Evolutionary Study of Web Spam: Webb Spam Corpus 2011 versus Webb Spam Corpus 2006

De Wang, Danesh Irani, and Calton Pu
College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0765
Email: {wang6, danesh, calton}@cc.gatech.edu

***Abstract*—With over 2.5 hours a day spent browsing websites online [1] and with over a billion pages [2], identifying and detecting web spam is an important problem. Although large corpora of legitimate web pages are available to researchers, the same cannot be said about web spam or spam web pages.**

We introduce the Webb Spam Corpus 2011 — a corpus of approximately $330,000$ spam web pages — which we make available to researchers in the fight against spam. By having a standard corpus available, researchers can collaborate better on developing and reporting results of spam filtering techniques. The corpus contains web pages crawled from links found in over 6.3 million spam emails. We analyze multiple aspects of this corpus including redirection, HTTP headers and web page content.

We also provide insights into changes in web spam since the last Webb Spam Corpus was released in 2006. These insights include: 1) spammers manipulate social media in spreading spam; 2) HTTP headers also change over time (e.g. hosting IP addresses of web spam appear in more IP ranges); 3) Web spam content has evolved but the majority of content is still scam.

***Index Terms*—web spam, evolutionary, spam corpus.**

I. Introduction

Web spam is defined as web pages that are created to manipulate search engines and deceive web users [3], [4]. Email has long been the primary method to spread web spam, although spammers are evolving with the times and quickly employing new techniques to spread web spam. One clear trend is the move towards social media due to the ease of sharing information providing more efficient and numerous channels for the growth of web spam. For example, web spam links in friend requests, inbox messages, and news feeds, are redirecting users to advertisement web sites or other types of malicious web sites. Further, social media sites have redefined the way links are shared with a tendency to share links using URL shorteners [5].

Apart from evolution of web and applications on the web being one of the reasons driving change in web spam, there is a constant evolution of spam as a reaction to defensive techniques introduced by researchers [6], [7]. Improvements in defensive techniques used in web spam are enabled by researchers having access to corpora of web spam and being able to collaborate on developing and reporting results on web spam filtering techniques.

In this paper we introduce the Webb Spam Corpus 2011, a new corpus of approximately 330,000 spam web pages. We compare this corpus with the previous, and first of its kind, web spam corpus [7] released in 2006. More concretely, we make the following contributions:

First, we create a new large-scale Web spam corpus – Webb Spam Corpus 2011 – which is a collection of approximately 330,000 spam web pages. Web spam links are extracted from spam email messages received between May 2010 to November 2010. Additionally, we also perform data cleansing to remove legitimate pages which may have been inadvertently collected (similar to the data cleansing performed in the prior Webb Spam Corpus by Webb et al. [8]).

Second, we analyze the Webb Spam Corpus 2011 from various perspectives. For example, we evaluate the new corpus on three main aspects: redirections, HTTP session information and content. Based on these aspects, we also make insightful observations. For example, when investigating legitimate web link attack in data cleansing, we found that social networks and search engines have become major targets of attacks.

Lastly, we studied the evolution of web spam by comparing Webb Spam Corpus 2011 with Webb Spam Corpus 2006. For redirections, Webb Spam Corpus 2011 has less redirection. Specifically, it has less "302 Found" redirections and location redirection but more iframe redirections. The host names in redirection chains have new category – social networks sites, which indicates that social media have been manipulated to spread Web spam through hosting profiles, like plug-ins, and widgets. For HTTP session information, the percentages of hosting IP addresses for web spam in the ranges of 63.* -69.* and 204.* -216.* have changed from 45.4% and 38.6% in Webb Spam Corpus to 28.1% and 21.7% respectively. Additionally, we compared the top 10 HTTP headers in the datasets. In terms of content, there are few exact content duplications between the datasets. We also compared the contents of the datasets from other content aspects: most popular words, top words based on information gain, and n-gram(n is from 2 to 3) sequences based on frequency.

