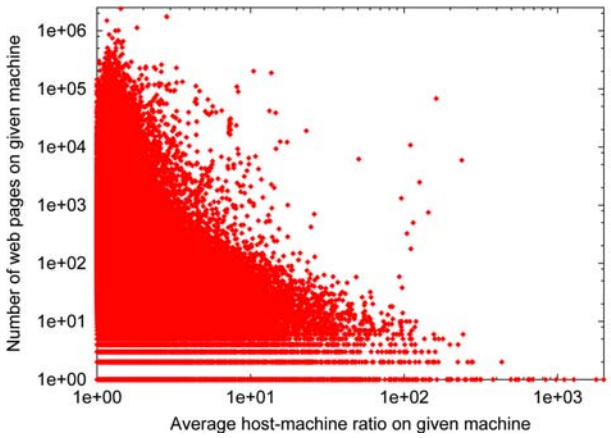

Figure 3: Distribution of "host-machine ratios" among all links on a page, averaged over all pages on a web site

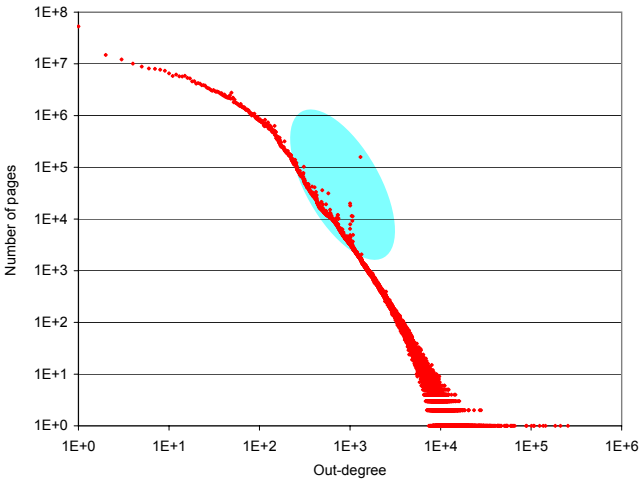

Figure 4: Distribution of out-degrees

longer, have more arcs, more digits, or the like. However, when we examined our data set DS2 for such correlations, we did not find any properties of the URL at large that are correlated to web spam.

However, we did find that various properties of the host component of a URL are indicative of spam. In particular, we found that host names with many characters, dots, dashes, and digits are likely to be spam web sites. (Coincidentally, 80 of the 100 longest host names we discovered refer to adult web sites, while 11 refer to financial-credit-related web sites.) Figure 1 shows the distribution of host name length. The horizontal axis shows the host name length in characters; the vertical axis shows how many host names with that length are contained in DS2.

Obviously, the choice of threshold values for the number of characters, dots, dashes and digits that cause a URL to be flagged as a spam candidate determines both the number of pages flagged as spam as well as the rate of false positives. 0.173% of all URLs in DS2 have host names that are at least 45 characters long, or contain at least 6 dots, 5 dashes, or 10 digits. The vast majority of these URLs appear to be spam.

4. HOST NAME RESOLUTIONS

One piece of folklore among the SEO community is that search engines (and Google in particular), given a query q, will rank a result URL u higher if u's host component contains q. SEOs try to exploit this by populating pages with URLs whose host components contain popular queries that are relevant to their business, and by setting up a DNS server that resolves those host names. The latter is quite easy, since DNS servers can be configured with wildcard records that will resolve any host name within a domain to the same IP address. For example, at the time of this writing, any host within the domain `highriskmortgage.com` resolves to the IP address 65.83.94.42.

Since SEOs typically synthesize a very large number of host names so as to rank highly for a wide variety of queries, it is possible to spot this form of web spam by determining how many host names resolve to the same IP address (or set of IP addresses). Figure 2 shows the distribution of host names per IP address. The horizontal axis shows how many host names map to a single IP address; the vertical axis indicates how many such IP addresses there are. A point at position (x, y) indicates that there are y IP addresses with the property that each IP address is mapped to by x hosts. 1,864,807 IP addresses in DS2 are referred to by one host name each (indicated by the topmost point); 599,632 IP addresses are referred to by two host names each; and 1 IP address is referred to by 8,967,154 host names (far-right point). We found that 3.46% of the pages in DS2 are served from IP addresses that are mapped to by more than 10,000 different symbolic host names. Casual inspection of these URLs showed that they are predominantly spam sites. If we drop the threshold to 1,000, the yield rises to 7.08%, but the rate of false positives goes up significantly.

