Characterizing Long-tail SEO Spam on Cloud Web Hosting Services

Xiaojing Liao
Georgia Institute of Technology
xliao@gatech.edu

Chang Liu
University of Maryland
liuchang@cs.umd.edu

Damon McCoy
New York University
mccoy@nyu.edu

Elaine Shi
Cornell University
runting@gmail.com

Shuang Hao
University of California, Santa Barbara
shuanghao@cs.ucsb.edu

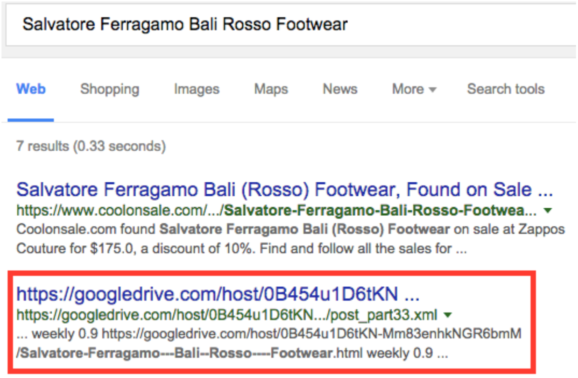
Raheem Beyah
Georgia Institute of Technology
rbeyah@ece.gatech.edu

ABSTRACT

The popularity of long-tail search engine optimization (SEO) brings with new security challenges: incidents of long-tail keyword poisoning to lower competition and increase revenue have been reported. The emergence of cloud web hosting services provides a new and effective platform for long-tail SEO spam attacks. There is growing evidence that large-scale long-tail SEO campaigns are being carried out on cloud hosting platforms because they offer low-cost, high-speed hosting services. In this paper, we take the first step toward understanding how long-tail SEO spam is implemented on cloud hosting platforms. After identifying 3,186 cloud directories and 318,470 doorway pages on the leading cloud platforms for long-tail SEO spam, we characterize their abusive behavior. One highlight of our findings is the effectiveness of the cloud-based long-tail SEO spam, with 6% of the doorway pages successfully appearing in the top 10 search results of the poisoned long-tail keywords. Examples of other important discoveries include how such doorway pages monetize traffic and their ability to manage cloud platform's countermeasures. These findings bring such abuse to the spotlight and provide some insights to eliminating this practice.

1. INTRODUCTION

Long-tail Search Engine Optimization (SEO) provides an opportunity for online advertisers to target niche markets. Instead of traditional SEO that targets a single keyword or shorter keyword phrases, long-tail SEO targets longer and more specific keyword phrases that tend to be directly related to specific products and locations. For example, a furniture marketing web page using long-tail SEO might target a more specific keyword phrase "contemporary Art Decoinfluenced semicircle lounge" rather than targeting "furniture". The advantages of long-tail SEO are that there is less competition for higher search rankings and it has been shown that specific searches are far more likely to convert to sales than generic searches [20]. As with most profitable online segments, long-tailed search results are being polluted by search engine spammers that manipulate search engine results using blackhat long-tail SEO techniques.



Salvatore Ferragamo Bali Rosso Footwear

7 results (0.33 seconds)

Salvatore Ferragamo Bali (Rosso) Footwear, Found on Sale ...
https://www.coolonsale.com/.../**Salvatore-Ferragamo-Bali-Rosso-Footwea**...
Coolonsale.com found **Salvatore Ferragamo Bali (Rosso) Footwear** on sale at Zappos Couture for $175.0, a discount of 10%. Find and follow all the sales for ...

https://googledrive.com/host/0B454u1D6tKN ...
https://googledrive.com/host/0B454u1D6tKN.../post_part33.xml
... weekly 0.9 https://googledrive.com/host/0B454u1D6tKN-Mm83enhkNGR6bmM **/Salvatore-Ferragamo---Bali--Rosso----Footwear**.html weekly 0.9 ...

Figure 1: Example of long-tail poisoning utilizing cloud hosting platform. The second result returned here are doorway page hosting on Google's cloud hosting platform Google Drive.

While long-tail SEO spamming has been an ongoing issue, the emergence of cloud web hosting services, such as Amazon S3 and Google Drive, provides a new and effective platform for dispersing long-tail SEO spam. The attractiveness of cloud hosting is that it offers fast, reliable and cheap (sometimes free) hosting. In addition, they provide a domain name that is shared by many of their users. This makes it infeasible to blacklist all content from a cloud hosting provider, which causes blacklist maintainers to expend more effort to build finer grained blacklists. Figure 1 shows an example of long-tail poisoning utilizing Google Drive, in which the second search result obtained from the long-tail keyword query "Salvatore Ferragamo Bali Rosso Footwear" is a doorway page with no useful content and affiliate links that are categorized as search spam by most search engines. Although there are indications of the presence of long-tail SEO spam in cloud hosting, characterizing the details of how such a spam attack is mounted, its effectiveness and spam-

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883008.

mers' ability to evade cloud platform's countermeasures have not been documented.

In this paper, we conduct the first measurement study of long-tail SEO spam hosted on cloud platforms. We bootstrapped our study by identifying spam *cloud directories* on cloud hosting platforms, in which doorway pages have largely homogeneous content in terms of their keywords and DOM structures. This enabled us to locate 930 spam cloud directories on Amazon S3 and 672 spam cloud directories on Google Drive, as well as other cloud platforms. Our analysis of the doorway pages' content revealed that they were utilizing relatively unsophisticated blackhat SEO techniques, such as keyword stuffing (which is the repetition of keyword phrases multiple times) and keyword spam (which includes unrelated keywords). Also, we found that the SEO spammers made use of evasion techniques, such as link shorteners and obfuscated client-side JavaScript to hide affiliate links when cloud platforms do not support server-side scripting.

In order to understand the effectiveness of these long-tail SEO spam campaigns, we monitored 236,368 long-tailed keyword searches over the course of one year. Based on our analysis, we observed that 6% of the cloud-hosted doorway pages polluted the top 10 search results of long-tail keywords, and 32% of the top 100 search results. These doorway pages indicate the high-level of effectiveness of polluting long-tailed search results. We also found that almost all of the doorway pages were monetized by including links to reputable affiliate programs such as Prosperent, ClickBank and VigLink.

To understand the profitability of long-tail SEO spam on cloud hosting platforms, we analyzed the estimated revenue and click-through rate for a single campaign, which showed spammers were earning a modest sum of approximately $400 USD each per month. In addition, we noted that their click-through rates were increasing by 20% over time. Finally, we monitored ongoing interventions by the cloud service providers. We found that service providers' efforts to detect and remove doorway pages had limited effectiveness, as long-tail SEO campaigns remained active. Doorway pages on cloud hosting platforms have an average lifetime of 7 weeks, which is *much longer* than those hosted on traditional platforms (i.e., 1 week [10]).

To the best of our knowledge, our study is the first to present a comprehensive understanding of long-tail SEO spam on cloud web hosting platforms and its effects. We summarize our main contributions as follows:

1. We propose a methodology to identify cloud directories containing long-tail SEO spam, which discovered 3,186 abusive cloud directories on 10 mainstream cloud platforms.
2. We conduct a measurement study of long-tail SEO spam on the cloud, which provides insights into its effectiveness, its use of cloud resources, network characteristics and revenue models.
3. Our empirical study shows that the cloud service provider's efforts to prevent these abusive usages are yet to be effective.

The rest of the paper is organized as follows: Section 2 presents the background information and adversary model for our research, while the method by which we collected data and identified spam cloud directories is discussed in Section 3. Section 4 reports the details of our analysis about long-tail SEO effectiveness on the cloud platforms, followed by an analysis of traffic monetization in Section 5. Section 6 reports the effectiveness of interventions conducted by cloud service providers, while Section 7 discusses the limitations of our technique and potential future work. The paper concludes with a look at related works and a brief summary of the paper's findings in Section 8 and Section 9.