The remainder of the paper is organized as follows. We motivate the problem further in Section II. Section III introduces corpus including the data collection and cleansing methods. Section IV compares Webb Spam Corpus 2011 with

COLLABORATECOM 2012, October 14-17, Pittsburgh, United States

DOI 10.4108/icst.collaboratecom.2012.250689

Webb Spam Corpus. We discuss related work in Section V and conclude the paper in Section VI.

II. Motivation

Web spam has received a lot of attention with search engines constantly adjusting techniques to identify web spam [9] and social networks trying to prevent web spam propagating through their networks [10]. With web links being one of the most popular and easiest ways to share information on the web, web spam will remain a problem.

One of the most common technique to fight web spam is using machine learning, more specifically supervised learning techniques, to build classifiers for web spam using headers, content, or link features. As a prerequisite to using such techniques or researching new ones, having access to a large amount of labeled web spam is important and thus we collect, cleanse, and release a corpus of web spam as an enabler for researchers to improve and develop new web spam techniques. A standard corpus released for any number of researchers to use, as is the case with our corpus, allows and encourages collaboration between researchers to share and improve on each others results.

Although the release of the previous Webb Spam Corpus achieved this a number of years ago, we found that web spam has changed significantly enough to warrant an update to the Webb Spam Corpus. Namely, as detection techniques improve, spammers evolve and introduce new techniques to avoid detection. A concrete example of this is popular tools such as URL shorteners (which reduce the length of a URL by mapping an identifier on a standard web link to a long URL) were quickly picked-up by spammers as a cheap method of obfuscation and redirection. Further, looking back at the year of 2006, social networks such as Facebook do not exist or are in early stage of startup or microblog sites such as Twitter. Thus, not only do we release the Webb Spam Corpus 2011, with real-time data collection, we also provide an analysis of evolution and major changes we have observed between the 2006 and 2011 version of the Webb Spam Corpus.

III. Webb Spam Corpus 2011

In this section, we introduce the data collection method, as well as the data cleansing process for the Webb Spam Corpus 2011.

A. Data Collection

1) Collection Method: We introduce the Webb Spam Corpus 2011 which is available for download for collaborative research investigation and reporting as an .arff file (Weka file format[1]) at the Webb Spam Corpus' home page—http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html. The two main parts involved in creating the Webb Spam Corpus 2011 are data collection and data cleansing. These steps are detailed below and a high-level overview of the process is provided in Figure 1.

[1]Weka: http://www.cs.waikato.ac.nz/ml/weka/

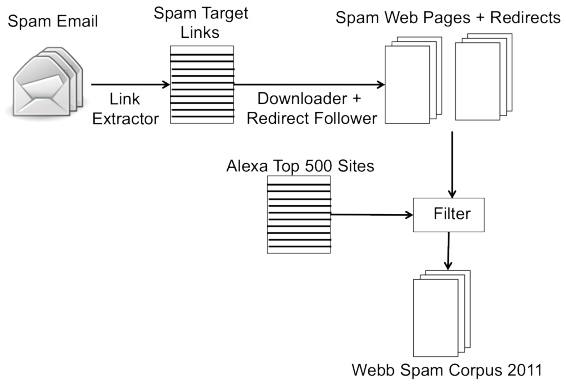

Fig. 1. Illustration of data collection and data cleansing process

2) Source URL and Actual URL: We distinguish URL links into two groups: source URLs and actual URLs. Here, source URLs are the original URLs extracted from email messages and are typically what the end user will see in the email message. Actual URLs are the final URLs or the URL of the web page that the user finally sees in their browsers. That is, this is the final URL after all redirects (http redirects, javascript redirects, meta-tag redirects, and more) have been followed. If a web page does not redirect a user, the actual URL could be the same as the source URL. To clarify this, the relationship between source URL and actual URL is shown in Figure 2:

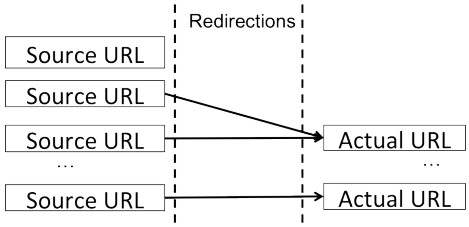

Fig. 2. The relationship between source URL and actual URL

The relationship between source URL and actual URL has the following characteristics:

a). One redirection chain leads from source URL to actual URL;
b). Many source URLs may redirect/map to a single actual URL;
c). Source URLs which were successfully accessed without resulting in a redirect is actual URL.

