Improving Cloaking Detection Using Search Query Popularity and Monetizability

Kumar Chellapilla
Microsoft Live Labs
One Microsoft Way
Redmond, WA, USA 98052
+1 (425) 707-7575
kumarc@microsoft.com

David Maxwell Chickering
Microsoft Live Labs
One Microsoft Way
Redmond, WA, USA 98052
+1 (425) 703-5426
dmax@microsoft.com

ABSTRACT

Cloaking is a search engine spamming technique used by some Web sites to deliver one page to a search engine for indexing while serving an entirely different page to users browsing the site. In this paper, we show that the degree of cloaking among search results depends on query properties such as popularity and monetizability. We propose estimating query popularity and monetizability by analyzing search engine query logs and online advertising click-through logs, respectively. We also present a new measure for detecting cloaked URLs that uses a normalized term frequency ratio between multiple downloaded copies of Web pages. Experiments are conducted using 10,000 search queries and 3 million associated search result URLs. Experimental results indicate that while only 73.1% of the cloaked popular search URLs are spam, over 98.5% of the cloaked monetizable search URLs are spam. Further, on average, the search results for top 2% most cloaked queries are 10x more likely to be cloaking than those for the bottom 98% of the queries.

1. INTRODUCTION

Cloaking is a hiding technique [11] used by some Web servers to deliver one page to a search engine for indexing while serving an entirely different page to users browsing the site. In short, cloaking is the classic “bait and switch” technique applied to the Web. The motivation behind cloaking is to distort search engine rankings in favor of the cloaked page. Cloaking is commonly used in conjunction with other Web spamming techniques. Spammers can present the ultimately intended content to the Web users (without traces of spam on the page), and, at the same time, send a spammed document to the search engine for indexing.

In order for cloaking to be effective, the Web server must be able to detect Web crawler clients reliably. This is typically achieved by examining the client’s: (a) user-agent string, and (b) IP address. A Web server can identify the Web client using the user-agent header in the HTTP request message. A few examples of user-agent strings are:

Internet Explorer:

```
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
```

MSNBot:

```
msnbot/1.0 (+http://search.msn.com/msnbot.htm)
```

GoogleBot:

```
Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)
```

However, the user-agent strings are not strictly standardized and it is really up to the requesting application what to include in the corresponding message field. For example, it is common for less popular Web browsers to mimic the user-agent strings of the dominant browser to ensure a consistent browsing experience. Nevertheless, search engine crawlers do identify themselves by a name distinct from the ones used by traditional Web browser applications.

A very reliable way of identifying the client requesting a Web page is through its IP address. Some spammers maintain lists of IP addresses used by search engines and identify Web crawlers based on their matching IPs. These IP lists are easily available online and are frequently updated.

The differences between the Web pages served to the search engine crawler vs the user typically include: (a) making some text on the page invisible (e.g. white-on-white, very small font size), (b) using style-sheets to hide text, (c) using javascript to alter page content when loaded in the Web browser, and (d) use of javascript or “meta-refresh” to redirect the user to another page.

Since anyone can be an author on the Web, cloaking practices naturally create a question of information reliability. Users accustomed to trusting print media (newspapers and books) may not be able, prepared or willing to think critically about the information obtained from the Web [10]. As a result, most Web search engines do not approve of cloaking and will permanently ban such sites from their databases.

In this paper, we investigate the distribution of cloaking based Web spam over two different query categories, namely popularity and monetizability. Popularity of a query is proportional to the frequency of occurrence in the search query logs. Monetizability can be defined to be proportional to the number of user clicks or the amount of revenue generated by user clicks on sponsored ads (paid advertisements) served alongside search results. Most major search engines serve online ads and keep track of their usage statistics. We mine these logs to obtain popularity and monetizability scores for search queries. In this paper, these

AIRWeb'06, August 10, 2006, Seattle, Washington, USA.

scores are used only to extract the top N queries from search and ad logs.

Section 2 presents some background on Web spam, online advertising, and cloaking. We argue that attracting online users to commercial Web sites for the purposes of increasing their monetization is a significant source of Web spam. Query popularity and monetizability are introduced in Section 3 along with strategies for combating Web spam. Sections 4 and 5 present cloaking detection experiments and their results. Section 6 concludes with a brief discussion of the experimental results presented in this paper and potential future work.

