Cloak and Dagger: Dynamics of Web Search Cloaking

David Y. Wang, Stefan Savage, and Geoffrey M. Voelker
Deptartment of Computer Science and Engineering
University of California, San Diego

ABSTRACT

Cloaking is a common "bait-and-switch" technique used to hide the true nature of a Web site by delivering blatantly different semantic content to different user segments. It is often used in search engine optimization (SEO) to obtain user traffic illegitimately for scams. In this paper, we measure and characterize the prevalence of cloaking on different search engines, how this behavior changes for targeted versus untargeted advertising and ultimately the response to site cloaking by search engine providers. Using a custom crawler, called *Dagger*, we track both popular search terms (e.g., as identified by Google, Alexa and Twitter) and targeted keywords (focused on pharmaceutical products) for over five months, identifying when distinct results were provided to crawlers and browsers. We further track the lifetime of cloaked search results as well as the sites they point to, demonstrating that cloakers can expect to maintain their pages in search results for several days on popular search engines and maintain the pages themselves for longer still.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Measurement, Security

Keywords

Cloaking, Search Engine Optimization, Web spam

1. INTRODUCTION

The growth of e-commerce in the late 20th century in turn created value around the attention of individual Internet users — described crassly by Caldwell as "The Great Eyeball Race" [3]. Since then, virtually every medium of Internet interaction has been monetized via some form of advertising, including e-mail, Web sites, social networks and on-line games, but perhaps none as successfully as *search*. Today, the top Internet search engines are a primary means for connecting customers and sellers in a broad range of markets, either via "organic" search results or sponsored search placements—together comprising a $14B marketing sector [16].

Not surprisingly, the underlying value opportunities have created strong incentives to influence search results—a field called "search engine optimization" or SEO. Some of these techniques are benign and even encouraged by search engine operators (e.g., simplifying page content, optimizing load times, etc.) while others are designed specifically to manipulate page ranking algorithms without regard to customer interests (e.g., link farms, keyword stuffing, blog spamming, etc.) Thus, a cat and mouse game has emerged between search engine operators and scammers where search operators try to identify and root out pages deemed to use "black hat" optimization techniques or that host harmful content (e.g., phishing pages, malware sites, etc.) while scammers seek to elude such detection and create new pages faster than they can be removed.

In this conflict, one of the most potent tools is *cloaking*, a "bait-and-switch" technique used to hide the true nature of a Web site by delivering blatantly different content to different user segments. Typically a cloaker will serve "benign" content to search engine crawlers and scam content to normal visitors who are referred via a particular search request. By structuring the benign version of this content to correspond with popular search terms—a practice known as keyword stuffing—Web spammers aggressively acquire unwitting user traffic to their scam pages. Similarly, cloaking may be used to prevent compromised Web servers hosting such scam pages from being identified (i.e., by providing normal content to visitors who are not referred via the targeted search terms). In response to such activity, search engine providers attempt to detect cloaking activity and delist search results that point to such pages.

In this paper, we study the dynamics of this cloaking phenomenon and the response it engenders. We describe a system called *Dagger*, designed to harvest search result data and identify cloaking in near real-time. Using this infrastructure for over five months we make three primary contributions. First, we provide a contemporary picture of cloaking activity as seen through three popular search engines (Google, Bing and Yahoo) and document differences in how each is targeted. Second, we characterize the differences in cloaking behavior between sites found using undifferentiated "trending" keywords and those that appear in response to queries for targeted keywords (in particular for pharmaceutical products). Finally, we characterize the *dynamic behavior* of cloaking activity including the lifetime of cloaked pages and the responsiveness of search engines in removing results that point to such sites.

The remainder of this paper is structured as follows. Section 2 provides a technical background on cloaking and the related work we build upon, followed by a description of Dagger's design and implementation in Section 3. Section 4 describes our results, followed in Section 5 by a discussion of our overall findings.

CCS'11, October 17–21, 2011, Chicago, Illinois, USA.

2. BACKGROUND

The term "cloaking", as applied to search engines, has an uncertain history, but dates to at least 1999 when it entered the vernacular of the emerging search engine optimization (SEO) market.[1] The growing role of search engines in directing Web traffic created strong incentives to reverse engineer search ranking algorithms and use this knowledge to "optimize" the content of pages being promoted and thus increase their rank in search result listings. However, since the most effective way to influence search rankings frequently required content vastly different from the page being promoted, this encouraged SEO firms to serve different sets of page content to search engine crawlers than to normal users; hence, cloaking.

In the remainder of this section we provide a concrete example of how cloaking is used in practice, we explain the different techniques used for visitor differentiation, and summarize the previous work that has informed our study.

2.1 An Example

As an example of cloaking, entering the query "bethenney frankel twitter" on Google on May 3, 2011 returned a set of search results with links related to Bethenney Frankel. (The appendix includes screenshots of this example, with Figure 11a showing a screenshot of the search results.) For the eighth result, the bold face content snippet returned by Google indeed indicates that the linked page includes the terms "bethenney frankel twitter" (in addition to other seemingly unrelated terms, a common indicator of keyword stuffing). Further, Google preview shows a snapshot of the page that the Google crawler sees. In this case it appears the link indeed leads to content with text and images related to "bethenney frankel". Upon clicking on this link on a Windows machine, however, we are sent through a redirect chain and eventually arrive at a landing page (Figure 11b). This page is an example of the fraudulent anti-virus scams that have become increasingly popular [5, 15] and shares little with "bethenney frankel twitter" (and indeed, both the previous snippet and any hints of the previewed snapshot do not appear). Finally, if we revisit the search result link but disguise our identity to appear as a search crawler (in this case by modifying the `User-Agent` string in our HTTP request header) we get redirected using a 302 HTTP code to the benign root page of the site (Figure 11c), presumably to avoid possible suspicion. This case represents a classic example of IP cloaking. The site is detecting Google IP addresses, as evidenced by the snippet and preview, and providing an SEO-ed page with text and images related to "bethenney frankel twitter" that it deceives Google into indexing. The core idea here is clear. The scammer is crafting content to reflect the transient popularity of particular terms (in this case due to Bethenney Frankel's sale of her cocktail line) and serving this content to search engine crawlers to capitalize on subsequent user queries, but then directs any user traffic to completely unrelated content intended to defraud them.

2.2 Types of Cloaking

For cloaking to work, the scammer must be able to distinguish between user segments based on some identifier visible to a Web server. The choice of identifier used is what distinguishes between cloaking techniques, which include *Repeat Cloaking*, *User Agent Cloaking*, *Referrer Cloaking* (sometimes also called "Click-through Cloaking"), and *IP Cloaking*.

[1]For example, in a September 2000 article in *Search Engine Watch*, Danny Sullivan comments that "every major performance-based SEO firm I know of does [use cloaking]" [19].

In the case of *Repeat Cloaking*, the Web site stores state on either the client side (using a cookie) or the server side (e.g., tracking client IPs). This mechanism allows the site to determine whether the visitor has previously visited the site, and to use this knowledge in selecting which version of the page to return. Thus first-time visitors are given a glimpse of a scam, in the hopes of making a sale, but subsequent visits are presented with a benign page stymieing reporting and crawlers (who routinely revisit pages).

In contrast, *User Agent Cloaking* uses the `User-Agent` field from the HTTP request header to classify HTTP clients as user browsers or search engine crawlers. User agent cloaking can be used for benign content presentation purposes (e.g., to provide unique content to Safari on an iPad vs. Firefox on Windows), but is routinely exploited by scammers to identify crawlers via the well-known `User-Agent` strings they advertise (e.g., Googlebot).

Referrer Cloaking takes the idea of examining HTTP headers even further by using the `Referer` field to determine which URL visitors clicked through to reach their site. Thus, scammers commonly only deliver a scam page to users that visit their site by first clicking through the search engine that has been targeted (e.g., by verifying that the `Referer` field is `http://www.google.com`). This technique has also been used, in combination with repeat cloaking and chains of Web site redirections, to create one-time-use URLs advertised in e-mail spam (to stymie security researchers). However, we restrict our focus to search engine cloaking in this paper.

Finally, one of the simplest mechanisms in use today is *IP Cloaking*, in which a scammer uses the IP address of the requester in determining the identity of the visitor. With an accurate mapping between IP addresses and organizations, a scammer can then easily distinguish all search engine requests and serve them benign content in a manner that is difficult to side-step. Indeed, the only clear way for search engine operators to mechanistically detect such cloaking is through acquiring fresh IP resources—but the signal of "delisting" seems to provide a clear path for efficiently mapping such address space even if it is initially unknown [1]. Although challenging to detect in principle since it would nominally require crawling from a Google IP address, in practice a crawler like Dagger can still detect the use of IP cloaking because cloakers still need to expose different versions of a page to different visitors (Section 3.3). As a result, we believe Dagger can detect all forms of cloaking commonly used in the Web today.