3) Source URL links: We start with a set of source URLs extracted from 6.3 million spam emails collected between May 2010 to November 2010 to a moderately sized email service provider. We only extract http and https URLs (although https links make up only 0.2% of all the spam links we extracted), using Perl's URI::Find::Schemeless and Html::LinkExtor modules to extract URLs from text and HTML respectively. We end up with 30.7 million web links (15.1 million unique links). Figure 3 shows the distribution of URL links in months. We also investigate the top level domains in source URLs and list top 10 TLDs in Table I. "RU" is top level domain for Russian Federation and "DE" is top level domain for Federal Republic of Germany. In this

study, we focus on English language web pages only, which are about 1.7 million web pages (before cleansing) which were crawled in March 2011.

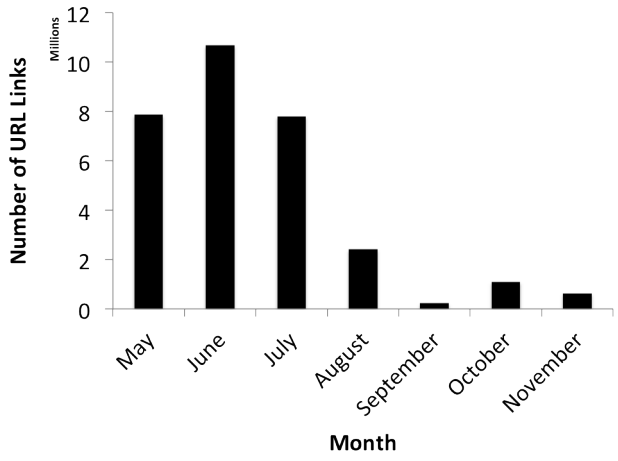

Fig. 3. Distribution of source URL links in months

TABLE I
LIST OF TOP LEVEL DOMAINS IN SOURCE URLS

Top Level Domain	Number of Unique Source URLs
RU	10,052,443
COM	3,063,766
NET	205,311
UK	191,583
INFO	168,192
DE	125,472
NL	117,099
PL	106,287
IP Addresses	13,263
Other	1,061,023

4) Web spam download: Once we have a set of source URLs, we proceed to download all the web pages. We use a custom crawler written using Perl's LWP::Parallel::UserAgent module to download corresponding web pages. We then follow any iFrame-redirects, http-redirects, javascript redirects (using Mozilla's Rhino), or meta-tag redirects. More details can be found in [7] which uses similar techniques. We keep the raw headers and HTML content of the page, and do not crawl or spider links from it. We downloaded a total of 1.7 million pages (including redirections) and in-total collected over 1 GB of data.

B. Data Cleansing

Data cleansing on Webb Spam Corpus 2011 is also split into two parts:

1) Removing False-positives: False-positives in corpus include legitimate URLs and error pages. Spammers often include legitimate URLs in spam emails to avoid spam rules or to appear legitimate [8]. Using Alexa's top 500 site list[2], we list top 10 legitimate actual URLs in Webb Spam Corpus 2011 shown in Figure 4.