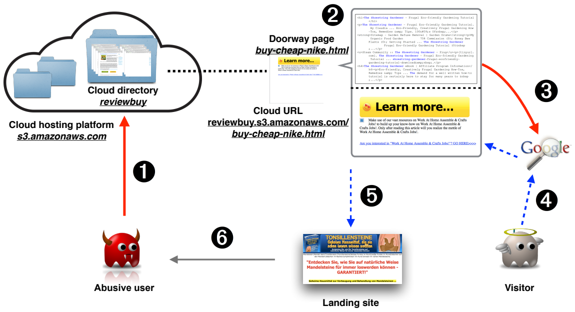


Figure 2: An example of long-tail SEO on a cloud hosting platform.

2. BACKGROUND

In this section, we present some background information on three areas of importance to this paper. First, we explain the basics of cloud hosting, followed by why long-tail SEO is beneficial to both web page hosts and potential audiences. Lastly, we present the adversarial model used in planning our study.

Cloud hosting. Cloud hosting is a type of "infrastructure as a service (IaaS)", which is rented by a cloud user to host her web page. These web pages are organized into *cloud directories* identified by unique, user-assigned keys that are mapped as unique sub-domains. The web page stored in the cloud directories can be served directly to users via file names in a relative path (i.e., *cloud URL*). This process is known as built-in site publishing [9]. For instance, an HTML file hosted in a cloud directory can be directly run in a browser and visited by the public as a web page via the cloud URL.

In recent years we have seen an increase in popularity of cloud hosting services. Pay-as-you-go cloud hosting is well received as an economic and flexible computing solution. As an example, Google Drive today offers a free web hosting service with 15GB of storage, and an additional 100GB for $1.99/month, and GoDaddy's web hosting starts from merely $1/month for 100GB. The pay-as-you-go feature on cloud web hosting enables multiple low-cost permanent or temporary websites such as start-up websites (e.g., yelp), research project websites (e.g., NASA/JPL's Mars Curiosity Mission) and political campaign websites (e.g., Obama for America Campaign 2012). Additionally, spam campaigns also utilize cloud web hosting for marketing promotion.

Long-tail SEO. Long-tail SEO optimizes *doorway pages* for longer and more specific keyword phrases (i.e., long-tail keyword). With long-tail keywords, a doorway page can attract exactly the audience looking for that specific product, and as a result, that audience will be far closer to point-of-purchase [20]. Also, compared with shorter keywords, competition for rankings can be less fierce, and the doorway page can more easily achieve a high search ranking.

For example, a doorway page to promote classic furniture is highly unlikely to appear near the top of an organic search for "furniture" because there is too much competition. But if the doorway page specializes in, say, contemporary art-deco furniture, then long-tail keywords like "contemporary Art Deco-influenced semi-circle lounge" are going to reliably find those consumers looking for that exact product.

Adversary model. In our research, we consider the abusive users who try to use cloud web hosting service for long-tail SEO spam. For this purpose, an abusive user could build her own cloud directories to store a large amount of *doorway pages*, which are optimized for long-tail keywords.

Figure 2 illustrates an example of long-tail SEO spam on cloud hosting service. An abusive user creates a cloud directory on the cloud hosting platform and uploads large amount of doorway pages for long-tail SEO spam (❶). To attract clicks, an abusive user would utilize blackhat SEO techniques to pollute the search engine's long-tail keywords (❷) and manipulate the search ranking (❸). When visiting the doorway page from the poisoned search engine results (❹), a visitor will be redirected to a landing site (❺) from which the abusive user will obtain a marketing commission (❻).

3. ABUSIVE CLOUD DIRECTORY IDENTIFICATION

In this section, we explain the methodology used in our study for abusive cloud directory identification. In the data collection stage, we first selected SEO targeted keywords to feed the search engine to identify the doorway pages on the cloud hosting service. Then, we utilized the directory structure of cloud hosting service to find other doorway pages. In the abusive cloud directory identification stage, since the long-tail SEO campaigns show high similarity in page contents in the same directories, we trained a classifier to identify the cloud directories hosting long-tail SEO spam.

3.1 Data Collection

In the data collection stage, we first collected the 'seed' web pages on the cloud hosting service. Specifically, we fed the SEO targeted keywords to the search engine, and used the Google Web Search API to pull the links that appeared in the search results. Second, since the web pages on the cloud platforms are organized into directories, we also crawled additional web pages in the same directories. Then, a web crawler followed the links in the page, collected their redirection chains, and stored the intermediate URL information in our local database.

Seed Data Collection. Selecting appropriate keyword phrases to feed the search engine is critical for obtaining representative results. To analyze the long-tail SEO spam in cloud hosting services, we first choose 'hot' keyword phrases and spammy keywords phrases. These keywords reflect what people are searching for and what SEOs are targeting. Further, we use the Google Web Search API to pull the top 100 search results for each term from the Google search engine. In this paper, we analyze the long-tail SEO spam on 10 leading cloud hosting services as listed in Table 1. This set of crawled pages is defined as a seed dataset D^s, which contains 32,177 cloud URLs and 20,328 cloud directories.

Table 1: List of cloud hosting platforms.

Cloud Platform	Domain
Heroku	herokuapp.com
Amazon S3	s3.amazonaws.com
Dropbox	kissr.com
Azure	azurewebsites.net
Google	googledrive.com
Openshift	rhcloud.com
Bitbucket	bitbucket.org
Sina	sinaapp.com
Baiduyun	duapp.com
Olympe	olympe.in

Table 2: Summary results of the datasets.

Name	# of URLs	# of cloud directories	# of keywords
D^s	32,177	20,328	1,500
D^d	1,073,642	15,774	NaN

For the first set of search terms, we employ popular trending keywords from Google Trend hot keywords [6]. We collect the top 20 popular search terms in 64 categories across various search interests including entertainment, education and technology. For the second set of search terms, we target some specific keywords which spammers also target. We utilized a spam trigger word list [4], which includes 200 spammy words such as "payday loan" and "casino no deposit". In addition, we gathered 20 pharmaceutical keywords, including a number of the most-prescribed and best-selling product terms from IMS Health [14]. Note that to restrict the search results to each cloud platform, we included the query "site:cloud service's domain name" (e.g., site:s3.amazonaws.com) before the aforementioned keyword phrases.

Directory Dataset Collection. On the cloud hosting service, the web pages are organized as directories. For example, a typical URL of a web page in cloud hosting service is as follows:

$$\text{scheme} : //\text{dir\_name.domain/file\_name}$$

where *scheme* is the protocol, e.g., HTTPS; the *dir_name* is the name of the directory shown as sub-domain; and the *file_name* is the path of the file in the cloud directory which is customized by the user. All pages from the same directory have the same *dir_name* component.

As the pages are organized as a directory in the cloud hosting service, the crawler further explores the web pages in the cloud directories which house the pages in a seed dataset D^s. Specifically, we extract the directory names from cloud URLs in the seed dataset, and then conduct another search engine query to restrict the search results to each cloud directory. Specifically, we use the keyword "site: dir_name.domain" (e.g., site:abc.s3.amazonaws.com) for the search engine query.

In this way, we generated an expanded dataset D^d, which contains 1,073,642 URLs. Ideally, the expanded dataset D^d should include all the cloud directories in the seed dataset D^s. However, as cloud platforms took action to delete the doorway pages during the course of our study, we found that 4,554 cloud directories expired. Table 2 shows the summary of the collected data.