Applying the same technique to DS1 flagged 2.92% percent of all pages in DS1 as spam candidates; manual inspection of a sample of 250 of these pages showed that 167 (66.8%) were spam, 64 (25.6%) were false positives (largely attributable to community sites that assign unique host names to each user), and 19 (7.6%) were "soft errors", that is, pages displaying a message indicating that the resource is not currently available at this URL, despite the fact that the HTTP status code was 200 ("OK").

It is worth noting that this metric flags about 20 times more URLs as spam than the hostname-based metric did.

Another item of folklore in the SEO community is that Google's variant of PageRank assigns greater weight to off-site hyperlinks (the rationale being that endorsing another web site is more meaningful than a self-endorsement), and even greater weight to pages that link to many different web sites (these pages are considered to be "hubs"). Many SEOs try to capitalize on this alleged behavior by populating pages with hyperlinks that refer to pages on many different hosts, but typically all of the hosts actually resolve to one or at most a few different IP addresses.

We detect this scheme by computing the average "host-machine-ratio" of a web site. Given a web page containing a set of hyperlinks, we define the host-machine-ratio of that page to be the size of the set of host names referred to by the link set divided by the size of the set of distinct machines that the host names resolve to (two host names are assumed to refer to distinct machines if they resolve to non-

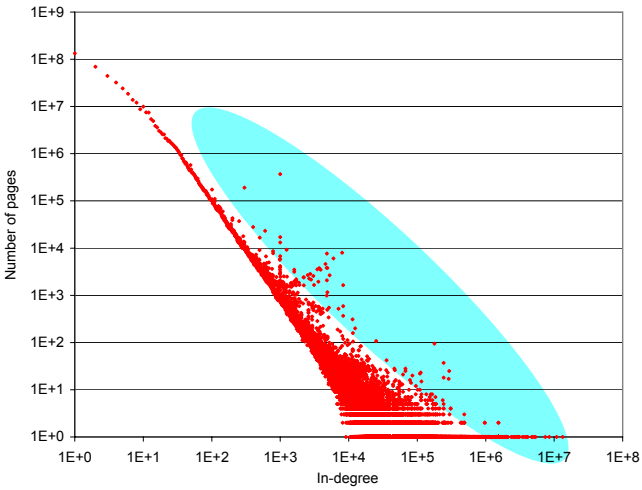

Figure 5: Distribution of in-degrees

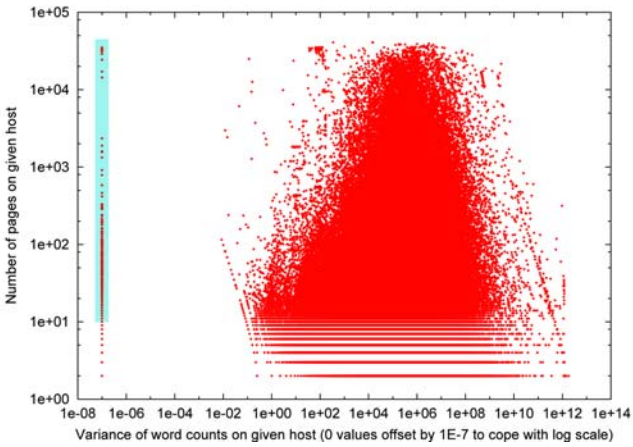

Figure 6: Variance of the word counts of all pages served up by a single host

identical sets of IP addresses). The host-machine-ratio of a machine is defined to be the average host-machine-ratio of all pages served by that machine. If a machine has a high host-machine-ratio, most pages served by this machine appear to link to many different web sites (*i.e.* have non-nepotistic, meaningful links), but actually all endorse the same property. In other words, machines with high host-machine-ratios are very likely to be spam sites.

Figure 3 shows the host-machine ratios of all the machines in DS2. The horizontal axis denotes the host-machine-ratio; the vertical axis denotes the number of pages on a given machine. Each point represents one machine; a point at position (x, y) indicates that DS2 contains y pages from this machine, and that the average host-machine-ratio of these pages is x. We found that host-machine ratios greater than 5 are typically indicative of spam. 1.69% of the pages in DS2 fulfill this criterion.