[2]http://www.alexa.com/topsites

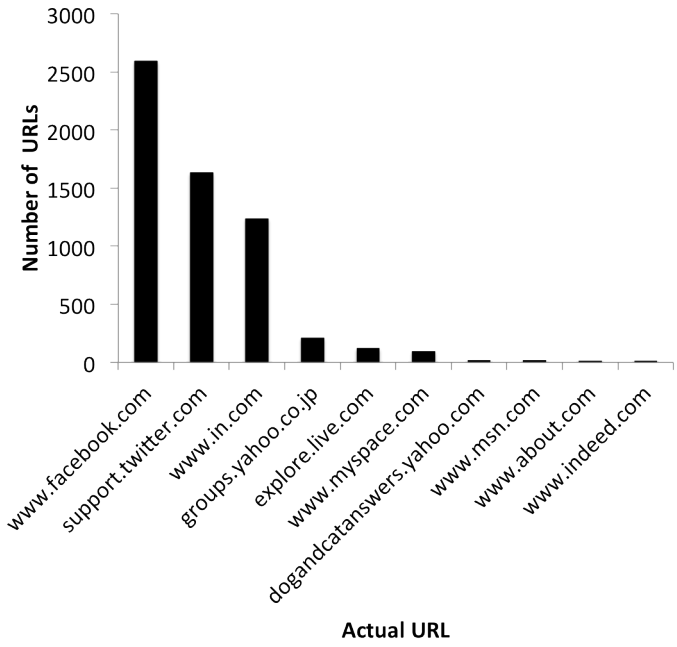

Fig. 4. Top 10 legitimate actual URLs in Webb Spam Corpus 2011

Figure 4 shows that 4 social networks websites (www.facebook.com, support.twitter.com, www.in.com, and www.myspace.com), 5 search engines (groups.yahoo.co.jp, explore.live.com, dogandcatanswers.yahoo.com, www.about.com, and www.indeed.com), and 1 information portal (www.msn.com) are in the top 10 list. It indicates that spammers are using popular social networks and search engines in legitimate URL attack. We removed $6,175$ legitimate actual URLs and $6,494$ legitimate source URLs in this process.

Besides legitimate URL links in spam emails, the downloaded web pages also contain other false-positives. Although these actual URLs may have been spam URLs, due to the delay in setting up our downloading and cleansing system, the spam URLs were crawled a few months after the source URLs were extracted. This resulted in a number of 404 HTTP errors or custom served "404 error web pages".

We eliminate such pages as well as previously mentioned false-positives leaving us with $673,489$ spam web pages in the corpus.

2) Removing Non-Textual Web Pages: Approximately 98% of web pages identify their "Content-type" as text/html. After cleansing false-positives in corpus, we discard non text/html pages. By removing non-textual web pages based on the attribute "Content-Type" in HTTP header information, we kept $673,313$ web pages including $342,478$ redirections.

C. Data Statistics

After finishing downloading all web pages, we investigate the distribution of top level domains and HTTP status codes. The purpose is to find which top level domain hosts the most web spam and the most common HTTP responses when we click through those spam URL links.

To obtain popular top level domains, we process the dataset in the following steps. First, we collect all top level domains from IANA Data[3], which contains 313 top level domains (last updated Jun 20, 2012). By matching all the source URLs in downloaded files with the top level domains list, we aggregated the count of web pages in the same top level domain. The 10 most popular top level domains are shown in Figure 5. We see that the three most popular top level domains COM, ORG, and NET almost represent more than 80% of the TLDs. Especially, the percentage of web pages which are belonging to top level domain COM is over 60%.

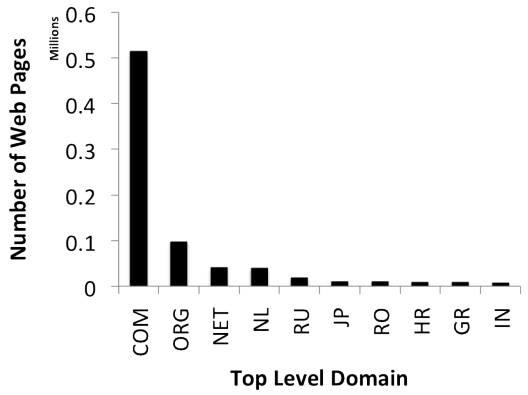

Fig. 5. Top 10 top level domains

For HTTP status codes, we aggregate all status codes based on the number of web pages and list the distribution of status codes shown in Figure 6. It shows that "200 OK" is the most common of status code in Webb Spam Corpus 2011 – over 70%. Also other status codes which are primarily used in redirection, such as "302 Found", "301 Moved Permanently", and "302 Moved Temporarily", are quite popular.