2. BACKGROUND AND RELATED WORK

2.1 Adversarial Aspects of Web Spam

One common definition of Web spam [9,11] is: "A Web page created for the sole purpose of attracting search engine referrals (to this page or some other "target" page)." Owing to such a broad definition, classifying a Web page as spam is inherently ambiguous. In many cases, determining whether a Web page is spam (or not) is ultimately a judgment call. For example, some Web pages have very little useful content, are badly formatted, and are borderline useless, but are not spam. Some other pages look fine in isolation, but in context are clearly spam.

Several approaches based on statistical analysis and machine learning have been proposed for detecting spam pages [4,9,12, 14,19-20]. However, none of them are guaranteed to succeed against all spammers. When search engines counteract Web spam using these approaches, they simply escalate the arms race: the approaches work for a short period of time while the spammers move to more successful and newer strategies.

Detecting Web spam is inherently an adversarial problem. Static machine learning based approaches do not do well against such adversarial problems. Spam classifiers need to be updated often or should be capable of learning online. Online learning requires a constant source of labeled data which can be expensive. Web spam is similar to other common adversarial problems such as e-mail spam and computer viruses that are contained but are not likely to be solved in the near future. The best systems continuously monitor performance and frequently update themselves.

2.2 Motivation behind Web Spam

Search engine optimization (SEO) is a legitimate way to improve traffic to commercial sites. The preferred approach is to ensure that the search engine spider can find and index the site and to improve overall site quality by offering added value to online users. Online advertising can be used to further improve traffic, but requires extra capital investment.

Most major search engines sell advertising keywords. Vendors can bid directly on these advertising keywords and have their commercial links served alongside search results. When search engines serve sponsored ads, they are clearly marked so that users can tell them apart from search results easily. Web spammers on the other hand act as intermediaries and sell search ranks for specific queries to businesses [12]. These ranks are achieved through various spamming techniques. Since these commercial sites are ranked inline with the other genuine search engine results, online users cannot tell them apart.

The overall motivation for most Web spamming approaches is monetizability. Simple conversion ratios such as impression-to-click and click-to-sale numbers determine how profitable an online business is. Many businesses can increase business revenue simply by increasing traffic to their site (all else being the same). The difference between white hat and black hat SEOs is mostly a difference of means rather than the ends. However, exceptions do exist. For example, Google bombers[1] [6] may not be completely motivated by money. However, we believe that non-monetary motivations for Web spam are secondary and as a result not as wide spread.

Similar monetizability arguments have been a rich source of robust approaches for fighting e-mail spam [3,7]. Understanding the monetization strategies of common Web spammers and designing approaches that increase their operating costs is a promising approach to combating Web spam.

2.3 Internet Advertising and the Generalized Second Price Auction

The search engine is not only a tool for searching the Web, but also an advertising platform for ones business and services of companies. Search engines sell online advertising through an auction process where advertisers bid for specific keywords and phrases. A brief description of the Generalized Second Price (GSP) auction [8] is presented below:

When a Web user enters a search query into a search engine, he gets back a page with results, containing both the links most relevant to the query and the sponsored links, i.e., paid advertisements. The presentation ensures that ads are clearly distinguishable from the actual search results. Different searches yield different sponsored links. Advertisers target their ads based on query keywords and/or phrases. For instance, if a travel agent buys the word "Hawaii," then each time a user performs a search on this word, a link to the travel agent will appear on the search results page. When a user clicks on the sponsored link, he is sent to the advertiser's Web page. The user click constitutes a referral to the advertiser from the search engine. The advertiser then pays the search engine for referring the user, hence the name—"pay-per-click" pricing.

The number of ads that the search engine can show to a user is limited, and different positions on the search results page have different desirabilities for advertisers. Preliminary eye tracking studies indicate a triangular region (Golden Triangle) of maximum visibility on the search results page [21]. The golden triangle is a right angled triangle aligned along the top of the first search result and the left side of the results page. It extends from the left top of the results page over to the top of the first result, then down to a point on the left side about three quarters of the way down the page. Generally, this area includes top sponsored links, top organic results and alternative results, including shopping, news or local suggestions. An ad shown at the top of a page is more likely to be clicked than an ad shown at the bottom.