To analyze the behavior of these cloud pages, we ran a dynamic crawler (as a Firefox add-on) to visit each cloud web page with the Referrer as google.com, and recorded the web activities it triggered, including network request, response, and browser events. For this purpose, we deployed 20 dynamic crawlers, which were hosted on Redhat Virtual Machines (VM) with distinct IP addresses.

3.2 Abusive Cloud Directory Classification

Automated spam page identification on large-scale web pages is an open research question and there are no clear rules for absolute positive identification [5][23]. From the quality guidelines from Google [8], the four categories that indicate spam pages are as follows: (a) Pages generated by an automated tool or automated processes, such as Markov chains. (b) Pages optimized for a specific keyword or phrase, that then funneled users to a single destination. (c) Pages with product affiliate links on which the product descriptions and reviews are copied directly from the original merchant, without any original content or added value. (d) Pages dedicated to embedding content such as video, images, or other media from other sites without substantial added value to the visitor.

A set of heuristics were used to developed a classifier, and to detect the cloud directories used for long-tail SEO. (1) The web pages in the abusive cloud directories were optimized for a series of similar long-tail keywords. This is because to promote a targeted content, the long-tail SEO web pages utilize several long-tail keywords generated for a specific content. For example, to promote the web pages for "green coffee bean", the corresponding long-tail keywords could be "green coffee bean capsules australia", "green coffee bean capsules uk" and "green coffee bean amazon uk". (2) The web pages in the abusive cloud directories show high similarity in content and sometimes funnel visitors to the same destination websites. This is because the abusive long-tail SEO web pages are typically generated from automatic tools with a limited number of templates, and thus the web pages in the cloud directories are very similar in their DOM (i.e., document object model) structure.

Our classification began by labeling the abusive cloud directories and non-abusive directories for training. To label the cloud directories for long-tail SEO spam, we sorted the cloud directories by the number of files in the directories and manually examined the web pages. In this way, we identified 100 abusive cloud directories (10 directories on 10 cloud platforms) meeting the aforementioned definition of long-tail SEO spam. To label the non-abusive directories, we extracted the second-level domains of the URLs embedded in the cloud web pages and sorted them by their frequency of appearance. We manually examined the pages and their corresponding cloud directories with the bottom 500 second-level domains from different directories to label the non-abusive directories. Also, for those pages without an embedded URL or JavaScript, we checked if their corresponding cloud directories were non-abusive. In this way, we label 100 non-abusive cloud directories.

We extracted features from the labeled dataset in an automated fashion. Specifically, we used two sources of inputs for features: the directory features and the web pages in the directories. For the cloud directory features, we observe that the file names in the abusive cloud directories show greater similarity. This is because keywords in URLs can increase the clickthrough rate in the search engine result pages [16], and the abusive user tends to make the long-tail keywords visible in the URLs. Hence, highly similar long-tail keywords in URLs show as similar file names in the abusive cloud directories. To calculate the file names' cosine-similarity, we extract the file names from the path component of the cloud URL, and then tokenize them into words using separators such as \`-′ and \`_′. Then, the words in each file name is converted into a sparse vector, and we calculate cosine-similarity for the vectors in the same cloud directories.

For the web page in the directories, the main reason we extract features from the raw HTML is that long-tail doorway pages in the abusive cloud directories shows high similarity in page content, such as meta keywords, page title and page template because of automatic page generation. To extract HTML source features, we follow a conventional n-gram approach. Particularly, we choose to build 3-gram features. The rationale is that a 3-gram can capture the structure for a sequence (e.g., affid=12345) very well. Each Meta keyword, URL and script in the web page is segmented into words so that each word is either one of the reserved characters in \`! * ′ ( ) ; : @ & = + \$, / ? \% # [ ] ′ , or contains no reserved characters. We convert each word into a sparse vector with the dimensions of the same number of 3-grams. On each dimension, the value is proportional to the frequency of the corresponding n-gram. Each vector is normalized to have the L_1 norm [26].

Subsequently, we trained a SVM (i.e., support vector machine) [26] classifier over the training set. We evaluated the predictive accuracy of the classifier by performing 10-fold cross-validation on the labeled dataset, yielding a 92% rate of successful classification. In the end, the algorithm classified 3,186 abusive cloud directories. To validate these predictions, we manually inspected additional subsets of unlabeled examples. Without loss of generality, we utilize Chernoff Bounds [22] to estimate the number of pages to be sampled. We set the trust interval $\delta = 0.01$ and the error probability $\lambda = 0.01$ to obtain the number of sampled cloud directories $n = 500$. After manually inspecting the sampled cloud directories, we find that around 12 of the cloud directories are false positives which is consistent with the predicted 92% rate.

3.3 Ethical concerns

In order to avoid unintentionally advertising for abusive actors, we do not include the actual names of abusive cloud directories and vendors. Instead of including the raw URL of spam directories and doorway pages, we adopt the naming convention of <cloud provider>_<affiliate program+ number> to minimize the impact on privacy. When including content from these doorway pages we redact all raw identifiers, such as URLs, identifying comments and other potentially identifying information. Also, we limit our analysis to public URLs that are indexed by a search engine for identification and measurement. We did not try to access the base directory listings in order to minimize the impact on privacy.

4. LONG-TAIL SEO ON THE CLOUD

In this section, we study the effectiveness of long-tail SEO spam on cloud web hosting services, i.e., the prevalence of long-tail SEO spam on cloud web hosting as well as their impact on organic long-tail keywords search results. We found that 6% of the long-tail SEO doorway pages we observed successfully poisoned the top-10 search results for long-tail

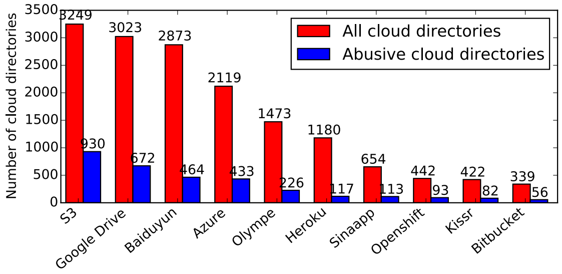


Figure 3: Number of abusive cloud directories on each cloud platform.

keywords included in our study. Then, we provide a perspective of the blackhat SEO techniques and the evasion techniques the abusive user adapted for the cloud web hosting platforms.

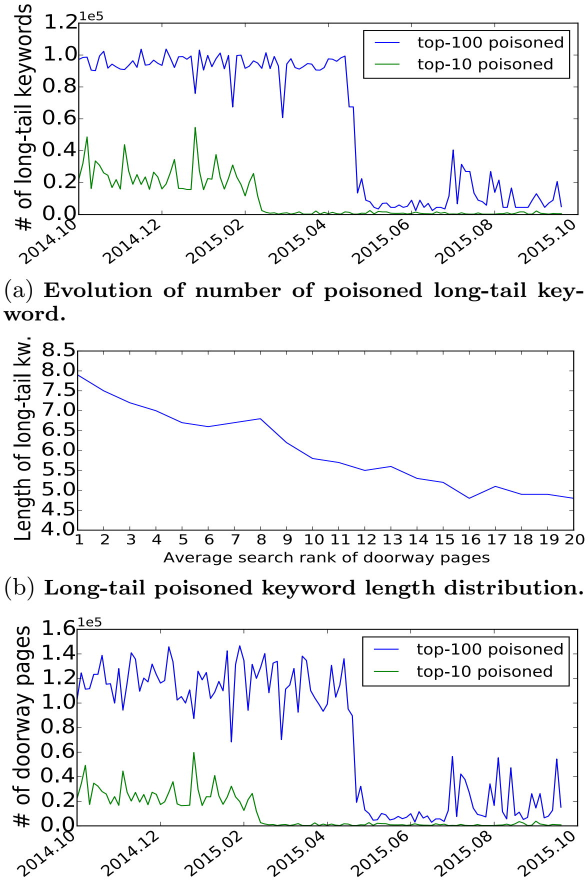
Overview. We start by discussing the prevalence of abusive cloud directories for long-tail SEO spam on cloud web hosting platforms. Of the 15,774 cloud directories we collected, we found that 3,186 directories (318,470 doorway pages) were long-tail SEO spam.