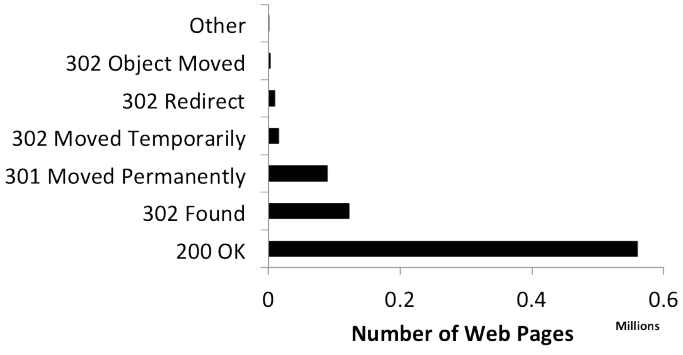

Fig. 6. Distribution of HTTP status codes

IV. Comparison between Two Datasets

We compare the Webb Spam Corpus 2011 with Webb Spam Corpus 2006 in three dimensions: redirections, HTTP session information, and content.

[3]http://data.iana.org/TLD/tlds-alpha-by-domain.txt

A. Redirections

Redirections are normally used by spammers to camouflage the actual spam URL links and avoid being blocked by URL blacklists. We look into redirections returned by source URLs in the Webb Spam Corpus 2011 shown in Table II.

TABLE II
Number of redirects returned by source URLs

Number of Redirects	Number of Source URLs
0	254,315
1	15,075
2	2,880
3	387
4	1361
5	86
6	58
7	46
8	31
9	27
10	26
11	19
12	13
13	15

To compare fairly with redirections in the Webb Spam Corpus 2006, we compute the percentage of source URLs versus number of redirections shown in Figure 7. It shows that Webb Spam Corpus 2011 has more source URLs returning no redirects (more source URLs which are also the actual URLs). The possible reasons are as follows: a) spammers are using less redirections for camouflaging actual spam URLs; b) Webb Spam Corpus 2011 has more URL links than Webb Spam Corpus 2006; c) there may exist false positives in Webb Spam Corpus 2011 before data cleansing.

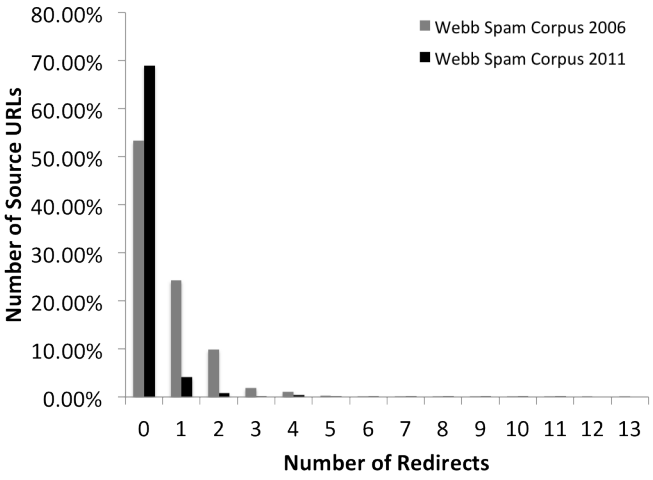

Fig. 7. Comparison based on percentage of source URLs vs number of redirections

We also aggregate source URLs based on the actual URLs they are mapping to and generate the distribution of number of source URLs that point to the same actual URL shown in Figure 8. It shows similar trend as the distribution of the number of source URLs that point to the same actual URL in Webb Spam Corpus 2006.