[1] Search engines associate the anchor text that is used to link to a page with that page. By referring to target pages with anchor terms that have a negative connotation, malicious sites can cause these targets to become search results for negative query terms [6].

Hence, search engines need a system for allocating the positions to advertisers, and auctions are a natural choice. Currently, the mechanisms most widely used by search engines are based on GSP.

In the simplest GSP auction, for a specific keyword, advertisers submit bids indicating the maximum price they are willing to pay. When a user enters a keyword, he receives search results along with sponsored links, the latter shown in decreasing order of bids. In particular, the ad with the highest bid is displayed at the top, the ad with the next highest bid is displayed in the second position, and so on. If a user subsequently clicks on an ad in position *k*, that advertiser is charged by the search engine an amount equal to the next highest bid, i.e., the bid of an advertiser in position $k + 1$. If a search engine offered only one advertisement per result page, this mechanism would be equivalent to the standard second price, or Vickrey-Clarke-Groves (VCG), auction [13]. With multiple positions available, the GSP generalizes the second price auction (hence the name). Here, a winner pays the next highest bidder's bid. Modified versions of GSP are used by Google AdWords[2], Yahoo Search Marketing (SM)[3] and MSN AdCenter[4]. For example, one common modification is to combine the advertisers bid price with the expected click-through-rate (CTR) to compute an expected monetization score. Sponsored links are presented in decreasing order of expected monetization.

2.4 Semantic and Syntactic Cloaking

Cloaking behavior that is aimed at manipulating the search engine is defined as semantic cloaking [19]. The exact definition of semantic cloaking varies from search engine to search engine. On the other hand, syntactic cloaking is a simpler and more basic variant of cloaking. Syntactic cloaking implies that different content is served to automated crawlers vs Web browsers, but not different content to every visitor. Dynamic Web pages that serve different pages to every visitor would not be syntactically cloaking, but could be semantically cloaking. In this paper, our operating definition for cloaking is more than just syntactic cloaking. Syntactic cloaking is definitely cloaking, but dynamic Web pages are also addressed to some extent (see Section 6).

3. POPULARITY AND MONETIZABILITY

Monitoring, evaluating, and understanding user behavior and preferences is crucial for search engine development, deployment, and maintenance. Search engines model and interpret user behavior to improve ranking, click spam detection, Web search personalization, and other tasks [1,2,17]. Further, for billing and reporting purposes every impression, user click, and referral relating to each sponsored link are also logged. We propose mining these logs to determine query popularity and monetizability. Such query categorization has been valuable for improving collaborative Web search [15-17].

3.1 Query Popularity

We define the popularity of a query to be proportional to the number of times it occurs in the query logs during a specific time period. Using this definition one can compute query lists such as the top 10 popular search queries for a day, a month, or even a year. Most major search engines publish these results online at different granularities. Table 1 presents a list of common sources of popular queries. The list of top 5000 most popular queries was computed from MSN Search query logs. In this paper, we examine the cloaking properties of search results from these top 5000 popular queries from Google[5], MSN Search[6], and Ask.com[7].

Table 1. Common sources of popular queries

Engine	URL
Google Zeitgeist	http://www.google.com/press/zeitgeist.html
Yahoo Buzz Index	http://buzz.yahoo.com/
MSN Search Insider	http://www.imagine-msn.com/insider/
Ask.com IQ	http://sp.ask.com/en/docs/iq/iq.shtml
AOL Hot Searches	http://hotsearches.aol.com/search/hotsearch.jsp
Dogpile Search Spy	http://www.dogpile.com/info.dogpl/searchspy/
Lycos 50	http://50.lycos.com/

3.2 Query Monetizability

Computing the monetizability of a query is not as straight forward as computing its popularity. Advertisers can bid for a single keyword, a keyword and additional search terms, or a phrase. The bidding process can be blind or open, i.e., each bidder's bid price and identity may or may not be disclosed to other bidders[8]. Three different types of matches are typically possible: broad match, phrase match, and exact match. Some providers support negative or excluded keywords also. The advertiser also picks the type of matching done between the user search query and the bids. A broad match occurs when the user query contains all of the keywords (in any order). Bid keywords may be expanded to include plurals and relevant variations. Phrase match occurs when all bid keywords occur in the prescribed order in the search query. Both broad and phrase matches allow extraneous query words. Exact matching occurs only when the search query matches the bid phrase exactly. No extraneous terms are allowed. The occurrence of negative or excluded keywords in the search query suppresses any matching. The matching sponsored links are ranked based on relevance, monetizability (combination of bid price and CTR), and other factors.