5. LINKAGE PROPERTIES

Web pages and the hyperlinks between them induce a graph structure. Using graph-theoretic terminology, the out-degree of a web page is equal to the number of hyperlinks embedded in the page, while the in-degree of a page is equal to the number of hyperlinks referring to that page.

Figure 4 shows the distribution of out-degrees. The x-axis denotes the out-degree of a page; the y-axis denotes the number of pages in DS2 with that out-degree. Both axes are drawn on a logarithmic scale. (The 53.7 million pages in DS2 that have out-degree 0 are not included in this graph due to the limitations of the log-scale plot.) The graph appears linear over a wide range, a shape characteristic of a Zipfian distribution. The blue oval highlights a number of outliers in the distribution. For example, there are 158,290 pages with out-degree 1301; while according to the overall distribution of out-degrees we would expect only about 1,700 such pages. Overall, 0.05% of the pages in DS2 have an out-degree that is at least three times more common than the Zipfian distribution would suggest. We examined a cross-section of these pages, and virtually all of them are spam.

Figure 5 shows the distribution of in-degrees. As in figure 4, the x-axis denotes the in-degree of a page, the y-axis denotes the number of pages in DS2 with that in-degree, and both axes are drawn on a logarithmic scale. The graph appears linear over an even wider range than the previous graph, exhibiting an even more pronounced Zipfian distribution. However, there is also an even larger set of outliers, and some of them are even more pronounced. For example, there are 369,457 web pages with in in-degree 1001 in DS2, while according to the overall in-degree distribution we would expect only about 2,000 such pages. Overall, 0.19% of the pages in DS2 have an in-degree that is at least three times more common than the Zipfian distribution would suggest. We examined a cross-section of these pages, and the vast majority of them are spam.

6. CONTENT PROPERTIES

As we mentioned earlier on, SEOs often try to boost their rankings by configuring web servers to generate pages on the fly, in order to perform "link spam" or "keyword stuffing." Effectively, these web servers spin an infinite web — they will return an HTML page for any requested URL. A smart SEO will generate pages that exhibit a certain amount of variance; however, many SEOs are naïve. Therefore, many auto-generated pages look fairly templatic. In particular, there are numerous spam web sites that dynamically generate pages which each contain exactly the same number of words (although the individual words will typically differ from page to page).

DS1 contains the number of non-markup words in each downloaded HTML page. Figure 6 shows the variance in word count of all pages drawn from a given symbolic host name. We restrict ourselves to hosts with a nonzero mean word count. The x-axis shows the variance of the word count, the y-axis shows the number of pages in DS1 downloaded from that host. Both axes are shown on a log-scale; we have offset data points with zero variance by 10^{-7}, in order to deal with the limitations of the log-scale. The blue oval highlights web servers that have at least 10 pages and no variance in word count. There are 944 such hosts serving 323,454 pages (0.21% of all pages). Drawing a random sample of 200 of these pages and manually assessing them showed that 55% were spam, 3.5% contained no text, and 41.5% were soft errors.

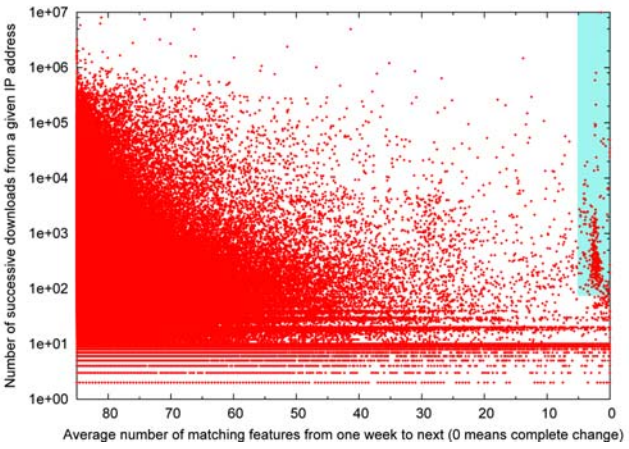

Figure 7: Average change week over week of all pages served up by a given IP address