Figure 3 illustrates the number of abusive cloud directories on each cloud platforms. Among them, Amazon S3 is the most popular (28%) in our dataset, followed by Google Drive (22%). The result shows that the abusive cloud directories for long-tail SEO is being hosted on cloud platforms. Note that of these 10 cloud platforms, eight of them provide free hosting services (e.g., 5GB for Amazon S3, 15GB for Google Drive), and therefore are ideal platforms for low-budget abusive users. These users also take advantage of the pay-as-you-go feature of cloud hosting to conduct low cost long-tail SEO, which does not require traditional SEO back linking techniques [17][18]. Lastly, long-tail SEO pages hosted on the cloud are more difficult to blacklist since cloud hosting domains also host a large amount of benign content.

Effectiveness of Long-tail SEO. To analyze the search engine poisoning impact of long-tail SEO spam on the cloud, we extracted 236,368 distinct long-tail keywords from doorway pages in the abusive cloud directories we identify, and then crawled the top 100 organic Google search results of the long-tail keywords from 10/2014 to 10/2015.

To extract the keywords, we implemented a stuffed keyword extraction tool based on n-grams. We define an N-gram as a contiguous sequence of n words in the HTML files. First, we extract the text from the DOM tree using an open-source tool `BeautifulSoup` and use white space as the token separator. Then, we calculate the frequency of each n-gram. In our implementation, we set the range of n from 3 to the length of page title l. After that, we compared the n-gram tokens' frequencies f where $n \in [3, l]$ and used the n-gram token with the largest keyword density $d = \frac{n \times f}{T}$ as the stuffed keywords, where n is the length of the keyword token, f is its frequency and T is the number of words in a page.

For example, the owner of the abusive cloud directories on Google Drive uploaded a keyword stuffing doorway page '1403682103503-aclarar-la-piel-para-siempre—oficial.html' with 775 word phrases. The page has the largest 3-gram token 'la piel para' with frequency 47, largest 4-gram token 'la piel para siempre' with frequency 42, largest 5-gram token 'aclarar la piel para siempre' with frequency 35 and largest 6-gram token 'aclarar la piel para siempre oficial' with frequency 12. The stuffed keyword extraction tool will extract the long-tail keyword 'aclarar la piel para siempre' with the largest percentage 22.5%.



(a) **Evolution of number of poisoned long-tail keyword.**

(b) **Long-tail poisoned keyword length distribution.**

(c) **Evolution of number of doorway pages.**

Figure 4: Effectiveness of Long-tail SEO spam.

Surprisingly, we found that the doorway pages in the abusive cloud directories successfully poisoned the highly specific long-tail keyword phrases. Figure 4(a) illustrates the evolution of the number of poisoned long-tail keywords over time. We define the long-tail keywords as poisoned if the abusive cloud directories appeared in the top 10 (i.e., indicated as top-10 poisoned in figures) or top 100 (i.e., indicated as top-100 poisoned in figures) organic search results. During the period from 10/2014.10 to 2/2015, 9% of the long-tail keywords were poisoned in the top 10 organic search results. This number jumped to 42% for top 100. In general, the trend exhibits a substantial decrease in the number of poisoned keywords, because cloud providers will remove the doorway pages. We also observe that non-English keywords were easier to be poisoned, such as 'como pintar con oleo', which has the relevant doorway page ranked as the first search result.

Figure 4(b) illustrates the average length of the poisoned long-tail keywords in search rank from 1 to 20. Overall, the average length of poisoned keywords increases while the search ranks of doorway pages become higher. This is because the shorter keywords have higher competition and is therefore difficult to be polluted. The average length of the keywords, whose corresponding doorway page's poisoned search rank is 1, is around eight. However, when the keyword length is 6, the average search rank of doorway pages decreases to 10.

Figure 4(c) shows the evolution of the number of doorway pages we found in the top 10 and top 100 organic search results for the poisoned long-tail keywords. On average, 6% of the doorway pages are ranked in the top 10, which is 32% in top 100. From Figure 4(c), we can see that the prevalence of the doorway pages in the organic search results. For an example of SEO effectiveness, 100 doorway pages in the abusive cloud directories googledrive_markethealth successfully poisoned 61 long-tail keywords' top 100 search results, which will redirect the visitors to the same online pharmacy vendor, a site that was reported as a scam website by reviewopedia [27]. Among the 61 poisoned keywords, the doorway pages appeared in 5 long-tail keywords' top 5 search results. Examples of the poisoned long-tail keywords include 'green coffee bean diet does it work' and 'green coffee bean cleanse australia'.

Blackhat SEO technique. We examined the blackhat SEO technique that the spam campaigns utilized to poison search results. Our research surprisingly revealed that using simple blackhat SEO technique (e.g., keyword stuffing), doorway pages were able to successfully poison the search results. In addition to blackhat SEO techniques, such as keyword stuffing and social fraud, targeted blackhat SEO techniques were also used, incorporating multiple cloud provider related elements such as adding products as unrelated keywords or misleading visitors by adding the cloud provider's logo.

Keyword poisoning is the deliberate manipulation of the search engine's index for specific keyword terms. It involves a number of methods such as keyword stuffing (i.e., the repetition of keywords in the meta tag and page contents), and traffic spam (i.e., adding unrelated keywords to manipulate the relevance).

Regarding the doorway pages' keyword densities that we obtain from Section 4, 84% of doorway pages have a keyword density larger than 15%, which is less than 3% for web pages in non-abusive cloud directories that we mention in Section 3. As an example of keyword stuffing, in the doorway pages uploaded in the abusive cloud directory googledrive_clickbank, keywords were repeated multiple times in the content of the pages. To hide the stuffed keywords from human readers, abusive users set white text on a white background or located the stuffed keywords behind figures in the doorway pages.

To measure the keywords relevance to identify traffic spam, we studied the doorway pages with more than one META keywords. We extract the keywords from the META tag of the doorway pages and query their semantic similarity using DISCO API. If the keywords have a large semantic gap (semantic similarity<0.05), we determine that the doorway page utilizes traffic spam techniques. Using this method we find that 48,922 doorway pages in 526 abusive cloud directories utilize traffic spam techniques to manipulate the page relevance. Interestingly, the abusive users include cloud platform-related information as the stuffed keywords or unrelated keywords, such as Google Plus and Youtube. For example, the doorway pages that masquerade as an online flower shop utilize "Proflowers Google Plus" or "lotus flower youtube" as the keyword phrases. We observed that 16% of the doorway pages on Google Drive use Google product terms as unrelated keywords, which is 5% on Amazon S3.

```
<script type="text/javascript">
  if (document.referrer != ""){
    var refer_url = document.referrer;
    var post_url = "http://www.gatherguideshare.info/product/" + data_loc
          ;
    document.write('<form name="form1" method="post" action="' + post_url
           + '"><input name="info" type="hidden" value="' + data_info +
          '" /><input name="refer" type="hidden" value="' + refer_url +
          '" /></form>');
    document.form1.submit();
  }
</script>}
```

Figure 5: Redirection cloaking used by amazon_gatguisha-20

Evasion technique. Given the prevalence of the doorway pages on cloud hosting platforms, we examine the evasion techniques used to avoid detection. We found that the spam campaigns adapted evasion techniques for cloud web hosting platform, such as link shorteners and obfuscated client-side JavaScript when cloud platforms do not support server-side scripting.