In this paper, we define the monetizabilty of a specific query to be proportional to the total revenue generated by sponsored ads

[2] http://adwords.google.com/select/

[3] http://searchmarketing.yahoo.com/

[4] http://adcenter.msn.com/

[5] http://www.google.com

[6] http://search.msn.com

[7] http://www.ask.com

[8] For example, Yahoo SM auctions are not blind while Google's AdWords and MSN's AdCenter's auctions are blind. Further, not only the bid-price but the bidding advertiser might be disclosed during these auctions.

served along side the search results (for that query) during a specific time period. The list of top 5000 most monetization queries over a single day were computed from MSN Search's advertisement logs. Note that the ad logs are used only to obtain the top 5000 monetizable queries and their ranks. For simplicity, we do not use their monetization scores. For each of these top 5000 monetizable queries, we examine the cloaking properties of the resulting top 200 search results from Google, MSN Search, and Ask.com.

4. DATA SETS

4.1 Query Data Sets

We use two lists of 5000 queries each in the experiments. The first list is the set of the top 5000 most popular search queries computed over one month. The second list is the set of the top 5000 most monetizable search queries over one day. The former was obtained by processing search query logs, while the latter was obtained by processing ad logs. Both logs were obtained from the MSN search engine. 826 queries (17%) were the same between the two lists.

4.2 URL Data Sets

For each query, the top 200 search results were obtained from three search engines: Google, MSN Search, and Ask.com. On every search engine, each unique query was looked up only once. Each query produced 600 search result URLs which typically contain several duplicates. Each set of 5000 queries generated 3 million URLs. Overall, the 5000 popular queries generated 1.49 million unique URLs (popular set), and the top 5000 monetizable queries generated 1.28 million unique URLs (monetizable set). Each unique URL was processed only once.

Current search engines already employ numerous but unknown anti-spam mechanisms. In our analysis, we assume that such Web spam filtering/removal techniques are uniformly applied to the 200 search results and the 5000 queries. This is likely the case for automated filtering that is applied to the whole index, for example during the crawling phase. Manual Web spam removal is one special case where this assumption may not be invalid. Given its expense, manual filtering is likely to be limited to the top few (top 10 or top 20) search results for the most popular queries. Since we are looking at the top 200 results for the top 5000 queries, we conjecture that the impact of filtering on the overall results reported in this paper is small.

5. CLOAKING DETECTION RESULTS

We use a modified version of the syntactic cloaking detection algorithm from [19]. For each URL, up to four copies of the Web page, denoted by C_1, B_1, C_2, and B_2, are downloaded and compared. There are several stages where an early out is possible making the modified procedure more efficient. During the download process, many of the non-cloaked pages are detected through simple HTML string comparisons, HTML to text conversion, and text string comparisons. Normalized term frequency difference (NTFD) is subsequently used to compute a cloaking score and used to further reduce the set of possibly cloaked URLs. Finally, using labeled data, a threshold for the cloaking score is chosen to classify remaining URLs. A flow chart depicting the different stages is presented in Figure 1.

5.1.1 Downloading Web Pages

The first copy of the URL (C_1) was obtained by mimicking a popular Web crawler (MSNBot) and the second (B_1) was obtained using a common Web browser's (Internet Explorer) agent string. The *user-agent* strings for MSNBot and Internet Explorer were set to those given in Section 1. These first and second copies were checked for identical HTML content (simple string comparison). If they were identical, the URL was marked as not cloaked. About 70–75% of the URLs fell under this category. The HTML content for the remaining 25–30% was converted to plain text and directly compared (simple string comparison). At this stage, about 13.5% of the URLs produce identical text streams and are marked as not-cloaked. The text streams are tokenized (using white space) and their term frequencies are computed. About 0.5% of the URLs produce identical term frequencies. The remaining URLs (about 12%) with differing text content were downloaded two more times to obtain a third (MSNBot, C_2) and a fourth (Internet Explorer, B_2) copy. These were then converted to text and their term frequencies calculated. Note that at the end of the download process those URLs with only (C_1, B_1) pair of pages are not-cloaked (by definition). The remaining URLs have four copies (C_1, B_1, C_2, and B_2) and need further processing.