2.3 Previous work

The earliest study of cloaking we are aware of is due to Wu and Davidson [24]. They first developed the now standard technique of crawling pages multiple times (using both user and crawler identifiers) and comparing the returned content. Using this approach, they refer to situations in which the content differs as "syntactic cloaking", whereas the subset of these differences that are deemed to be driven by fraudulent intent (using some other classifier [25]) are termed "semantic cloaking".

Chellapilla and Chickering used a similar detection framework to compare syntactic cloaking on the most popular and monetizable search terms [4]. In this study, monetizability corresponds to the amount of revenue generated from users clicking on sponsored ads returned with search results for a search term. Using logs from a search provider, they found that monetized search terms had a higher prevalence of cloaking (10%) than just popular terms (6%) across the top 5000 search terms. While we do not have access to logs of search provider revenue, we also demonstrate significant differences in cloaking prevalence as a function of how search terms are selected.

Up to this point, all studies had focused exclusively on user agent

cloaking. In 2006, Wang et al. extended this analysis to include referrer cloaking (called click-through cloaking by them), where pages only return cloaked content if accessed via the URL returned in search results [21]. Targeting a handful of suspicious IP address ranges, they found widespread referrer cloaking among the domains hosted there. In a related, much more extensive study, Wang et al. used this analysis in a focused study of redirection spam, a technique by which "doorway" Web pages redirect traffic to pages controlled by spammers [22]. Finally, Niu et al. performed a similar analysis incorporating referrer cloaking but focused exclusively on forum spamming [14].

These prior studies have used a range of different inputs in deciding whether multiple crawls of the same page are sufficiently different that cloaking has occurred. These include differences in the word distribution in the content of the two pages [4, 24], differences in the links in the page [24], differences in HTML tags [12] or differences in the chain of domain names in the redirection chain [21, 22]. A nice summary of these techniques as well as the different algorithms used to compare them is found in [12]. Our approach represents a combination of these techniques and some new ones, driven by a desire to support crawling and detection in near-real time (Section 3).

Finally, the use of cloaking is predominately driven by so-called "black hat" search engine optimization (SEO) in which the perpetrators seek to attract traffic by using various methods to acquire higher ranking in search engine results. Two contemporaneous studies touch on different aspects of this behavior that are echoed in our own work. John et al. examine one very large scale "link farm" attack used to create high rankings for "trending" search terms and thus attract large amounts of undifferentiated traffic [8]. While this particular attack did not use cloaking per se (except implicitly in the use of JavaScript redirection) we encounter similar structures driving the popularity of cloaked sites in our measurements. Leontiadis et al. perform a broad study of SEO-advertised pharmaceutical sites (driven by focused searches on drug-related terms) and note that cloaking is commonly used by compromised sites and, in part, serves the purpose of "hiding" the existence of such sites from normal visitors [10]. We explore both kinds of search traffic streams in our analysis (driven by trending and focused search terms) and find significant differences between them.

In general, most studies of cloaking have focused on a single snapshot in time. Moreover, the prevalence of cloaking reported in all such studies is difficult to compare due to variations in detection approach, differences in how search terms are selected and changes in scammer behavior during different time periods. Thus, one of our primary contributions is in extending these previous efforts to examine the dynamics of cloaking over time, uncovering the fine-grained variability in cloaking, the lifetime of cloaked search results (bounding the response time of search engines to cloaked results) and the duration of the pages they point to. Ultimately, these *dynamics* in the ranking of cloaked sites drive the underlying economics of their attack.

3. METHODOLOGY

Dagger consists of five functional components: collecting search terms, fetching search results from search engines, crawling the pages linked from the search results, analyzing the pages crawled, and repeating measurements over time. In this section, we describe the design and implementation of each functional component, focusing on the goals and potential limitations.

3.1 Collecting Search Terms

The first challenge in data collection is building a meaningful test set for measuring cloaking. Since our goal is to understand the dynamics of scammers utilizing cloaking in search results, we want to target our data collection to the search terms that scammers also target rather than a random subset of the overall search space. In particular, we target two different kinds of cloaked search terms: those reflecting popular terms intended to gather high volumes of undifferentiated traffic, and terms reflecting highly targeted traffic where the cloaked content matches the cloaked search terms.

For our first set of search terms, as with previous work we seed our data collection with popular *trending* search terms. We also enhance this set by adding additional sources from social networks and the SEO community. Specifically, we collect popular search terms from Google Hot Searches, Alexa, and Twitter, which are publicly available and provide real-time updates to search trends at the granularity of an hour.[2] We extract the top 20 popular search trends via Google Hot Searches and Alexa, which reflect search engine and client-based data collection methods, respectively, while the 10 most popular search terms from Twitter adds insight from social networking trends. These sources generally compliment each other and extend our coverage of terms. We found that terms from Google Hot Searches only overlapped 3–8% with trending terms from both Twitter and Alexa. Note that, for trending terms, the page being cloaked is entirely unrelated to the search terms used to SEO the page. A user may search for a celebrity news item and encounter a search result that is a cloaked page selling fake anti-virus.

For our second set of search terms, we use a set of terms catering to a specific domain: pharmaceuticals. We gathered a generic set of pharmaceutical terms common to many spam-advertised online pharmacy sites, together with best-selling product terms from the most popular site [11]. Unlike trending search terms, the content of the cloaked pages actually matches the search terms. A user may search for "viagra" and encounter a cloaked page that leads to an online pharmacy site selling Viagra.

We construct another source of search terms using keyword suggestions from Google Suggest. Google Suggest is a search term autocomplete service that not only speeds up user input, but also allows users to explore similar long-tail queries. For example, when users enter "viagra 50mg", they are prompted with suggestions such as "viagra 50mg cost" and "viagra 50mg canada". Specifically, we submit search terms from Google Hot Searches and the online pharmacy site to Google Suggest and use the result to create dynamic feeds of search terms for trending and pharmaceutical searches, respectively. While investigating SEO community forums, we found various software packages and services using popular search term services as seeds for extending the terms they target using suggestion services. Combined with a suggestion service, each search term forms the basis of a cluster of related search terms from lists of suggestions [23]. The main attraction of a suggestion service is that it targets further down the tail of the search term distribution, resulting in less competition for the suggestion and a potentially more advantageous search result position. Moreover, long-tail queries typically retain the same semantic meaning as the original search term seed. Furthermore, recent studies have shown superior conversion rates of long-tail queries [20].

3.2 Querying Search Results

Dagger then submits the terms, every four hours for trending

[2]We originally considered using an even broader range of search term sources, in particular Yahoo Buzz, Ask IQ, AOL Hot Searches, eBay Pulse, Amazon Most Popular Tags, Flickr Popular Tags, and WordPress Hot Topics. Since we detected no more than a few cloaked results in multiple attempts over time, we concluded that scammers are not targeting these search terms and no longer considered those sources.

queries and every day for pharmaceutical queries, to the three most actively used search engines in the US: Google, Yahoo, and Bing. With results from multiple search engines, we can compare the prevalence of cloaking across engines and examine their response to cloaking. From the search results we extract the URLs for crawling as well as additional metadata such as the search result position and search result snippet.

At each measurement point, we start with the base set of search terms and use them to collect the top three search term suggestions from Google Suggest.[3] For trending searches, Google Hot Searches and Alexa each supply 80 search terms every four-hour period, while Twitter supplies 40. Together with the 240 additional suggestions based on Google Hot Searches, our search term workload is 440 terms. Note that while overlap does occur within each four-hour period and even between periods, this overlap is simply an artifact of the search term sources and represents popularity as intended. For example, a term that spans multiple sources or multiple periods reflects the popularity of the term. For pharmaceutical queries, we always use a set of 230 terms composed of the original 74 manually-collected terms and 156 from Google Suggest.

Next, we submit the search terms and suggestions to the search engines and extract the top 100 search results and accompanying metadata. We assume that 100 search results, representing 10 search result pages, covers the maximum number of results the vast majority of users will encounter for a given query [17]. Using these results Dagger constructs an index of URLs and metadata, which serves as the foundation for the search space that it will crawl and analyze. At this point, we use a whitelist to remove entries in the index based on regular expressions that match URLs from "known good" domains, such as `http://www.google.com`. This step reduces the number of false positives during data processing. In addition, whenever we update this list, we re-process previous results to futher reduce the number of false positives. Afterwards, we group similar entries together. For example, two entries that share the same URL, source, and search term are combined into a single entry with the same information, except with a count of two instead of one to signify the quantity. As a result, for each search engine, instead of crawling 44,000 URLs for trending search results (440 search terms $\times$ 100 search results), on average Dagger crawls roughly 15,000 unique URLs in each measurement period.

3.3 Crawling Search Results

For each search engine, we crawl the URLs from the search results and process the fetched pages to detect cloaking in parallel to minimize any possible time of day effects.

Web crawler. The crawler incorporates a multithreaded Java Web crawler using the `HttpClient 3.x` package from Apache. While this package does not handle JavaScript redirects or other forms of client-side scripting, it does provide many useful features, such as handling HTTP 3xx redirects, enabling HTTP header modification, timeouts, and error handling with retries. As a result, the crawler can robustly fetch pages using various identities.