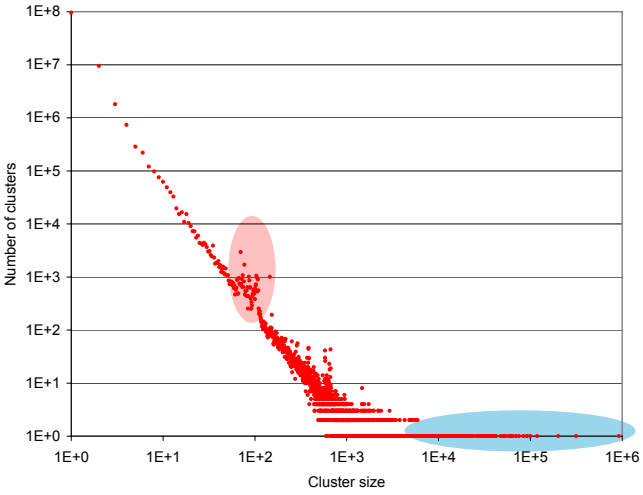

Figure 8: Distribution of sizes of clusters of near-duplicate documents

7. CONTENT EVOLUTION PROPERTIES

Some spam web sites that dynamically generate a page for any requested URL do so without actually using the URL in the generation of the page. This approach can be detected by measuring the evolution of web pages and web sites. Overall, the web evolves slowly, 65% of all pages will not change at all from one week to the next, and only about 0.8% of all pages will change completely [8]. In contrast, spam pages that are created in response to an HTTP request, independent of the requested URL, will change completely on every download. Therefore, we can detect such spam sites by looking for web sites that display a high rate of average page mutation.

Figure 7 shows the average amount of week-to-week change of all the web pages on a given server. The horizontal axis denotes the average week-to-week change amount; 0 denotes complete change, 85 denotes no change. The vertical axis denotes the number of pairs of successive downloads served up by a given IP address (change from week 1 to week 2, week 2 to week 3, etc.). The data items are represented as points; each point represents a particular IP address. The blue oval highlights IP addresses for which almost all pages change almost completely every week. There are 367 such servers, which account for 1,409,353 pages in DS1 (0.93% of all pages). Sampling 106 of these pages and manually assessing them showed that 103 of them (97.2%) were spam, 2 pages were soft errors, and 1 page was a (pornographic) false positive.

One might think that our technique would conflate news sites with spam sites, given that news changes often. However, we did not find any news pages among the spam candidates returned by this method. We attribute this to the fact that most news sites have fast-changing index pages, but essentially static articles. Since we measure the *average* amount of change of all pages from a particular site, news sites will not show up prominently.

8. CLUSTERING PROPERTIES

Section 6 argued that many spam sites serve large numbers of pages that all look fairly templatic. In some cases, pages are formed by inserting varying keywords or phrases into a template. Quite often, the individual pages created from the template hardly vary. We can detect this by forming clusters of very similar pages, for example by using the "shingling" algorithm due to Broder *et al.* [3]. The full details of our clustering algorithm are described elsewhere [9].

Figure 8 shows the distribution of the sizes of clusters of near-duplicate documents in DS1. The x-axis shows the size of the cluster (*i.e.* how many web pages are in the same near-equivalence class), the y-axis shows how many clusters of that size exist in DS1. Both axes are drawn on a log-scale; as so often, the distribution is Zipfian.

The distribution contains two groups of outliers. Examining the outliers highlighted by the red oval did not uncover any spam site; these outliers were due to genuine replication of popular content across many distinct web sites (*e.g.* mirrors of the PHP documentation). However, the clusters highlighted by the blue oval turned out to be predominantly spam: 15 of the 20 largest clusters were spam, accounting for 2,080,112 pages in DS1 (1.38% of all pages).

9. RELATED WORK

Henzinger *et al.* [10] identified web spam as one of the most impiortant challenges to web search engines. Davison [7] investigated techniques for discovering nepotistic links, *i.e.* link spam. Move recently, Amitay *et al.* [1] identified feature-space based techniques for identifying link spam. Our paper, in contrast, presents techniques for detecting not only link spam, but more generally spam web pages.

All of our techniques are based on detecting anomalies in statistics gathered through web crawls. A number of papers have presented such statistics; but focused on the trend rather than the outliers.

Broder *et al.* investigated the link structure of the web graph [4]. They observed that the in-degree and the out-degree distributions are Zipfian, and mentioned that outliers in the distribution were attributable to web spam. Bharat *et al.* have expanded on this work by examining not only the link structure between individual pages, but also the higher-level connectivity between sites and between top-level domains [2].

Cho and Garcia-Molina [6] studied the fraction of pages

on 270 web servers that changed day over day. Fetterly *et al.* [8] expanded on this work by studying the *amount* of week-over-week change of 150 million pages (parts of the results described in this paper are based on the data set collected during that study). They observed that the much higher than expected change rate of the German web was due to web spam.