Each of the copies (C_1, B_1, C_2, and B_2) was asynchronously crawled using different crawler threads. For example, all C_1 copies were crawled by the first crawler thread. Similarly, all B_1, C_2, and B_2 copies were crawled by the first browser thread, the second crawler thread, and the second browser thread, respectively. The ordering of initiating URLs downloads was the same for all four threads (with the exception of early out scenarios where URLs were skipped by the C_2, and B_2 threads).

In the event of a download failure, the download was reattempted once. URLs that failed download twice were dropped from analysis. For both the popular and monetizable query URL sets, less than 3% of the URLs failed to download. Overall, on average of about 2.1 downloads are done per unique URL.

5.1.2 Normalized Term Frequency Difference

A simple normalized term frequency difference (NTFD) between the four copies was used in computing a cloaking score. Let T_1 and T_2 be sets of terms from two Web pages after conversion and tokenization. Note that T_1 and T_2 may contain repeats. The normalized term frequency difference is computed as

$$D(T_1, T_2) = \frac{\left|\left(T_1 \setminus T_2\right) \cup \left(T_2 \setminus T_1\right)\right|}{\left|\left(T_1 \cup T_2\right)\right|} = 1 - 2\frac{\left|\left(T_1 \cap T_2\right)\right|}{\left|\left(T_1 \cup T_2\right)\right|}$$

Where |.| is the set cardinality operator and all set operations are extended to work with sets with repeated terms. ($T_1 \setminus T_2$) is the set of terms in the first page but not in the second page, ($T_2 \setminus T_1$) is the set of terms in the second page but not in the first page, and ($T_1 \cup T_2$) is the aggregation of terms in both pages. The normalization by the ($T_1 \cup T_2$) term reduces any bias that stems from the size of the Web page. The NTFD score for any pair of Web pages lies in [0,1]. In essence, for the same $D(T_1,T_2)$, value, larger Web pages are allowed to have more terms that are different between the two pages. We note that the normalized term frequency difference is symmetric, i.e.,

$$D(T_1,T_2) = D(T_2,T_1)$$

The above term-based page-difference score is quite simple and disregards the semantic and layout structure of page content. Further, all sections of the Web page (navigation, header, footer, advertisements, etc) are treated equally[9].

We note that this score differs significantly from that proposed in [19]. Instead of using the cardinality of all the terms in the web pages, a "bag of words" method is used in [19] for analyzing the Web pages. They parse the HTML into terms and only count each unique term once no matter how many times this term appears. Further, they do not normalize the term set difference which could potentially bias the score against large Web pages.

5.1.3 Cloaking Test

As described in Section 5.1.1, many of the URLs are marked as non-cloaking during the download process itself. The remaining URLs end up with four downloaded versions (C_1, B_1, C_2, and B_2). The NTFD score for these four Web page versions is used to obtain a cloaking score, S, given by

$$S = \frac{\Delta_D}{\Delta_S}$$

Where Δ_D is the smaller of the NTFD values for the two cross-pairs of Web pages (C_1,B_1) and (C_2,B_2), and Δ_S is the larger of the NTFD values for the two similar-pairs of Web pages (C_1,C_2) and (B_1,B_2). Mathematically,

$$\Delta_D = \min(D(C_1,B_1), D(C_2,B_2))$$

$$\Delta_S = \max(D(C_1,C_2), D(B_1,B_2))$$

The simple divide-by-zero cases are resolved as follows: (a) If $\Delta_S = 0$ and $\Delta_D = 0$, the URL is marked as non-cloaked ($S = 0$), (b) If $\Delta_S = 0$ and $\Delta_D > 0$, the URL is marked as cloaked ($S = \infty$). At this stage all of the dynamic Web pages are identified using:

$$0 < S < \infty \Rightarrow \text{ dynamic URLs}$$

A subsequent threshold test is used to find cloaked pages:

$$0 < t < S \Rightarrow \text{ cloaking spam}$$

For each of the URL sets (popular and monetizable) 2000 URLs were randomly sampled from the set of dynamic URLs and manually labeled as spam or no-spam.