Multiple crawls. For each URL we crawl the site three times, although only the first two are required for analysis. We begin disguised as a normal search user visiting the site, clicking through the search result using Internet Explorer on Windows. Then we visit the site disguised as the Googlebot Web crawler. These crawls download the views of the page content typically returned to users and crawlers, respectively. Finally, we visit the site for a third time as a user who does not click through the search result to download the view of the page to the site owner. As with previous approaches, we disguise ourselves by setting the `User-Agent` and the `Referer` fields in the HTTP request header. This approach ensures that we can detect any combination of user-agent cloaking and referrer cloaking. Moreover, our third crawl allows us to detect pure user-agent cloaking without any checks on the referrer. We found that roughly 35% of cloaked search results for a single measurement perform pure user-agent cloaking. For the most part, these sites are not malicious but many are involved in black-hat SEO operations. In contrast, pages that employ both user-agent and referrer cloaking are nearly always malicious (Section 4.5).

IP cloaking. Past studies on cloaking have not dealt with IP address cloaking, and the methodology we use is no different. However, because the emphasis of our study is in detecting the situation where cloaking is used as an SEO technique in scams, we do not expect to encounter problems caused by IP cloaking. In our scenario, the cloaker must return the scam page to the user to potentially monetize the visit. And the cloaker must return the SEO-ed page to the search crawler to both index and rank well. Even if the cloaker could detect that we are not a real crawler, they have few choices for the page to return to our imitation crawler. If they return the scam page, they are potentially leaving themselves open to security crawlers or the site owner. If they return the SEO-ed page, then there is no point in identifying the real crawler. And if they return a benign page, such as the root of the site, then Dagger will still detect the cloaking because the user visit received the scam page, which is noticeably different from the crawler visit. In other words, although Dagger may not obtain the version of the page that the Google crawler sees, Dagger is still able to detect that the page is being cloaked (see Appendix A for an illustrated example).

To confirm this hypothesis, we took a sample of 20K cloaked URLs returned from querying trending search terms. We then crawled those URLs using the above methodology (three crawls, each with different `User-Agent` and `Referer` fields). In addition, we performed a fourth crawl using Google Translate, which visits a URL using a Google IP address and will fool reverse DNS lookups into believing the visit is originating from Google's IP address space. From this one experiment, we found more than half of current cloaked search results do in fact employ IP cloaking via reverse DNS lookups, yet in every case they were detected by Dagger because of the scenario described above.

3.4 Detecting Cloaking

We process the crawled data using multiple iterative passes where we apply various transformations and analyses to compile the information needed to detect cloaking. Each pass uses a comparison-based approach: we apply the same transformations onto the views of the same URL, as seen from the user and the crawler, and directly compare the result of the transformation using a scoring function to quantify the delta between the two views. In the end, we perform thresholding on the result to detect pages that are actively cloaking and annotate them for later analysis.

While some of the comparison-based mechanisms we use to detect cloaking are inspired from previous work, a key constraint is our real-time requirement for repeatedly searching and crawling to uncover the time dynamics of cloaking. As a result, we cannot use a single snapshot of data, and we avoided intensive offline training for machine learning classifiers [4, 12, 24, 25]. We also avoided running client-side scripts, which would add potentially unbounded latency to our data collection process. Consequently, we do not directly measure all forms of redirection, although we do capture the same end result: a difference in the semantic content of the same URL [21, 22]. Since we continuously remeasure over time, man-

[3] We only use three suggestions to reduce the overall load on the system while still maintaining accuracy, as we found no significant difference in our results when using five or even ten suggestions.

ual inspection is not scalable outside of a couple of snapshots [14]. Moreover, even an insightful mechanism that compares the structure of two views using HTML tags, to limit the effects of dynamic content [12], must be applied cautiously as doing so requires significant processing overhead.

The algorithm begins by removing any entries where either the user or crawler page encountered an error during crawling (a non-200 HTTP status code, connection error, TCP error, etc.); on average, 4% of crawled URLs fall into this category.

At this point, the remaining entries represent the candidate set of pages that the algorithm will analyze for detecting cloaking. To start, the detection algorithm filters out nearly identical pages using text shingling [2], which hashes substrings in each page to construct signatures of the content. The fraction of signatures in the two views is an excellent measure of similarity as we find nearly 90% of crawled URLs are near duplicates between the multiple crawls as a user and as a crawler. From experimentation, we found that a difference of 10% or less in sets of signatures signifies nearly identical content. We remove such pages from the candidate set.

From this reduced set, we make another pass that measures the similarity between the snippet of the search result and the user view of the page. The snippet is an excerpt from the page content obtained by search engines, composed from sections of the page relevant to the original query, that search engines display to the user as part of the search result. In effect, the snippet represents ground truth about the content seen by the crawler. Often users examine the snippet to help determine whether to follow the link in the search result. Therefore, we argue that the user has an implicit expectation that the page content should resemble the snippet in content.

We evaluate snippet inclusion by first removing noise (character case, punctuation, HTML tags, gratuitous whitespace, etc.) from both the snippet and the body of the user view. Then, we search for each substring from the snippet in the content of the user view page, which can be identified by the character sequence '...' (provided in the snippet to identify non-contiguous text in the crawled page). We then compute a score of the ratio of the number of words from unmatched substrings divided by the total number of words from all substrings. The substring match identifies similar content, while the use of the number of words in the substring quantifies this result. An upper bound score of 1.0 means that no part of the snippet matched, and hence the user view differs from the page content originally seen by the search engine; a lower bound score of 0.0 means that the entire snippet matched, and hence the user view fulfills the expectation of the user. We use a threshold of 0.33, meaning that we filter out entries from the candidate set whose user view does not differ by more than two-thirds from the snippet. We chose this threshold due to the abundance of snippets seen with three distinct substrings, and 0.33 signifies that the majority of the content differs between the snippet and user view. In practice, this step filters 56% of the remaining candidate URLs.

At this point, we know (1) that the views are different in terms of unstructured text, and (2) that the user view does not resemble the snippet content. The possibility still exists, however, that the page is not cloaked. The page could have sufficiently frequent updates (or may rotate between content choices) that the snippet mismatch is misleading. Therefore, as a final test we examine the page structures of the views via their DOMs as in [12]. The key insight for the effectiveness of this approach comes from the fact that, while page content may change frequently, as in blogs, it is far less likely for the page structure to change dramatically.

We compare the page structure by first removing any content that is not part of a whitelist of HTML structural tags, while also attempting to fix any errors, such as missing closing tags, along the way. We compute another score as the sum of an overall comparison and a hierarchical comparison. In the overall comparison, we calculate the ratio of unmatched tags from the entire page divided by the total number of tags. In the hierarchical comparison, we calculate the ratio of the sum of the unmatched tags from each level of the DOM hierarchy divided by the total number of tags. We use these two metrics to allow the possibility of a slight hierarchy change, while leaving the content fairly similar. An upper bound score of 2.0 means that the DOMs failed to match any tags, whereas a lower bound score of 0.0 means that both the tags and hierarchy matched. We use a threshold of 0.66 in this step, which means that cloaking only occurs when the combination of tags and hierarchy differ by a third between the structure of the user and crawler views. We chose this threshold from experimentation that showed the large majority of cloaked pages scored over 1.0. Once we detect an entry as actively cloaking, we annotate the entry in the index for later processing.

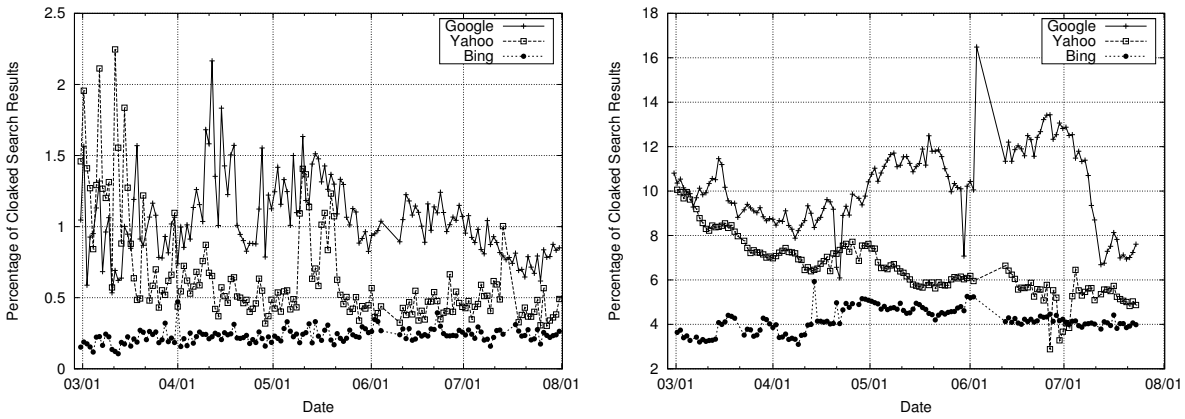
When using any detection algorithm, we must consider the rate of false positives and false negatives as a sign of accuracy and success. Because it is infeasible to manually inspect all results, we provide estimates based on sampling and manual inspection. For Google search results, we found 9.1% (29 of 317) of cloaked pages were false positives, meaning that we labeled the search result as cloaking, but it is benign; for Yahoo, we found 12% (9 of 75) of cloaked pages were false positives. It it worth noting that although we labeled benign content as cloaking, they are technically delivering different data to users and crawlers. If we consider false positives to mean that we labeled the search result as cloaking when it is not, then there are no false positives in either case. In terms of false negatives, when manually browsing collections of search results Dagger detected cloaked redirection pages for the majority of the time. The one case where we fail is when the site employs advanced browser detection to prevent us from fetching the browser view, but we have only encountered this case a handful of times.