As the illicit practices of doorway pages and manipulating search rankings can lead to the pages being removed from the Google index [8], the attackers utilize evasion mechanisms to avoid detection. However, as most of the cloud hosting platforms (e.g., Google Drive, Amazon S3) do not support server-side scripting and a simple client-side script for evasion is easily detected, abusive users make several changes to adapt to the cloud web hosting platform. Many evasion mechanisms were used by the abusive users, such as obfuscation, link shortening and redirection cloaking.

1) Mixed redirect cloaking. Cloaking refers to deceiving search engines by providing different content to the search engine crawlers compared to users clicking on search results. Cloaking on the traditional platform includes client-side cloaking (e.g., use client-side scripting to store cookies) and server-side cloaking (e.g., using server-side scripting to track IP). Compared to client-side cloaking, server-side cloaking is more concealed and much more likely to circumvent detection [32]. As most of the cloud hosting platforms do not support server-side scripting, we observed mixed redirection cloaking, which combined the client-side cloaking on doorway pages and server-side cloaking on the external server. The abusive user utilized mixed redirection cloaking as shown in Figure 5. When the user visits the webpage, a POST request is dynamically generated by the Javascript implementation to report the document.referrer to the external server gatherguideshare.info. The external sever then operates the server-side redirection cloaking based on the document.referrer, i.e., the external sever will respond with status code 302 for the POST request to redirect normal visitors (e.g., those who visit doorway pages by clicking through search engine results) to gatherguideshare.info, while search engine crawlers receive content crafted to rank well for targeted query terms (e.g., "compaq armada dock station").

2) Obfuscation. Obfuscation is the deliberate act of creating code that is difficult for humans to understand. Obfuscation, as another way to circumvent static analysis of

Table 3: Top 5 affiliate networks where most abusive cloud directories belong to.

Affiliate network	# of doorway pages	# of cloud directories	Volume
Amazon	7,663	72	2.4%
viglink	6,272	64	2.0%
prosperent	5,077	52	1.6%
Clickbank	4,689	51	1.4%
MarketHealth	4,177	43	1.3%

Table 4: Example of regular expressions for campaign ID identification.

Campaign ID Regex
`prosperent.com/store/product/[0-9]{6}-[0-9]{4}-[0-9]`
`redirectingat.com?id=[0-9]{5}X[0-9]{7}`
`\w.\w.hop.clickbank.net`
`247rxshop.com/?affid=[0-9]{8}`
`paydaylendersearch.com/[a-z0-9]{8}-[a-z0-9]{4}-[a-z0-9]{4}-[a-z0-9]{4}-[a-z0-9]{12}`

the client-side illicit script, is also widely used in the doorway pages on cloud platforms. For example, the redirection cloaking code we mentioned in Figure 5 was obfuscated by the character code. Note that by combining the cloaking and obfuscation techniques, the doorway pages from `s3.amazonaws.com_gatguisha-20` have a longer lifetime (more than 15 weeks we observed) than other doorway pages on the same cloud platform (average 7 weeks, detailed in Section 6). Other forms of obfuscation were also found, such as word substitution, which separates key phrases (e.g., campaign ID) into fragments with random order.

URL shortening is another evasion technique used by the abusive users to circumvent static analysis from the cloud service provider. For the doorway pages in cloud directory `googledrive_filepost`, a shortened URL was generated dynamically by the Javascript code, which requests bit.ly URL shorten API `https://api-ssl.bitly.com/v3/shorten` for each doorway pages. In the Javascript implementation, a long URL was first generated with the parameter in "asin" tag of each page, and the fixed domain and path. Then, a bitly URL shortener API was called to return the shortened URL for the original one. Note that the fixed domain and path were also obfuscated by the BASE64 code and the shortened URL was generated at run time.

In addition to `bit.ly`, multiple URL shorteners are utilized by the abusive users such as `t.co` (0.6% of doorway pages), `goo.gl` (1.5% of doorway pages) and `tinyurl.com` (5% of doorway pages).

5. TRAFFIC MONETIZATION

In this section, we study how the long-tail SEO campaigns monetize traffic. We find that almost all of the long-tail spam campaigns are monetized by sending visitors to affiliate programs. Further, we identify five large long-tail spam campaigns and surprisingly find that they are mainly working for reputable affiliate networks (e.g., prosperent.com) or for reputable online vendors' promotion (e.g., Amazon). Traffic monetization techniques used by the long-tail spam campaigns were analyzed, followed by a revenue analysis.

Affiliate structure of long-tail spam campaigns. Different from traditional platforms, doorway pages on cloud platforms share a common second-level domain and trusted name servers belonging to cloud platforms. Thus, to look at the network topology, we built network topology graphs G_{ta} for the abusive cloud directories. In the graphs, each IP of the redirectors and landing servers is regarded as a node, and the abusive cloud directories are set as the starting nodes. Each edge corresponds to a redirection between two nodes.

We manually review the network topology G_{ta} of abusive cloud directories and found that many of them were organized as affiliate programs. For the graph G_{ta} with 6,012 nodes and 47,398 edges, we surprisingly only found that *the long-tail SEO campaigns on the cloud web hosting platform show great connectivity in network topology.* As hubs in the graph G_{ta}, the top 3 nodes with the largest in-degree in G_{ta} are `amazon.com`, `prosperent.com` and `viglink.com`. The top three nodes with the largest out-degree in G_{ta} are `clickbank.net`, `redirectingat.com` and `dotomi.com`. By reverse DNS lookup, we found that *each of them belongs to reputable affiliate networks.*

Next, we utilize the hubs in the graph to give an overview of the affiliate structure of the abusive cloud directories. Table 3 provides an overview of the top 5 affiliate networks that host the most abusive cloud directories. Overall, the majority of the abusive cloud directories come from abusive users that work for reputable affiliate networks. These affiliate networks mostly collaborate with well-known online vendors that have a policy, based on abuse reports from individuals, to prohibit their affiliates from using the service in conjunction with network abuse or spam. We observe that the largest amount (2.4%) of doorway pages belong to 72 abusive cloud directories working for the reputable affiliate network `Amazon`. Thus, it appears that even though most of the reputable affiliate networks have policies that govern the affiliates to prevent abuse and search engine attacks, illicit practices are still found in these reputable affiliate networks.

Identifying the affiliate ID of the abusive users would help to identify spam campaigns, prevent their spread and efficiently remove the doorway pages. Our idea is to extract the affiliate IDs from the redirection chains of the doorway pages. To do so, we designed a semi-automatic common substring-based algorithm to generate regular expressions to extract affiliate IDs, as follows: (1) We extract a common string from each directory by the cross-comparison between the redirection chain inner pages and the redirection chains among the pages in the directory using a generalized suffix tree [11]. Note that we consider each order of the parameters in the URL query string, i.e., for both cases `example.com/?a=1&b=2` and `example.com/?b=2&a=1`, we calculate their common strings. (2) We generate the regular expressions by mapping the digits and English alphabets in the URL parameter into formal language. (3) We manually check the correctness of these regular expressions such as accessing the affiliate network for marketing URL information and manually inspecting the sampled pages. In this way, we labeled 2,360 abusive cloud directories with 342 affiliate IDs (a.k.a., abusive entities), which were associated with 225,008 long-tail SEO doorway pages. We present the cumulative distribution of the number of cloud directories per abusive entities in Figure 6. Figure 6 shows that 80% of the abusive entities are associated with more than one cloud directory. Moreover, 14% of abusive entities distribute doorway pages on different cloud platforms. This might be because the abusive entities are concerned about being de-

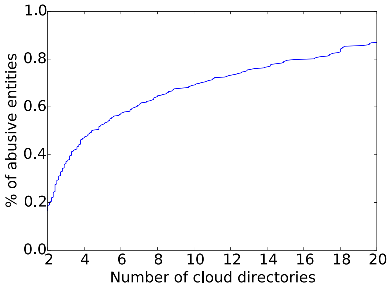


Figure 6: Cumulative distribution of number of cloud directories per abusive entities.

tected by cloud platforms, and so they distribute doorway pages into different directories, and cloud platforms.