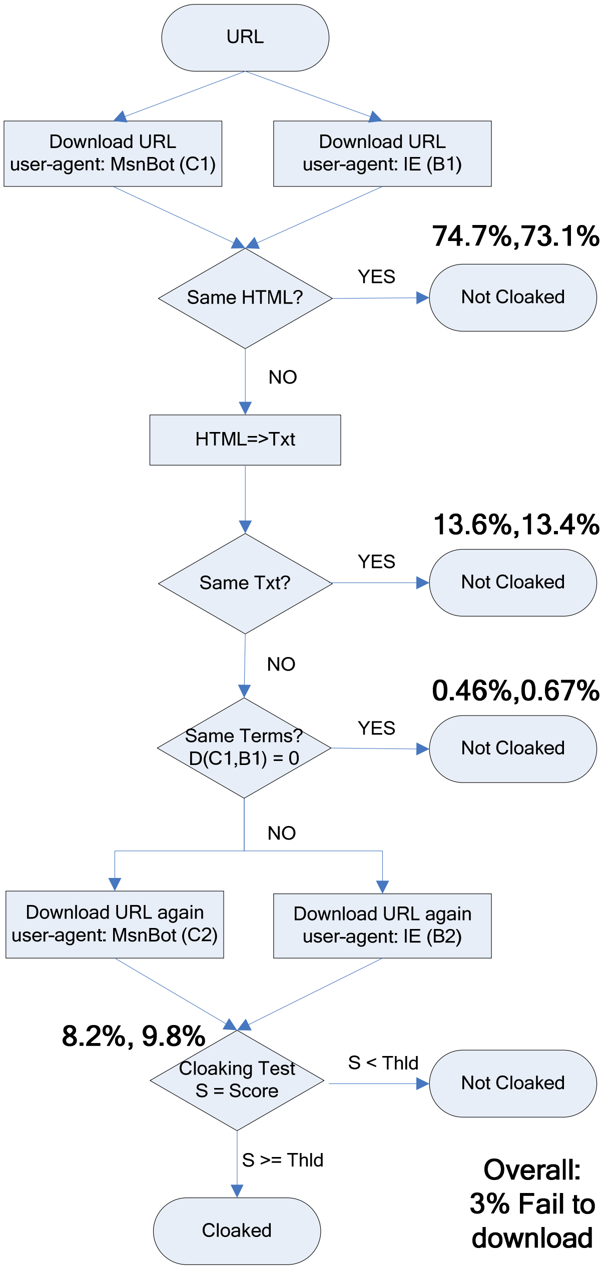

Figure 1. Cloaking detection procedure. The pair of percentages indicate number of classified / unclassified URLs at each stage for the popular and monetizable URLs, respectively.

[9] We plan to explore more advanced methods for computing page difference scores based on page and link content in future work.

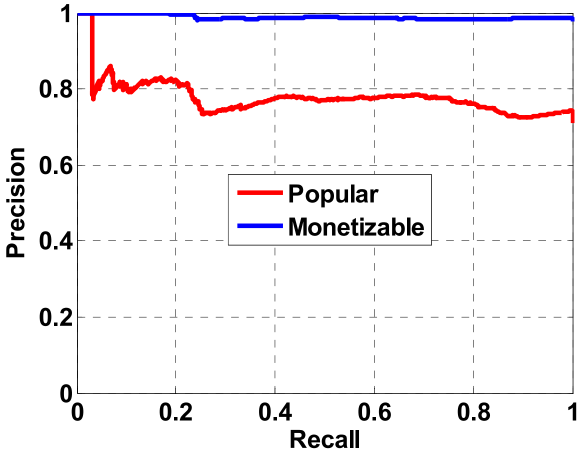

Figure 2. Precision-Recall curve for popular and monetizable URL sets as a function of the cloaking score threshold, *t*.

Figure 2 shows the precision-recall curve for various values of the threshold *t*. The precision and recall values and their associated thresholds are also presented in Table 2. As the value of *t* increases, recall gradually decreases. Precision starts out high at low values of recall and quickly reaches a final value around 75% for popular URLs and a value of 98.5% for monetizable URLs.