3.5 Temporal Remeasurement

The basic measurement component captures data related to cloaking at one point in time. However, because we want to study temporal questions such as the lifetime of cloaked pages in search results, Dagger implements a temporal component to fetch search results from search engines and crawl and process URLs at later points in time. In the experiments in this paper, Dagger remeasures every four hours up to 24 hours, and then every day for up to seven days after the original measurement.

The temporal component performs the same basic data collection and processing steps as discussed in the previous components. To measure the rate at which search engines respond to cloaking, we fetch results using the original search term set and construct a new index from the results that will capture any churn in search results since the original measurement. Then we analyze the churn by searching for any entry with a URL that matches the set of cloaked pages originally identified and labeled. Note that there still exists the possibility that for every cloaked page removed from the new index, another cloaked page, which originally was not a part of the index, could have taken its place. Therefore, this process does not measure how clean the search results are at a given time, just whether the original cloaked pages still appear in the search results.

To measure the duration pages are cloaked, the temporal component selects the cloaked URLs from the original index. It then performs the measurement process again, visiting the pages as both a user and a crawler, and applying the detection algorithm to the results. There still exists the possibility that pages perform cloaking at random times rather than continuously, which we might not de-



(a) Trends (b) Pharmacy

Figure 1: Prevalence of cloaked search results in Google, Yahoo, and Bing over time for trending and pharmaceutical searches.

tect. However, we consider these situations unlikely as spammers need sufficient volume to turn a profit and hence cloaking continuously will result in far greater traffic than cloaking at random.

4. RESULTS

This section presents our findings from using Dagger to study cloaking in trending and pharmaceutical search results in Google, Yahoo, and Bing. We use Dagger to collect search results every four hours for two months, from March 1, 2011 through August 1, 2011, crawling over 47 million search results. We examine the prevalence of cloaking in search results over time, how cloaking correlates to the various search terms we use, how search engines respond to cloaking, and how quickly cloaked pages are taken down. We also broadly characterize the content of cloaked pages, the DNS domains with cloaked pages, and how well cloaked pages perform from an SEO perspective. Where informative, we note how trending and pharmaceutical cloaking characteristics contrast, and also comment on the results of cloaking that we observe compared with results from previous studies.

4.1 Cloaking Over Time

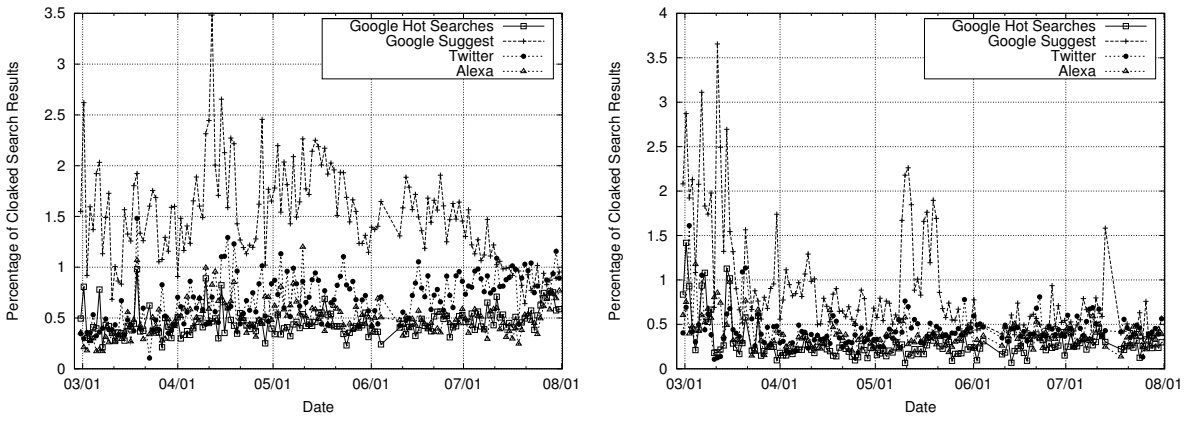
Figure 1a shows the prevalence of cloaking over time for trending search results returned from each search engine. We show the percentage of the cloaked search results averaged across all searches made each day. Recall from Section 3.3 that we crawl the top 100 search results every four hours for 183 unique trending search terms (on average) collected from Google Hot Searches, Google Suggest, Twitter, and Alexa, resulting on average in 13,947 unique URLs to crawl after de-duping and whitelisting. Although we crawl every four hours, we report the prevalence of cloaking at the granularity of a day for clarity (we did not see any interesting time-of-day effects in the results). For example, when cloaking in Google search results peaked at 2.2% on April 12, 2011, we found 2,094 out of 95,974 cloaked search results that day.

Initially, through May 4th, we see the same trend for the prevalence of cloaked search results among search engines: Google and Yahoo have nearly the same amount of cloaked results (on average 1.11% on Google and 0.84% on Yahoo), whereas Bing has 3–4× fewer (just 0.25%). One explanation is that Bing is much better at detecting and thwarting cloaking, but we find evidence that cloakers simply do not appear to target Bing nearly as heavily as Google and Yahoo. For instance, cloaked results often point to link farms, a form of SEO involving collections of interlinked pages used to boost page rank for sets of search terms. For large-scale link farms that we have tracked over time, we consistently find them in Google and Yahoo results, but not in Bing results.

Similarly, Figure 1b shows the prevalence of cloaking over time when searching for pharmaceutical terms. We crawl the top 100 search results daily for 230 unique pharmaceutical search terms collected from a popular spam-advertised affiliate program, further extended with Google Suggest, resuling in 13,646 unique URLs to crawl after de-duping and whitelisting. (Note that the gap in results in the first week of June corresponds to a period when our crawler was offline.) Across all days, we see the same relative ranking of search engines in terms of cloaking prevalence, but with overall larger quantities of cloaked results for the same respective time ranges: on average 9.4% of results were cloaked on Google, 7.7% results on Yahoo, and 4.0% on Bing.

The difference in quantities of cloaked results for trending and pharmaceutical terms reflects the differences between these two types of searches. In trending searches the terms constantly change, with popularity being the one constant. This dynamic allows cloakers to target many more search terms and a broad demographic of potential victims—anyone by definition searching using a popular search term—at the cost of limited time to perform the SEO needed to rank cloaked pages highly in the search results. In contrast, pharmaceutical search terms are static and represent product searches in a very specific domain. Cloakers as a result have much more time to perform SEO to raise the rank of their cloaked pages, resulting in more cloaked pages in the top results. Note, though, that these targeted search terms limit the range of potential victims to just users searching in this narrow product domain. Section 4.7 further explores the effects of SEO on cloaked results.

Looking at trends over time, cloakers were initially slightly more successful on Yahoo than Google for trending search terms, for instance. However, from April 1 through May 4th, we found a clear shift in the prevalence of cloaked search results between search engines with an increase in Google (1.2% on average) and a decrease in Yahoo (0.57%). We suspect this is due to cloakers further concentrating their efforts at Google (e.g., we uncovered new link farms performing reverse DNS cloaking for the Google IP range). In addition, we saw substantial fluctuation in cloaking from day to day. We attribute the variation to the adversarial relationship between cloakers and search engines. Cloakers perform blackhat SEO to artifically boost the rankings of their cloaked pages. Search engines refine their defensive techniques to detect cloaking either directly (analyzing pages) or indirectly (updating the ranking algorithm). We interpret our measurements at these time scales as



(a) Google (b) Yahoo

Figure 2: Prevalence of cloaked search results over time associated with each source of trending search terms.

Search Term	% Cloaked Pages
viagra 50mg canada	61.2%
viagra 25mg online	48.5%
viagra 50mg online	41.8%
cialis 100mg	40.4%
generic cialis 100mg	37.7%
cialis 100mg pills	37.4%
cialis 100mg dosage	36.4%
cialis 10mg price	36.2%
viagra 100mg prices	34.3%
viagra 100mg price walmart	32.7%

Table 1: Top 10 pharmaceutical search terms with the highest percentage of cloaked search results, sorted in decreasing order.