Dissecting traffic monetization techniques. Long-tail spam campaigns monetize traffic through multiple vectors, such as search redirection and social fraud.

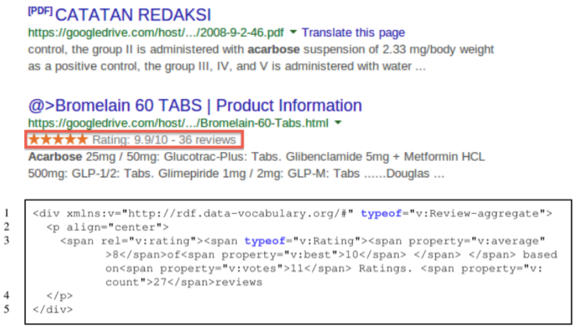
Search redirection. The search redirection technique has been regarded as the blackhat technique to lure traffic. When visitors click the link in the search engine result pages, they will be redirected to a different site rather than the one pointed to by the link. Traditional search redirection attackers prompt a site server to conduct request redirection via code injection. On the cloud platform, the abusive users utilize client-side script to redirect visitors, such as iframe, JavaScript (e.g., windows.location) and POST request.

We determine the search redirection with the dynamic crawler (see Section 3), and further analyze the semantics consistency of the source page and the landing page. Specifically, we utilized Yahoo content analysis API [34] to extract a series of keywords from the source page and the landing page. If the keyword sets did not intersect, the search redirection shows semantic inconsistency.

In this way, we find that 63,900 doorway pages in 769 abusive cloud directories utilized search redirection to monetize traffic. Among them, 23% of the doorway pages redirected visitors to a semantic inconsistency landing page. For example, 280 doorway pages were uploaded in the abusive cloud directory `googledrive.com_mediaupdate41` for malware distribution. The doorway pages masquerade as web pages that sell flowers to funnel visitors to a malware distribution website.

Social fraud. Social fraud techniques mislead Google users by manipulating the search snippets. Google's Rich Snippets technique [7] allows users to summarize the content of a page such as a product's review. Rich Snippets help visitors recognize the relevancy of their search and trigger potential clicking. However, Rich Snippets can be directly inserted in the page without validation. Abusive users can thus leverage this technique to provide fake review scores or irrelevant reviews. For example, the doorway page in the abusive cloud directories `s3.amazonaws.com_markethealth` makes up irrelevant reviews using rich snippets which shows in the search result to attract clicks. Figure 7 shows Rich Snippets that have been abused.

Revenue Analysis. To understand the economic motives behind the abusive activities, we analyzed the revenue received by these users. We utilize the following revenue model which was proposed in prior research[1][23]: $R(t) = N_v(t) \cdot P_a \cdot R_a$, where the total revenue $R(t)$ during the time period t is calculated from the total number of actions taken



[PDF] CATATAN REDAKSI
https://googledrive.com/host/.../2008-9-2-46.pdf ▾ Translate this page
control, the group II is administered with **acarbose** suspension of 2.33 mg/body weight
as a positive control, the group III, IV, and V is administered with water ...

@>Bromelain 60 TABS | Product Information
https://googledrive.com/host/.../Bromelain-60-Tabs.html ▾
★★★★★ Rating: 9.9/10 - 36 reviews
Acarbose 25mg / 50mg: Glucotrac-Plus: Tabs. Glibenclamide 5mg + Metformin HCL
500mg: GLP-1/2: Tabs. Glimepiride 1mg / 2mg: GLP-M: Tabs ......Douglas ...

```
<div xmlns:v="http://rdf.data-vocabulary.org/#" typeof="v:Review-aggregate">
  <p align="center">
    <span rel="v:rating"><span typeof="v:Rating"><span property="v:average"
          >8</span>of<span property="v:best">10</span> </span> </span> based
          on<span property="v:votes">11</span> Ratings. <span property="v:
          count">27</span>reviews
  </p>
</div>
```

Figure 7: Fake product score shown in the search result page.

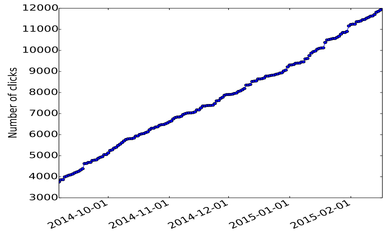


Figure 8: The cumulative number of clicks of the doorway pages on the abusive cloud directory `googledrive.com_filepost`.

(i.e., click-through number [23], $N_v(t) \cdot P_a$) and the average revenue per action R_a.

To investigate the increase rate of the click-through number (i.e., $N_v(t) \cdot P_a$), we track the number of URL clicks for the doorway pages uploaded by the abusive user who works for a shady affiliate network called *filepost.ml*. The affiliate network *filepost.ml* hires affiliates to promote its fake free e-book download website which lures visitors to finish many cost-per-action affiliate programs. The abusive user hosts 384 doorway pages in one cloud directory and hides the marketing URL by using the URL shortener `bitly.com`. As `Bitly` provides an API to count the number of clicks for its shortened URL, we obtain the click number of the marketing URL in the abusive doorway pages.

Figure 8 shows the cumulative click-through number of the 384 doorway pages from Sep. 5, 2014 to Feb. 18, 2015 . Hosting the 384 doorway pages on Google Drive, the abusive entity will see around a 1,800 click increase every month. The click increase rate is around 20% per month. Utilizing the same revenue model and parameter setting $R_a = \$0.265$ as prior works [1][23], we can estimate the revenue for `googledrive.com_filepost` in October 2014 of $R(1\ month) = (5249 - 3754) \times 0.265 = \$396$, which increased to $665 in January 2015. Note that with the evasion technique we mentioned in Section 4, abusive cloud directories have extremely long lifetimes (i.e., more than 40 weeks), which helps the abusive users gain more profit.