Earlier, we used that same data set to examine the evolution of clusters of near-duplicate content [9]. In the course of that study, we observed that the largest clusters were attributable to spam sites, each of which served a very large number of near-identical variations of the same page.

10. CONCLUSIONS

This paper described a variety of techniques for identifying web spam pages. Many search engine optimizers aim to improve the ranking of their clients' web sites by trying to inject massive numbers of spam web pages into the corpus of a search engine. For example, raising the PageRank of a web page requires injecting many pages endorsing that page into the search engine. The only way to effectively create a very large number of spam pages is to generate them automatically.

The basic insight of this paper is that many automatically generated pages differ in one way or another from web pages authored by a human. Some of these differences are due to the fact that many automatically generated pages are too "templatic", that is, they have little variance in word count or even actual content. Other differences are more intrinsic to the goal of the optimizers: pages that are ranked highly by a search engine must, by definition, differ from average pages. For example, effective link-spam requires pages to have a high in-degree, while effective keyword spam requires pages to contain many popular terms.

This paper describes a number of properties that we have found to be indicative of spam web pages. These properties include:

- various features of the host component of a URL,
- IP addresses referred to by an excessive number of symbolic host names,
- outliers in the distribution of in-degrees and out-degrees of the graph induced by web pages and the hyperlinks between them,
- the rate of evolution of web pages on a given site, and
- excessive replication of content.

We applied all the techniques that did not require link information (that is, all techniques except for the in- and out-degree outlier detection and the host-machine-ratio technique) in concert to the DS1 data set. The techniques flagged 7,475,007 pages as spam candidates according to at least one technique (4.96% of all pages in DS1, out of an estimated 8.1% $\pm$ 2% true spam pages). The false positives, without excluding overlap between the techniques, amount to 14% of the flagged pages. Most of the false positives are due to imprecisions in the host name resolution technique. Judging from the results we observed for DS2, the techniques that we could not apply to DS1 (since it does not include linkage information) could have flagged up to an additional 1.7% of the pages in DS1 as spam candidates.

Our next goal is to benchmark the individual and combined effectiveness of our various techniques on a unified data set that contains the full text and the links of all pages. A more far-reaching ambition is to use semantic techniques to see whether the actual words on a web page can be used to decide whether it is spam.

Techniques for detecting web spam are extremely useful to search engines. They can be used as a factor in the ranking computation, in deciding how much and how fast to crawl certain web sites, and, in the most extreme scenario, they can be used to excise low-quality content from the engine's index. Applying these techniques enables engines to present more relevant search results to their customers while reducing the index size. More speculatively, the techniques described in this paper could be used to assemble a large collection of spam web pages, which can then be used as a training set for machine-learning algorithms aimed at detecting a more general class of spam pages.

11. REFERENCES

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel and A. Soffer. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In *14th ACM Conference on Hypertext and Hypermedia*, Aug. 2003.

[2] K. Bharat, B. Chang, M. Henzinger, and M. Ruhl. Who Links to Whom: Mining Linkage between Web Sites. In *2001 IEEE International Conference on Data Mining*, Nov. 2001.

[3] A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic Clustering of the Web. In *6th International World Wide Web Conference*, Apr. 1997.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. Graph Structure in the Web. In *9th International World Wide Web Conference*, May 2000.

[5] A. Broder, M. Najork and J. Wiener. Efficient URL Caching for World Wide Web Crawling. In *12th International World Wide Web Conference*, May 2003.

[6] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *26th International Conference on Very Large Databases*, Sep. 2000.

[7] B. Davison. Recognizing Nepotistic Links on the Web. In *AAAI-2000 Workshop on Artificial Intelligence for Web Search*, July 2000.

[8] D. Fetterly, M. Manasse, M. Najork and J. Wiener. A large-scale study of the evolution of web pages. In *12th International World Wide Web Conference*, May 2003.

[9] D. Fetterly, M. Manasse and M. Najork. On the Evolution of Clusters of Near-Duplicate Web Pages. In *1st Latin American Web Congress*, Nov. 2003.

[10] M. Henzinger, R. Motwani, C. Silverstein. Challenges in Web Search Engines. *SIGIR Forum* 36(2), 2002.

[11] L. Page, S. Brin, R. Motwani and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Libraries Technologies Project, Jan. 1998.