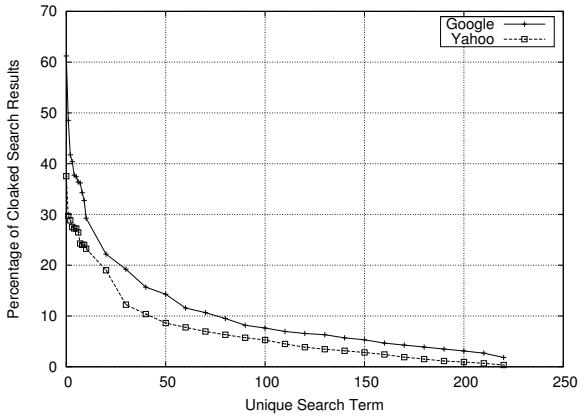


Figure 3: Distribution of percentage of cloaked search results for pharmaceutical search terms, sorted in decreasing order.

simply observing the struggle between the two sides. Finally, we note that the absolute amount of cloaking we find is less than some previous studies, but such comparisons are difficult to interpet since cloaking results fundamentally depend upon the search terms used.

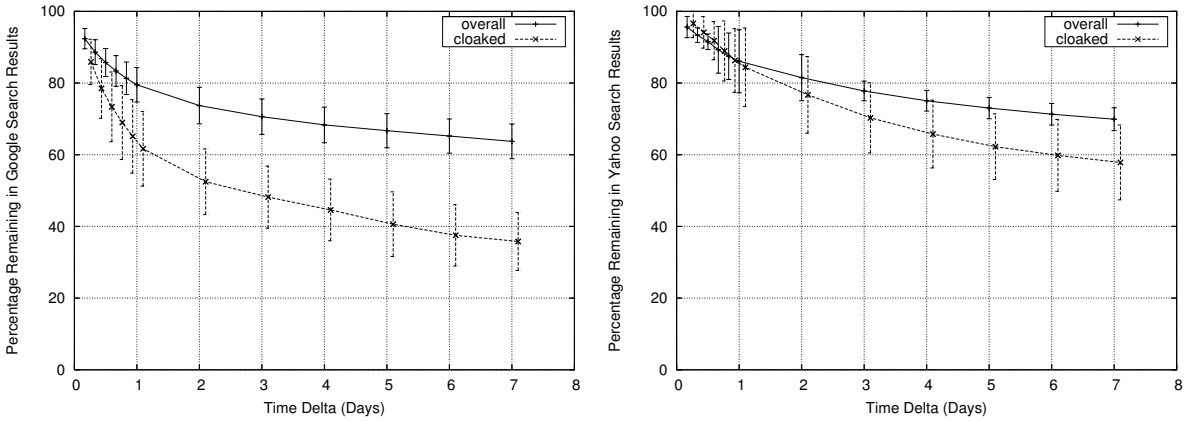
4.2 Sources of Search Terms

Cloakers trying to broadly attract as many visitors as possible target trending popular searches. Since we used a variety of sources for search terms, we can look at how the prevalence of cloaking correlates with search term selection.

Figures 2a and 2b show the average prevalence of cloaking for each source on search results returned from Google and Yahoo, respectively, for trending searches; we do not present results from Bing due to the overall lack of cloaking. Similar to Figure 1, each point shows the percentage of cloaked links in the top 100 search results. Here, though, each point shows the average percentage of cloaked results for a particular source, which normalizes the results independent of the number of search terms we crawled from each source. (Because different sources provided different numbers of search terms, the percentages do not sum to the overall percentages in Figure 1.)

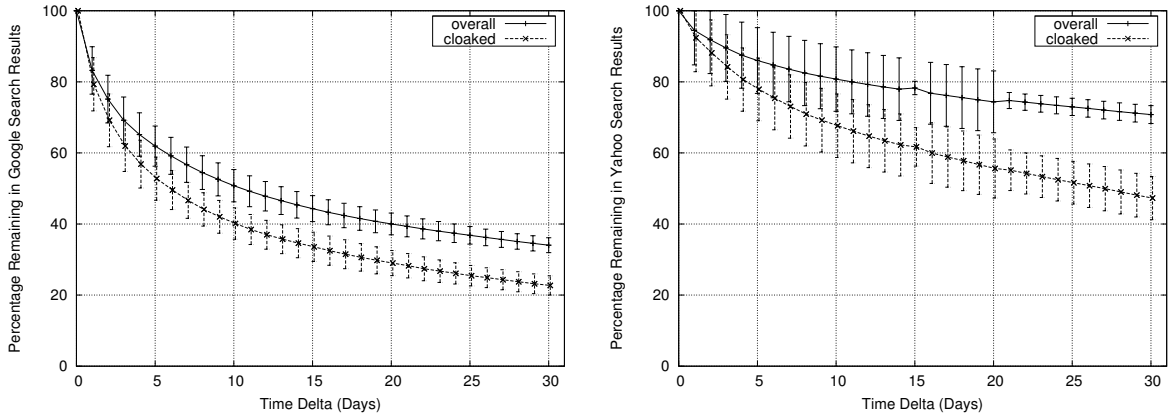
From the graphs, we see that, through May 4th, using search terms from Google Suggest, seeded initially from Google Hot Searches, uncovers the most cloaking. For Google search results, averaged across the days, Google Suggest returns 3.5× as many cloaked search results as Google Hot Searches alone, 2.6× as Twitter, and 3.1× as Alexa. Similarly, even when using Yahoo, Google Suggest returns 3.1× as many cloaked search results as Google Hot Searches alone, 2.6× as Twitter, and 2.7× as Alexa. As discussed in Section 3.1, cloakers targeting popular search terms face stiff SEO competition from others (both legitimate and illegitimate) also targeting those same terms. By augmenting popular search terms with suggestions, cloakers are able to target the same semantic topic as popular search terms. Yet, because the suggestion is essentially an autocomplete, it possesses the long-tail search benefits of reduced competition while remaining semantically relevant.

The above results demonstrate that the prevalence of cloaking in search results is highly influenced by the search terms. As another perspective, for each measurement period that crawls the search terms at a given point in time, we can count the number of cloaked results returned for each search term. Averaging across all measurement periods, 23% and 14% of the search terms accounted for 80% of the cloaked results from Google and Yahoo, respectively. For reference, Table 1 lists the results for the top 10 search terms on Google and Figure 3 shows the distribution of the percentage of cloaked search results for pharmaceutical search terms. The query "viagra 50mg canada" is the pharmaceutical term with the largest percentage of cloaked search results on Google with 61%. Yet, the median query "tramadol 50mg" contains only 7% of cloaked search results. Note that the percentages sum to much more than 100% since different search terms can return links to the same cloaked



(a) Google (b) Yahoo

Figure 4: Churn in the top 100 cloaked search results and overall search results for trending search terms.



(a) Google (b) Yahoo

Figure 5: Churn in the top 100 cloaked search results and overall search results for pharmaceutical search terms.

pages. As an example in Figure 3, the sixth point shows the average percentage of cloaked search results, across all measurements, for the search term with the sixth highest percentage of cloaked search results. We plot the first 10 points with the most significant percentage of cloaked search results, then plot every 10th search term, for clarity. From these results we see high variance in the percentage of cloaked search results.

4.3 Search Engine Response

Next we examine how long cloaked pages remain in search results after they first appear. For a variety of reasons, search engines try to identify and thwart cloaking. Although we have little insight into the techniques used by search engines to identify cloaking,[4] we can still observe the external effects of such techniques in practice.

We consider cloaked search results to have been effectively "cleaned" by search engines when the cloaked search result no longer appears in the top 100 results. Of course, this indicator may not be directly due to the page having been cloaked. The search engine ranking algorithms could have adjusted the positions of cloaked pages over time due to other factors, e.g., the SEO techniques used by cloakers may turn out to be useful only in the short term. Either way, in this case we consider the cloaked pages as no longer being effective at meeting the goals of the cloakers.

[4]Google has a patent in the area [13], but we have not seen evidence of such a client-assisted approach used in practice.

To measure the lifetime of cloaked search results, we perform repeated search queries for every search term over time (Section 3.5). We then examine each new set of search results to look for the cloaked results we originally detected. The later search results will contain any updates, including removals and suppressions, that search engines have made since the time of the initial measurement. To establish a baseline we also measure the lifetime of all our original search results, cloaked or not. This baseline allows us to differentiate any churn that occurs naturally with those attributable to "cleaning". We perform these repeated searches on each term every four hours up to 24 hours and then every day up to seven days.