6. INTERVENTION

In this section, we monitored ongoing interventions by the cloud service providers. Since the abusive users violate the usage policies of cloud platforms [8] and poison search engine results to degrade users' experience, cloud providers tend to remove doorway pages entirely from the cloud plat-

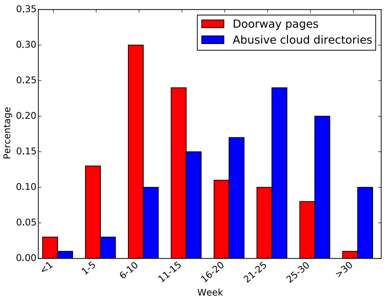


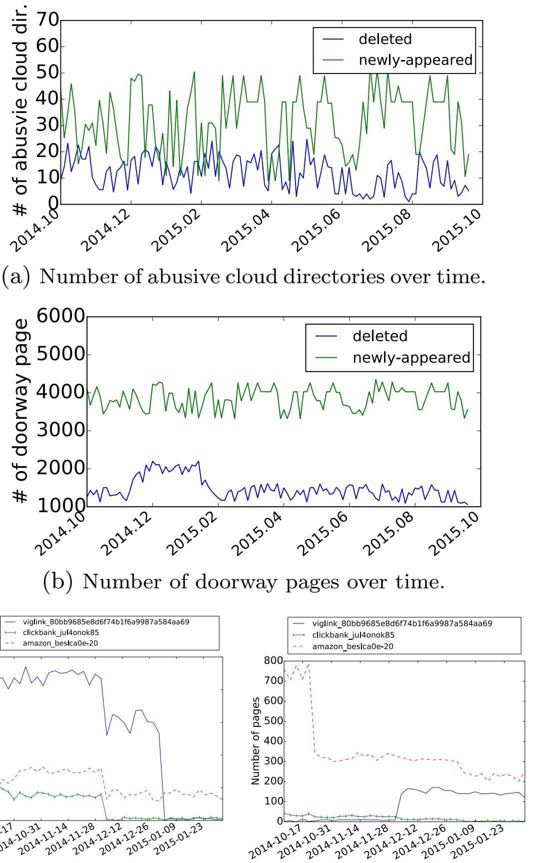
Figure 9: Lifetime of doorway pages and abusive cloud directories.

forms. However, as our empirical analysis shows, these cloud providers' efforts are far from effective.

To measure the average lifetime of the doorway pages and the abusive cloud directories, we re-crawled the active doorway pages every three days and used the lifetime as the time between the first and last time the crawler observed a page. Figure 9 illustrates the percentage of doorway pages and abusive cloud directories in different lifetime ranges. We found that the average lifetime of the doorway pages is around seven weeks, which is much longer than those hosted on the compromised sites (i.e., around one week [10]). Moreover, the average lifetime of the abusive cloud directories was extremely long (around 20 weeks). In the empirical study, we observed that though the cloud providers found and removed the doorway pages, they did not aggressively remove the corresponding abusive directories, or the doorway pages from the same abusive entities in different cloud directories.

Then, we analyzed the abuse situation of doorway pages in the cloud web hosting platform, i.e., the evolution of the newly-appeared doorway pages and abusive cloud directories. Hence, we resubmitted the hot and shady keywords to the search engine every three days from 2014.10 to 2015.10. At each measurement point, we crawled the data in the same way as mentioned in Section 3. In this way, we built the time-period dataset D^t. At each measurement point, the average number of URLs we crawled was around 500K associated with 3K cloud directories.

Figure 10(a) shows the evolution of the number of newly-appearing abusive doorway pages, compared with the number of deleted doorway pages we found in Section 3. The evolution of the abusive cloud directories is shown in Figure 10(b). From 10(a), we can observe that large amounts of doorway pages newly appear, which has a higher rate of increase than deletion rate by the cloud provider. Also, 23% of the newly-appeared doorway pages were associated with the known abusive directories, and the rest of them belonged to the newly-appeared abusive directories. Also, from Figure 10(b), we observed that the deletion rate of the abusive cloud directories is much smaller than that of doorway pages. This shows that the detection method used by cloud platform did not identify a large enough amount of doorway pages for each abusive cloud directory or remove the abusive cloud directories. Moreover, we observe an increased deletion rate from 2014.12 to 2015.02, because the cloud provider *Google Drive* took more efficient action to remove doorway pages. As the doorway pages can be easily spread on the cloud web hosting platform, the abuse situation will become worse if the detection method is not effective enough.



(a) Number of abusive cloud directories over time.

(b) Number of doorway pages over time.

(c) Number of active doorway pages in Top 10/100 search ranking per measurement point.

(d) Number of deleted doorway pages in Top 10/100 search ranking per measurement point.

Figure 10: The increasing trends of active pages, deleted pages and doorway pages over time are shown in Figure 10(a). The corresponding accounts are shown in Figure 10(b). Number of doorway pages for the three campaigns over time are shown in Figure 10(c) and 10(d).

Figure 10(c) shows the prevalence of doorway pages for three abusive cloud campaigns. The trend line shows the number of doorway pages in the 1,520 'hot' keywords top 10 search results restricted to their cloud platform. For the three abusive campaigns, the number of doorway pages appeared in the top 10 search results did not change much in October and November. For the abusive user `amazon_beslca0e-20`, more doorway pages poisoned the top 10 search results in November. Figure 10(d) shows the number of doorway pages that appeared in the deleted page set over time. Even though the deleted doorway pages will be removed from the top 10 search results, the active doorway pages from the same abusive cloud directories stay at the same measurement point, which means that cloud provider did not detect and remove all the doorway pages from the same campaign.

7. DISCUSSION

In this section, we discuss the limitations of our study and potential mitigation strategies.

Limitation. As mentioned earlier, long-tail SEO spam identification on large-scale cloud data is difficult, especially for a third party. Our design has a number of limitations imposed by our vantage point, over-restricted features and manual validation. First, our methodologies' vantage points are limited to Google's search results. While Google is the mainstream search engine targeted for search engine poisoning, the results we crawled were limited to the cloud web pages that are indexed by Google. Second, the insight for feature extraction is that the abusive user tends to put several doorway pages in cloud directories for long-tail SEO spam, and the doorway pages are auto-generated and hence show similarity. While this insight was validated by our pre-measurement study on training data, there may be small numbers of abusive users intentionally increasing the keyword and DOM source diversity to evade detection. Hence, the over-restricted feature design may bias our technique to low false positives but relatively smaller coverage. Third, we use manual inspection to validate abusive cloud directories, which is laborious and may include false positive.

Mitigation strategies. Based on the results of our measurement study, we have identified several potentially effective mitigation strategies to reduce the impact of long-tail SEO spam on the cloud hosting platforms. First, search engines, could detect the highly similar low-quality content of these doorway pages and penalize them in the search rankings. This would cause these spammers to expend more resources creating less duplicated, higher quality content. Second, the cloud providers could increase the cost of establishing accounts on their services and more aggressively detect and remove spammy cloud hosting accounts. While this can result in an escalating detection and evasion arms-race, our analysis of these doorway pages found that identifying affiliate IDs can be done automatically and these can be used to detect and remove large numbers of accounts hosting doorway pages. Finally, the affiliate networks could monitor HTTP refers and identify other indications that their affiliates are engaging in SEO spam. We found that most of the affiliate networks currently have reactive policies, such as abuse reporting to restrict illicit practices of affiliates. A more proactive policy might help to mitigate the surge of long-tail SEO spam on cloud hosting platforms.

8. RELATED WORK

In this section, we discuss the related prior studies and their relationships with our work.

Cloud Security. Previous cloud security studies primarily focused on confidentiality of data or attacks targeting the cloud computing infrastructure. Ristenpart *et al.* introduced the vulnerability of information leakage by sharing physical resources with co-resident malicious virtual machines in the Amazon EC2 service [28]. Zhang *et al.* discovered a new cache-based side-channel attack to collect sensitive data or hijack user accounts on commercial clouds [35]. Recently, researchers have paid attention to the fraudulent use of cloud-based services. Mulazzani *et al.* demonstrated that Dropbox can be exploited to hide files in the cloud and serve as a covert channel for attackers [24]. Han *et al.* conducted a measurement study of malicious and dedicated cloud-based domains used in malicious infrastructure [13]. Unlike these works, we analyze *long-tail SEO spam on cloud web hosting services* which promote illicit sites and have a negative impact on end users, including those not using these cloud services.