All three commonly used F–measures: F_1, $F_{0.5}$, and F_2, reach the highest value at a threshold of 0.0, where the recall is 100% and the precision is 73.12% and 98.54% for popular and monetizable URLs, respectively. This clearly indicates that the cloaking score is a very good indicator of cloaking spam. A threshold of 0 implies that all pages marked as dynamic using

$$0 < S < \infty \Rightarrow \text{ dynamic URLs}$$

can be classified as cloaking spam. Overall, we estimate that 5.99% (=8.2*0.731) of popular query results and 9.66% (=9.8*0.985) of all monetizable queries employ cloaking spam.

Table 2. Precision, Recall, and Thresholds for classifying URLs as cloaking spam using their cloaking score

Recall	Precision (threshold, *t*)	
	Popular URLs	Monetizable URLs
10	85.74 (19.93)	100.00 (15.11)
20	81.72 (1.98)	99.91 (1.28)
30	75.33 (1.10)	98.77 (0.97)
40	76.65 (0.94)	98.56 (0.87)
50	77.39 (0.78)	98.79 (0.77)
60	77.81 (0.53)	98.72 (0.56)
70	77.88 (0.27)	98.59 (0.32)
80	75.86 (0.11)	98.34 (0.07)
90	73.26 (0.02)	98.46 (0.004)
100	73.12 (0.00)	98.54 (0.000)

Note that the above percentages are mean values over all 5000 queries. Figure 3 shows the distribution of cloaking spam URLs over different queries. Both popular and monetizable queries were *independently* sorted such that the percentage curves are monotonically decreasing with increasing sorted query rank. Note that these two query sets are not the same. They have only 17% of the queries in common. We note that, on average, the top 100 (2%) most cloaked queries have 10x as many cloaking URLs in their search results than the bottom 4900 queries (98%). This skewed distribution gives an effective way of monitoring and detecting cloaked URLs. By starting with the most cloaked queries once can efficiently and quickly identify cloaked URLs.

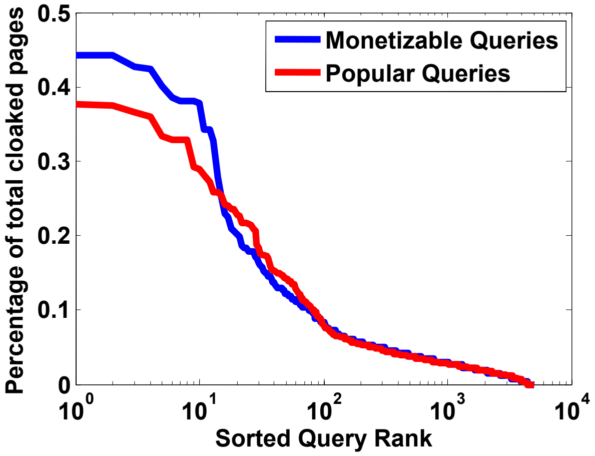

Figure 3. The distribution of cloaking spam URLs over different queries. Both popular and monetizable queries were independently sorted such that the percentage curves are monotonically decreasing with increasing sorted query rank.

6. DISCUSSION AND FUTURE WORK

Cloaking is a search engine spamming technique that reduces the reliability of Web page information that is widely accessible through the search engine. Web sites that deliver one page to a search engine for indexing while serving an entirely different page to users browsing the site inherently hurt the search engine's credibility and waste internet users' time.

In this paper, we showed that the degree of cloaking among search results depends on query properties such as popularity and monetizability. Query popularity and monetizability were estimated based on whether a given query belonged to the popular set of URLs or monetizable set of URLs (or both). We also presented a new cloaking detection algorithm based on normalized term frequency difference scores and demonstrated its effectiveness in identifying cloaking spam pages on a dataset of 3 million URLs obtained using 10,000 search queries.

The proposed cloaking detection algorithm has a very high accuracy in detecting cloaked spam pages in monetizable query results. Moderate accuracy is also achieved for popular queries. By combining a matching model (query ⇔ bid keywords) similar to that used in for serving online advertisements with search query and advertising logs, one can estimate the popularity and monetizability of arbitrary queries. Such estimates may be valuable in prioritizing URLs to be tested for cloaking spam. We hope to pursue these ideas in our future work.

7. ACKNOWLEDGMENTS

We would like to thank Ahmad Abdulkader and Chris Meek for very useful and productive discussions on Cloaking detection. We would also like to thank Chau Luu for help with our spam labeling efforts.

8. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, R. Ragno (2006), “Learning User Interaction Models for Predicting Web Search Result Preferences,” To appear in SIGIR’2006: 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, Seattle.

[2] E. Agichtein, E. Brill, S. Dumais (2006), “Improving Web Search Ranking by Incorporating User Behavior,” To appear in SIGIR’2006: 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, Seattle.

[3] A. Back (2002), “Hash cash - a denial of service counter-measure,” Technical. Report. Available at: http://citeseer.ist.psu.edu/back02hashcash.html

[4] A. Benczúr, K. Csalogány, T. Sarlós and M. Uher (2005), “SpamRank – Fully Automatic Link Spam Detection,” In 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.

[5] B. Davison (2000), “Recognizing Nepotistic Links on the Web,” In AAAI-2000 Workshop on Artificial Intelligence for Web Search, July 2000.

[6] I. Drost and T. Scheffer (2005), “Thwarting the negritude ultramarine: Learning to identify link spam.” In Proceedings of European Conference on Machine Learning, pages 96-107, Oct. 2005.

[7] C. Dwork, A. Goldberg, and M. Naor (2003), “On Memory-Bound Functions for Fighting Spam,” Proceedings of the 23rd Annual International Cryptology Conference (CRYPTO 2003), pages 426-444, Santa Barbara, CA, August 2003.

[8] B. Edelman, M. Ostrovsky, and M. Schwarz (2005), “Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords,” November 2005, NBER Working Paper No. W11765 Available at SSRN: http://ssrn.com/abstract=847037

[9] D. Fetterly, M. Manasse, and M. Najork (2004), “Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages,” In Proceedings of WebDB, pages 1-6, June 2004.

[10] L. Graham and P. T. Metaxas (2003). “Of course it’s true; I saw it on the internet!”: Critical thinking in the internet era. Commun. ACM, 46(5): 70–75, 2003.

[11] Z. Gyöngyi and H. Garcia-Molina (2005), “Web spam taxonomy,” In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb’05), Chiba, Japan, 2005.

[12] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen, “Combating Web Spam with TrustRank,” In 30th International Conference on Very Large Data Bases, Aug. 2004.

[13] V. Krishna (2002). *Auction Theory*, Academic Press, 2002.

[14] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly (2006), “Detecting Spam Web Pages through Content Analysis,” In Proceedings of the World Wide Web Conference 2006 (WWW'06). pp. 83-92, Edinburgh, United Kingdom, May 23-26, 2006.

[15] D. Shen, J-T. Sun, Q. Yang, Z. Chen (2006), “Building Bridges for Web Query Classification,” In Proceedings of the 29th ACM International Conference on Research and Development in Information Retrieval (SIGIR'06). Seattle, USA, August 6-11, 2006.

[16] D. Shen, R. Pan, J-T. Sun, J. J. Pan, K. Wu, J. Yin and Q. Yang (2005), “Q2C@UST: Our Winning Solution to Query Classification in KDD Cup 2005,” SIGKDD Explorations. Volume 7, Issue 2, December 2005.

[17] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz (1999), “Analysis of a Very Large Web Search Engine Query Log. SIGIR Forum,” 33(1):6--12, 1999.

[18] J-T. Sun, X. Wang, D. Shen, H-J. Zeng, Z. Chen (2006), “Mining Clickthrough Data for Collaborative Web Search,” In Proceedings of the World Wide Web Conference 2006 (WWW'06). pp. 947-948, Edinburgh, United Kingdom, May 23-26, 2006.

[19] B. Wu and B. D. Davison (2005) “Cloaking and Redirection: A Preliminary Study,” In Proceedings of AIRWeb'05, May 10, 2005, Chiba, Japan.

[20] Baoning Wu and Brian D. Davison. (2006), “Detecting Semantic Cloaking on the Web,” Accepted in 15th International World Wide Web Conference, Industrial Track, Edinburgh, Scotland, May 22-26, 2006.

[21] “Did-it, Enquiro, and Eyetools Uncover Google’s Golden Triangle: *New Eye Tracking Study verifies the importance of page position and rank in both Organic and PPC search results for visibility and click through in Google*.” Available at http://www.prweb.com/releases/2005/3/prweb213516.htm