Figures 4a and 4b show the lifetime of cloaked and overall search results for Google and Yahoo for trending searches. Each point shows the average percentage of search results that remain in the top 100 for the same search terms over time. The error bars denote the standard devation across all searches, and we plot the points for "cloaked" slightly off-center to better distinguish error bars on different curves. The results, for both search engines, show that cloaked search results rapidly begin to fall out of the top 100 within the first day, with a more gradual drop thereafter. In contrast, search results in general have similar trends, but decline more gradually. For Google, nearly 40% of cloaked search results have a lifetime of a day or less, and over the next six days only an additional 25% drop from the top 100 results. In contrast, for the baseline only 20% of overall search results have a lifetime of a day of less, and

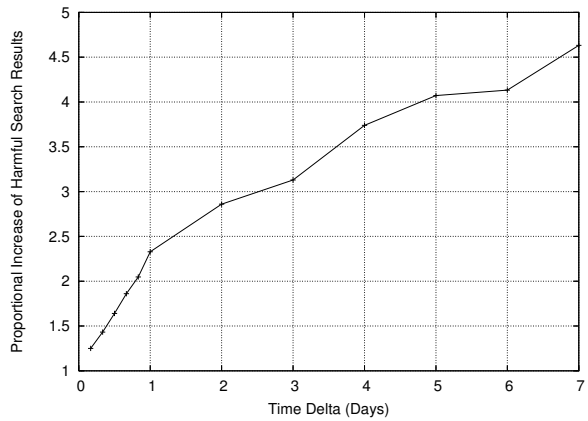


Figure 6: Increase in harmful trending search results over time on Google as labeled by Google Safe Browsing.

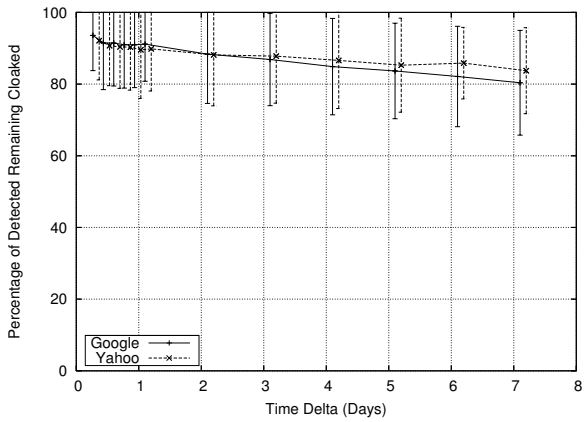


Figure 7: Duration pages are cloaked.

an additional 15% are cleaned after the next six days. Yahoo exhibits a similar trend, although with less rapid churn and with a smaller separation between cloaked and the baseline (perhaps reflecting differences in how the two search engines deal with cloaking). Overall, though, while cloaked pages do regularly appear in search results, many are removed or suppressed by the search engines within hours to a day.

Figures 5a and 5b show similar results for pharmaceutical searches. Note that the maximum time delta is 30 days because, unlike trending terms, the pharmacy search terms do not change throughout the duration of our experiment and we have a larger window of observation. While we still see similar trends, where cloaked search results drop more rapidly than the churn rate and Google churns more than Yahoo, the response for both Google and Yahoo is slower for pharmaceutical terms than for trending terms. For example, whereas 45% and 25% of cloaked trending search results were "cleaned" for Google and Yahoo, respectively, within two days, only 30% and 10% of cloaked pharmacy search results were "cleaned" for Google and Yahoo, respectively.

As another perspective on "cleaning", Google Safe Browsing [7] is a mechanism for shielding users by labeling search results that lead to phishing and malware pages as "harmful". These harmful search results sometimes employ cloaking, which Google Safe Browsing is able to detect and bypass. This insight suggests that the rate that Google is able to discover and label "harmful" search results correlates with the rate at which they can detect cloaking. We can measure this Safe Browsing detection by repeatedly querying for the same terms as described in Section 3.5 and counting the number of "harmful" search results.

As observed in Section 4.2, the prevalence of cloaking is volatile and depends heavily on the specific search terms. The prevalence of detected harmful pages is similarly volatile; although 37% of the results on average on Google are marked as harmful for the terms we search for, there is substantial variance across terms. Therefore, we normalize the change over time in the number of harmful search results labeled by Google Safe Browsing relative to the first measurement. Figure 6 shows the average normalized change in the number of harmful labels over time, across all queries on trending search terms. The number of harmful labels increases rapidly for the first day, with nearly 2.5× more labels than the original measurement, and then increases steadily over the remaining six days, where there are nearly 5× more labels than the original query. This behavior mirrors the results on cleaning above.

4.4 Cloaking Duration

Cloakers will often subvert existing pages as an SEO technique to capitalize on the already established good reputation of those pages with search engines. We have seen that the search engines respond relatively quickly to having cloaked pages in search results. Next we examine how long until cloaked pages are no longer cloaked, either because cloakers decided to stop cloaking or because a subverted cloaked page was discovered and fixed by the original owner.

To measure the duration that pages are cloaked, we repeatedly crawl every cloaked page that we find over time, independent of whether the page continues to appear in the top 100 search results. We then apply the cloaking detection algorithm on the page (Section 3.4), and record when it is no longer cloaked. As in Section 4.3, we crawl each page every four hours up to 24 hours and then every day up to seven days.

Figure 7 shows the time durations that pages are cloaked in results returned by Google and Yahoo. Each point shows the percentage of all cloaked pages for each measurement period that remain cloaked over time, and the error bars show the standard deviations across measurement periods. We see that cloaked pages have similar durations for both search engines: cloakers manage their pages similarly independent of the search engine. Further, pages are cloaked for long durations: over 80% remain cloaked past seven days. This result is not very surprising given that cloakers have little incentive to stop cloaking a page. Cloakers will want to maximize the time that they might benefit from having a page cloaked by attracting customers to scam sites, or victims to malware sites. Further, it is difficult for them to recycle a cloaked page to reuse at a later time. Being blacklisted by Google Safe Browsing, for instance, requires manual intervention to regain a positive reputation. And for those cloaked pages that were subverted, by definition it is difficult for the original page owners to detect that their page has been subverted. Only if the original page owners access their page as a search result link will they realize that their page has been subverted; accessing it any other way will return the original contents that they expect.

4.5 Cloaked Content

Since the main goal of cloaking as an SEO technique is to obtain user traffic, it is natural to wonder where the traffic is heading. By looking at the kind of content delivered to the user from cloaked search results, not only does it suggest why cloaking is necessary

Category	% Cloaked Pages
Traffic Sale	81.5%
Error	7.3%
Legitimate	3.5%
Software	2.2%
SEO-ed Business	2.0%
PPC	1.3%
Fake-AV	1.2%
CPALead	0.6%
Insurance	0.3%
Link farm	0.1%

Table 2: Breakdown of cloaked content for manually-inspected cloaked search results from Google for trending search terms. Note that "Traffic Sale" pages are the start of redirection chains that typically lead to Fake-AV, CPALead, and PPC landing pages.

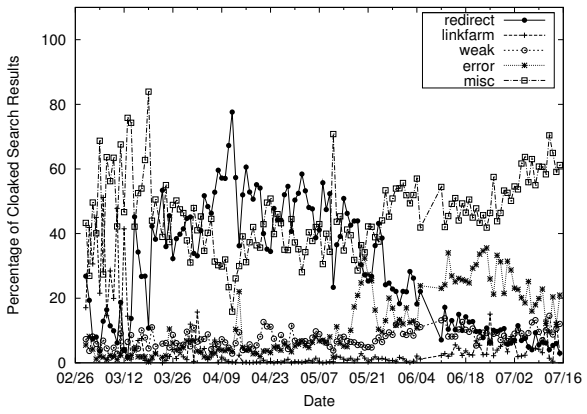


Figure 8: Proportional distribution of cloaked search results in Google over time for trending searches.

for hiding such content, but it also reveals the motives cloakers have in attracting users.

We have no fully automated means for identifying the content behind cloaked search results. Instead, we cluster cloaked search results with the exact same DOM structure of the pages as seen by the user when clicking on a search result. We perform the clustering on all cloaked search results from Google across all measurement points for trending searches. To form a representative set of cloaked pages for each cluster, we select a handful of search results from various measurement times (weekday, weekend, dayime, morning, etc.) and with various URL characteristics. We then manually label pages in this representative set to classify the content of the pages being cloaked.

We manually label the content of each cluster into one of ten categories: traffic sales, pay-per-click (PPC), software, insurance, Fake-AV, CPALead,[5] link farm, SEO-ed business, error, and legitimate. Traffic sales are cloaked search results with the sole purpose of redirecting users through a chain of advertising networks, mainly using JavaScript, before arriving at a final landing page. Although we are unable to follow them systematically, from manually examining thousands of traffic sales, we observed these search results directing users primarily to Fake-AV, CPALead, and PPC pages. Occasionally cloaked search results do not funnel users through a redirection chain, which is how we are able to classify the PPC, software, insurance, Fake-AV, and CPALead sets. The link farm set contain benign pages that provide many backlinks to boost the rankings of beneficiary pages. The SEO-ed business refers to businesses that employ black-hat SEO techniques, such as utilizing free hosting to spam a large set of search results for a single term. The errors are pages that have client side requirements we were unable to meet, i.e., having an Adobe Flash plugin. Finally, the legitimate set refers to pages that display no malicious behavior but were labeled as cloaking due to delivering differing content, as is the case when sites require users to login before accessing the content.