Affiliate Program. Recent studies have investigated spam affiliates that send spam through their own email delivery infrastructure and receive a cut of the final revenue for every purchase they bring to the spam-advertised sites [3][19][29]. McCoy *et al.* analyzed customer demand and overhead in the spam cost model by using transaction logs of pharmaceutical affiliate programs [21]. Caballero *et al.* infiltrated malware distribution affiliates and measured the pay-per-install market [2]. We supplement prior research by *characterizing the affiliates and affiliate networks abusing cloud web hosting services.*

Search poisoning. Miscreants use search poisoning attacks to falsely increase the rank for their web sites. Previous studies have examined the lexical patterns of the page content [25][30], the hyper-link structure from site to site [12][33], or the combination of the aforementioned features as well as network-level features [31]. deSEO used URL signatures to identify malicious SEO campaigns of fake pages hosted on compromised web servers [15]. Leontiadis *et al.* did an in-depth analysis of search poisoning attacks which redirected traffic to online pharmacies and found that the conversion rate was higher than email spam [17]. They further used data collected over four years to investigate the evolution of search engine poisoning, which showed that search poisoning attacks have steadily grown. A potential bottleneck is the relatively small set of traffic redirectors was highlighted by Leontiadis *et al.* [18]. In this paper, we focus on long-tail search-result manipulation based on cloud-hosted pages.

9. CONCLUSION

To the best of our knowledge, a comprehensive overview of the long-tail SEO spam on the cloud web hosting platform, and measurement study of the abusive activities are still open research challenges. In this paper, we conduct the first study to measure, and analyze the long-tail SEO spam on cloud web hosting platforms. Specifically, we identified 3,186 abusive cloud directories for long-tail SEO spam from analyzing approximately 15,774 cloud directories over 10 cloud platforms. Then, we conducted an in-depth measurement study of the abusive cloud directories for long-tail SEO spam. As a result of our measurement, we uncover that the abusive users take advantage of the pay-as-you-go feature of cloud hosting to conduct low cost long-tail SEO. Our measurement study provides insights into long-tail SEO spam effectiveness, blackhat SEO techniques they used, and network characteristics of the long-tail SEO campaigns. Moreover, the intervention of the cloud provider is analyzed, which is shown to be far from effective. Our findings for the long-tail SEO spam on cloud hosting platforms enable us to deeply understand the abusive long-tail SEO spam, which enable an important step toward effective mitigating of this new type of security threat.

10. ACKNOWLEDGMENT

This work was supported by the National Science Foundation (grants CNS-1619620, CNS-1314857, CNS-1453634, CNS-1518765, CNS-1514261); a Packard Fellowship; a Sloan Fellowship; two Google Faculty Research Awards and a VMWare Research Award. We thank our anonymous reviewers for their useful comments.

11. REFERENCES

[1] Sumayah Alrwais, Kan Yuan, Eihal Alowaisheq, Zhou Li, and X Wang. Understanding the dark side of domain parking. In *Proceedings of the 23rd USENIX Security Symposium*, 2014.

[2] Juan Caballero, Chris Grier, Christian Kreibich, and Vern Paxson. Measuring Pay-per-Install: The Commoditization of Malware Distribution. In *Proc. 20th USENIX Security Symposium*, San Francisco, CA, August 2011.

[3] Chris Kanich and Christian Kreibich and Kirill Levchenko and Brandon Enright and Vern Paxson and Geoffrey M. Voelker and Stefan Savage,. Spamalytics: an Empirical Analysis of Spam Marketing Conversion. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, Arlington, VA, October 2008.

[4] Comm100. Spammy words. http://emailmarketing.comm100.com/email-marketing-ebook/spam-words.aspx, 2015. [Online].

[5] Matthew F Der, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. Knock it off: profiling the online storefronts of counterfeit merchandise. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1759–1768. ACM, 2014.

[6] Google. Google Trend. http://www.google.com/trends/hottrends, 2014. [Online].

[7] Google. Rich snippets guidelines. https://support.google.com/webmasters/answer/2722261?hl=en, 2014. [Online].

[8] Google. Webmaster Guidelines. https://support.google.com/webmasters/answer/35769?hl=en&ref_topic=6002025, 2014. [Online].

[9] Google. Publish website content. https://developers.google.com/drive/web/publish-site, 2015. [Online].

[10] Chris Grier, Lucas Ballard, Juan Caballero, Neha Chachra, Christian J Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Damon McCoy, Antonio Nappa, Andreas Pitsillidis, et al. Manufacturing compromise: the emergence of exploit-as-a-service. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 821–832. ACM, 2012.

[11] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[12] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating Web Spam with TrustRank. In *Proc. 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, Canada, September 2004.

[13] Xiao Han, Nizar Kheir, and Davide Balzarotti. The role of cloud services in malicious software: Trends and insights. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 187–204. Springer, 2015.

[14] IMS Health. IMS Health. http://www.imshealth.com/portal/site/imshealth, 2014. [Online].

[15] John P John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martin Abadi. deSEO: Combating Search-Result Poisoning. In *Proc. 20th USENIX Security Symposium*, San Francisco, CA, August 2011.

[16] Amy N Langville and Carl D Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

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

[18] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A Nearly Four-Year Longitudinal Study of Search-Engine Poisoning. In *Proc. 21st Conference on Computer and Communications Security (CCS)*, Scottsdale, AZ, October 2014.

[19] Kirill Levchenko, Neha Chachra, Brandon Enright, Mark Felegyhazi, 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 *Proc. IEEE Symposium on Security and Privacy*, Oakland, CA, May 2011.

[20] Alan A Lew. Long tail tourism: New geographies for marketing niche tourism products. *Journal of Travel & Tourism Marketing*, 25(3-4):409–419, 2008.

[21] Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M. Voelker, Stefan Savage, and Kirill Levchenko. PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs. In *Proc. 21st USENIX Security Symposium*, Bellevue, WA, August 2012.

[22] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[23] Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. Fashion crimes: trending-term exploitation on the web. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 455–466. ACM, 2011.

[24] Martin Mulazzani, Sebastian Schrittwieser, Manuel Leithner, and Markus Huber. Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space. In *Proc. 20th USENIX Security Symposium*, San Francisco, CA, August 2011.

[25] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting Spam Web Pages through Content Analysis. In *Proc. 15th International Wordl Wide Web Conference (WWW)*, Edinburgh, Scotland, May 2006.

[26] John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, 1998.

[27] Reviewopedia. Reviewopedia. http://www.reviewopedia.com/, 2015. [Online].

[28] Thomas Ristenpart, Eran Tromer, Hovav Shacham, and Stefan Savage. Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds. In *Proc. 21st Conference on Computer and*

Communications Security (CCS), Chicago, IL, November 2009.

[29] Dmitry Samosseiko. The Partnerka âĂŤ What Is It, and Why Should You Care? . In *Proc. of Virus Bulletin Conference*, Geneva, Switzerland, September 2009.

[30] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking Web Spam with HTML Style Similarities. *ACM Transactions on the Web*, 2(1), 2008.

[31] John Wadleigh, Jake Drew, and Tyler Moore. The e-commerce market for lemons: Identification and analysis of websites selling counterfeit goods. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1188–1197. International World Wide Web Conferences Steering Committee, 2015.

[32] David Y Wang, Stefan Savage, and Geoffrey M Voelker. Cloak and dagger: dynamics of web search cloaking. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 477–490. ACM, 2011.

[33] Baoning Wu and Brian D. Davison. Identifying Link Farm Spam Pages. In *Proc. 14th International Wordl Wide Web Conference (WWW)*, Chiba, Japan, May 2005.

[34] Yahoo. Yahoo! Content Analysis API. `https://developer.yahoo.com/contentanalysis`, 2015. [Online].

[35] Yinqian Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Cross-Tenant Side-Channel Attacks in PaaS Clouds. In *Proc. 21st Conference on Computer and Communications Security (CCS)*, Scottsdale, AZ, October 2014.