Table 2 shows the breakdown of cloaked search results after manually inspecting the top 62 clusters, out of 7671 total, which were sorted in decreasing order of cluster size. These 62 clusters account for 61% of all cloaked search results found in Google for trending searches across all measurement points. From this, we see that about half of the time a cloaked search result leads to some form of abuse. Further, over 49% of the time, cloaked search results sell user traffic through advertising networks, which will eventually lead to Fake-AV, CPALeads, or PPC.

Interestingly, the DOM structure of the largest cluster, which alone acounted for 34% of cloaked search results, was a single JavaScript snippet that performs redirection as part of traffic sales. Other large clusters that accounted for 1–3% of cloaked search results also consist primarily of JavaScript that performs redirection as part of traffic sales.

Despite the fact that the clusters we have examined account for 61% of all cloaked search results, there still exists 39% that have not been categorized and likely do not share the same distribution. While incrementally clustering, we noted that the top clusters grew larger and larger as more and more data was added. This suggests the presence of long-term SEO campaigns, as represented by the top clusters, that constantly change the search terms they are targeting and the hosts they are using. Therefore, since the uncategorized search results fall within the long tail, they are unlikely to be actively involved in direct traffic sales. Instead, we speculate that they fall in the link farm or legitimate sets given that those groups have the most difficult time in forming large clusters because they are not being SEO-ed as heavily across search terms.

The kinds of pages being cloaked is also dynamic over time. Figure 8 shows the classification of cloaked page content for search results from Google using trending terms, from March 1, 2011 through July 15, 2011. We classify the HTML of cloaked pages, using the file size and substrings as features, into one of the following categories: Linkfarm, Redirect, Error, Weak, or Misc. The Linkfarm category represents content returned to our "Googlebot" crawler that contains many outbound links hidden using CSS properties. The Redirect category represents content returned to a user that is smaller than 4 KB and contains JavaScript code for redirection, or an HTML meta refresh. The Error category represents user content that is smaller than 4 KB and contains a blank page or typical error messages. The Weak category contains user content below 4 KB in file size not already classified; similarly, the Misc category contains user content larger than 4 KB not already classified. As an example, on March 7th approximately 7% of the cloaked content detected were linkfarms, 53% were redirects, 6% were errors, 3% were weak and 31% were misc.

Looking at the trends over time reveals the dynamic nature of the content being hidden by cloaking. In particular, we saw a surge in redirects from March 15th to June 5th. During this period, the average distribution of redirects per day increased from 11.4% to

[5]Cost-per-action pages that ask a user to take some action, such as filling out a form, that will generate advertising revenue.

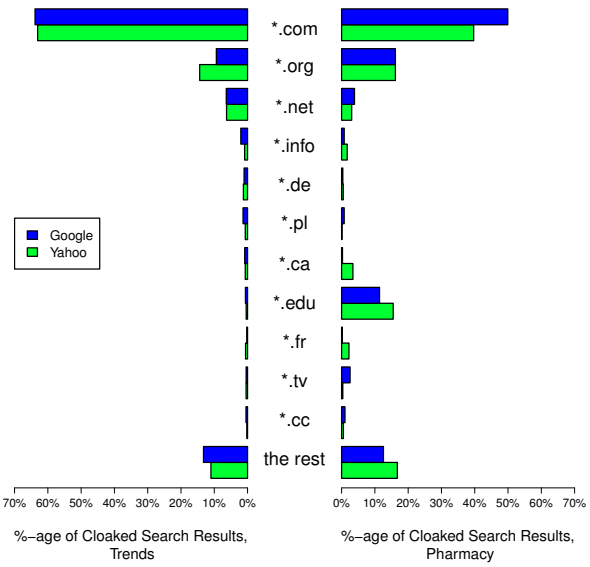


Figure 9: Histogram of the most frequently occuring TLDs among cloaked search results.

41.3% and later dropped off to 8.7%. Interestingly, as redirects begin to fall off, we see a corresponding increase in errors. During the high period of redirects, errors represented 8.0% of the average distribution, but afterwards represented 24.3%. One explanation of this correlation is that the infrastructure supporting redirects begins to collapse at this point. Anecdotally, this behavior matches the general time frame of the undermining of key Fake-AV affiliate programs [9], frequently observed at the end of the redirect chains.

4.6 Domain Infrastructure

Analyzing the underlying intent of cloaked pages confirmed again that cloakers are attempting to attract traffic by illegitimately occupying top search result positions for trending and pharmacy search terms and their suggestions. The implication is that the key resource that spammers must possess, to effectively utilize cloaking in their scams, is access to Web sites and their domains. Ideally, these sites should be established sites already indexed in search engines. Otherwise, solely using traditional SEO tactics, such as link farming, will have limited success in obtaining high search result positions. Recent reports confirm that many pages have been targeted and infected by well known exploits to their software platforms, and subsequently used to cloak hidden content from search engines [18].

In this section, we examine the top level domains (TLDs) of cloaked search results. Figure 9 shows histograms of the most frequently occuring TLDs among all cloaked search results, for both Google and Yahoo. We see that the majority of cloaked search results are in `.com`. Interestingly, cloaked search results from pharmaceutical queries utilize domains in `.edu` and `.org` much more frequently, where together they represent 27.6% of all cloaked search results seen in Google and 31.7% in Yahoo. For comparison, `.edu` and `.org` together represent just 10% in Google and 14.8% in Yahoo for trending searches. Cloakers spamming pharmaceutical search terms are using the "reputation" of pages in these domains to boost their ranking in search results similar to the recent accusations against `overstock.com` [6].

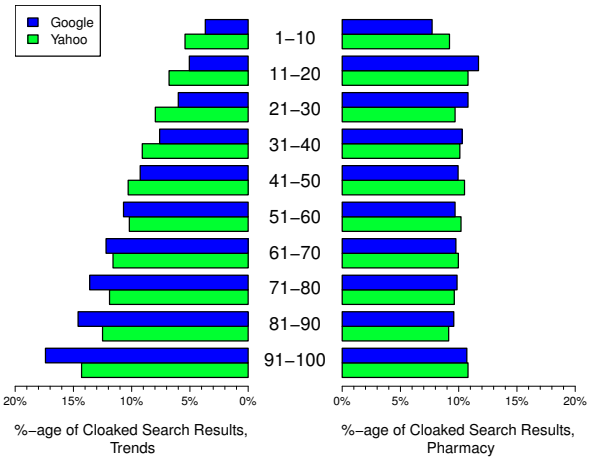


Figure 10: Distribution of cloaked search result positions.

4.7 SEO

Finally, we explore cloaking from an SEO perspective by quantifying how successful cloaking is in high-level spam campaigns. Since a major motivation for cloaking is to attract user traffic, we can extrapolate SEO performance based on the search result positions the cloaked pages occupy. For example, a campaign that is able to occupy search result positions between 1–20 is presumably much more successful than one that is only able to occupy search result positions between 41–60.

To visualize this information, we calculate the percentage of cloaked search results found between ranges of search result positions for each measurement. Then we take the average across all measurements. Again, we only focus on Google and Yahoo due to the lack of cloaked search results in Bing. For clarity, we bin the histogram by grouping every ten positions together, the default number of search results per page.

Figure 10 shows the resulting histograms for trending search terms and pharmaceutical terms, side by side. For trending searches we see a skewed distribution where cloaked search results mainly hold the bottom positions; for both Google and Yahoo, positions further away from the top contain more cloaking. Compared to the positions 1–10 on the first page of results, the number of cloaked search results are 2.1× more likley to hold a search result positon between 31–40 in Google, and 4.7× more likely to be in position 91–100; results for Yahoo are similar. In some ways, this distribution indicates that cloaking is not very effective for trending search terms. It does not lead to a higher concentration in the most desirable search result positions (top 20), likely due to the limited amount of time available to SEO. Although cloakers will have fewer opportunities for their scams as a result, presumably it still remains profitable for cloakers to continue the practice.

Interestingly, we see a very different trend in pharmaceutical searches where there is an even distribution across positions. The number of cloaked pages are just as likely to rank in the first group of search results (positions 1–10) as any other group within the top 100. Wu and Davison [24] had similar findings from 2005. One possible explanation is that the differences again reflect the differences in the nature of cloaked search terms. Cloaking the trending terms by definition target popular terms that are very dynamic, with limited time and heavy competition for performing SEO on those search terms. Cloaking pharmacy terms, however, is a highly focused task on a static set of terms, providing much longer time frames for performing SEO on cloaked pages for those terms. As

a result, cloakers have more time to SEO pages that subsequently span the full range of search result positions.

5. CONCLUSION

Cloaking has become a standard tool in the scammer's toolbox and one that adds significant complexity for differentiating legitimate Web content from fraudulent pages. Our work has examined the current state of search engine cloaking as used to support Web spam, identified new techniques for identifying it (via the search engine snippets that identify keyword-related content found at the time of crawling) and, most importantly, we have explored the dynamics of cloaked search results and sites over time. We demonstrate that the majority of cloaked search results remain high in rankings for 12 hours and that the pages themselves can persist far longer. Thus, cloaking is likely to be an effective mechanism so long as the overhead of site placement via SEO techniques is less than the revenue obtained from 12 hours of traffic for popular keywords. We believe it is likely that this holds, and search engine providers will need to further reduce the lifetime of cloaked results to demonetize the underlying scam activity.

Acknowledgments

We thank Damon McCoy for insightful comments and discussion of this work, and we also thank the anonymous reviewers for their valuable feedback. This work was supported in part by National Science Foundation grants NSF-0433668 and NSF-0831138, by the Office of Naval Research MURI grant N000140911081, and by generous research, operational and/or in-kind support from Google, Microsoft, Yahoo, Cisco, HP and the UCSD Center for Networked Systems (CNS).

6. REFERENCES

[1] John Bethencourt, Jason Franklin, and Mary Vernon. Mapping Internet Sensors with Probe Response Attacks. In *Proceedings of the 14th USENIX Security Symposium*, Baltimore, MD, July 2005.

[2] Andrei Z. Broder. On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29, June 1997.

[3] Lee G. Caldwell. *The Fast Track to Profit*. Pearson Education, 2002.

[4] Kumar Chellapilla and David Maxwell Chickering. Improving Cloaking Detection Using Search Query Popularity and Monetizability. In *Proceedings of the SIGIR Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, August 2006.

[5] Marco Cova, Corrado Leita, Olivier Thonnard, Angelos Keromytis, and Marc Dacier. An Analysis of Rogue AV Campaigns. In *Proceedings of the 13th International Symposium on Recent Advances in Intrusion Detection (RAID)*, September 2010.

[6] Amir Efrati. Google Penalizes Overstock for Search Tactics. `http://online.wsj.com/article/SB10001424052748704520504576162753779521700.html`, February 24, 2011.

[7] Google Safe Browsing API. `http://code.google.com/apis/safebrowsing/`.

[8] John P. John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martin Abadi. deSEO: Combating Search-Result Poisoning. In *Proceedings of the 20th USENIX Security Symposium*, August 2011.

[9] Brian Krebs. Huge Decline in Fake AV Following Credit Card Processing Shakeup. `http://krebsonsecurity.com/2011/08/huge-decline-in-fake-av-following-credit-card-processing-shakeup/`, August 2011.

[10] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade. In *Proceedings of the 20th USENIX Security Symposium*, August 2011.

[11] Kirill Levchenko, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Andreas Pitsillidis, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium and Security and Privacy*, Oakland, CA, May 2011.

[12] Jun-Lin Lin. Detection of cloaked web spam by using tag-based methods. *Expert Systems with Applications*, 36(4):7493–7499, 2009.

[13] Marc A. Najork. System and method for identifying cloaked web servers, United States Patent number 6,910,077. Issued June 21, 2005.

[14] Yuan Niu, Yi-Min Wang, Hao Chen, Ming Ma, and Francis Hsu. A Quantitative Study of Forum Spamming Using Contextbased Analysis. In *Proceedings of 15th Network and Distributed System Security (NDSS) Symposium*, February 2007.

[15] Moheeb Abu Rajab, Lucas Ballard, Panayiotis Mavrommatis, Niels Provos, and Xin Zhao. The Nocebo Effect on the Web: An Analysis of Fake Anti-Virus Distribution. In *Proceedings of the 3rd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET'10)*, April 2010.

[16] Search Engine Marketing Professional Organization (SEMPO). State of Search Engine Marketing Report Says Industry to Grow from $14.6 Billion in 2009 to $16.6 Billion in 2010. `http://www.sempo.org/news/03-25-10`, March 2010.

[17] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[18] Julien Sobrier. Tricks to easily detect malware and scams in search results. `http://research.zscaler.com/2010/06/tricks-to-easily-detect-malware-and.html`, June 3, 2010.

[19] Danny Sullivan. Search Engine Optimization Firm Sold For $95 Million. `http://searchenginewatch.com/2163001`, September 2000. Search Engine Watch.

[20] Jason Tabeling. Keyword Phrase Value: Click-Throughs vs. Conversions. `http://searchenginewatch.com/3641985`, March 8, 2011.

[21] Yi-Min Wang and Ming Ma. Detecting Stealth Web Pages That Use Click-Through Cloaking. Technical Report MSR-TR-2006-178, Microsoft Research, December 2006.

[22] Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. Spam Double-Funnel: Connecting Web Spammers with Advertisers. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*, pages 291–300, May 2007.

[23] Wordtracker. Five Reasons Why Wordtracker Blows Other

Keywords Tools Away. `http://www.wordtracker.com/find-the-best-keywords.html`.

[24] Baoning Wu and Brian D. Davison. Cloaking and Redirection: A Preliminary Study. In *Proceedings of the SIGIR Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2005.

[25] Baoning Wu and Brian D. Davison. Detecting Semantic Cloaking on the Web. In *Proceedings of the 15th International World Wide Web Conference*, pages 819–828, May 2006.

APPENDIX

A. CLOAKING ILLUSTRATION

Figure 11 uses browser screenshots to illustrate the effects of cloaking when searching on Google on May 3, 2011 with the terms "bethenney frankel twitter", a topic popular in user queries at that time. The top screenshot shows the first search results page with the eighth result highlighted with content snippets and a preview of the linked page; this view of the page corresponds to the cloaked content returned to the Google crawler when it visited this page. These search results are also examples of what Dagger obtains when querying Google with search terms (Section 3.2).

The middle screenshot shows the page obtained by clicking on the link using Internet Explorer on Windows. Specifically, it corresponds to visiting the page with the `User-Agent` field set to:

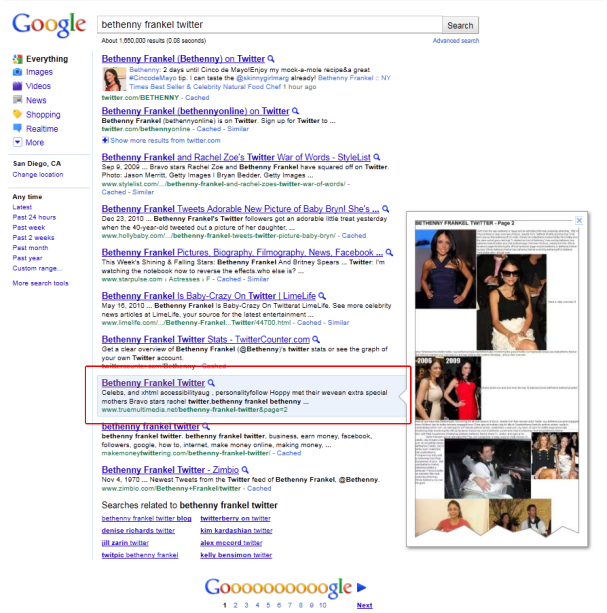
```
Mozilla/5.0 (compatible; MSIE 8.0; Windows NT 5.2;
Trident/4.0; Media Center PC 4.0; SLCC1; .NET CLR
3.0.04320)
```

and the `Referer` field indicating that the click comes from a search result for the terms "bethenney frankel twitter". This page visit corresponds to when Dagger crawls the URL as a search "user" (Section 3.3), displaying an advertisement for Fake-AV.

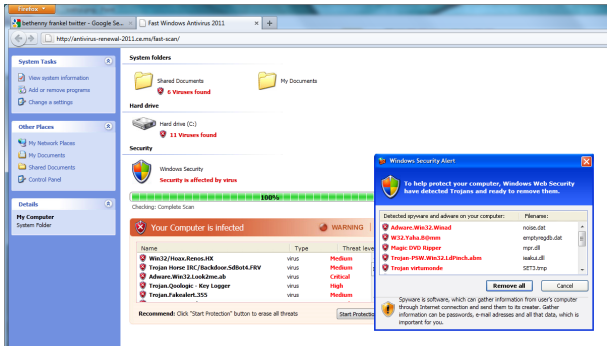
The bottom screenshot shows the page obtained by crawling the link while mimicking the Google crawler. Specifically, it corresponds to visiting the page with the `User-Agent` field set to:

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

and corresponds to when Dagger visits the URL as a search "crawler". Note that this page also uses IP cloaking: the content returned to Dagger, which does not visit it with a Google IP address, is different from the content returned to the real Google search crawler (as reflected in the snippet and preview in the top screenshot). Nevertheless, because the cloaked page does return different content to search users and the Dagger crawler, Dagger can still detect that the page is cloaked.



(a) Google Search Results Page



(b) User from Windows

(c) Crawler

Figure 11: Example of cloaking in practice: (a) the first search results page for the query "bethenney frankel twitter", including the Google preview; (b) the page obtained by clicking on the highlighted link from a Windows machine (with `Referer` indicating a search result); and (c) the same page visited but with the `User-Agent` field in the HTTP request header set to mimic that of the Google search crawler.