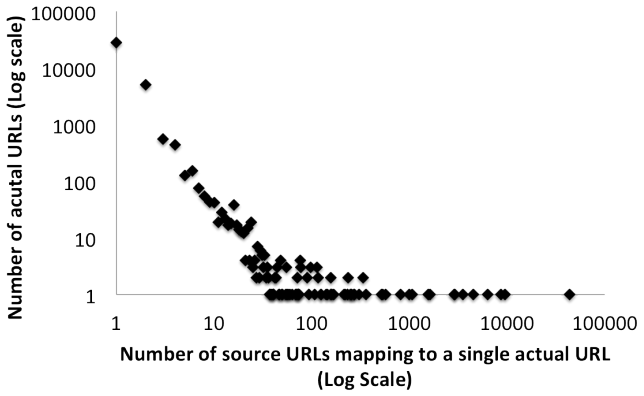

Fig. 8. Distribution of the number of source URLs that point to the same actual URL

Redirections have different categories including HTTP redirect, frame redirect, iFrame redirect, meta-refresh redirect and location redirect [8]. For HTTP redirect, it also has some subcategories based on response status such as "301 Moved" HTTP redirect and "302 Found" HTTP redirect. We compare the redirection distribution of two datasets which is shown in Figure 9.

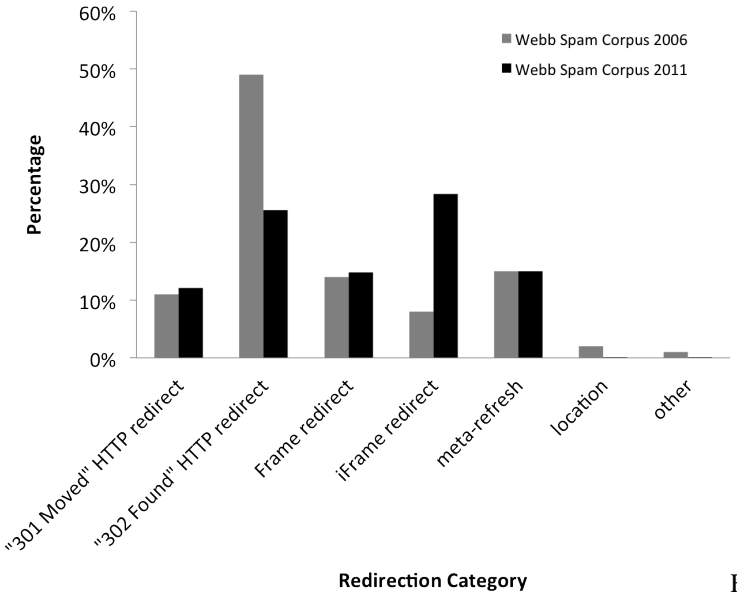

Fig. 9. Comparison between redirection distributions of the two datasets

Figure 9 shows HTTP redirect in Webb Spam Corpus 2011 still occupies the majority of redirections, accounting for 41.7% of the redirections (25.6% for "Found" redirects, 12.1% for "Moved Permanently" redirects, and 4.0% for other HTTP redirects). HTML frame and HTML iFrame redirects account for 14.8% and 28.4% respectively. Redirection using meta-refresh tags account for 15.0% and location redirect accounts for less than 1% of all redirects.

We observe that Webb Spam Corpus 2011 has fewer "302 Found" redirections and location redirection. But it has more iFrame redirections. Meanwhile, we found that Webb Spam Corpus 2011 has other HTTP redirects which occupies 4% redirections. The response status examples of other HTTP redirects includes: a) "302 Object moved"; b) "302 Moved Temporarily"; c) "302 Redirect".

Besides showing the distribution of redirections, we also look into the common host names in redirection chains which will tell us what kinds of websites have been taken advantage of by the spammers. The most common host names in redirection chains including HTTP redirection, frame redirection, iFrame redirection, and meta-refresh redirection are shown in Table III.

TABLE III
MOST COMMON HOST NAMES IN REDIRECTION CHAINS

Top 5 host names in redirection chain	
Host name	**Count**
domdex.com	59,004
www.facebook.com	37,580
domains.googlesyndication.com	9,934
bodisparking.com	9,530
potentbusy.com	9,431
Top 5 host names of HTTP redirection	
Host name	**Count**
mrs45.hosteur.com	9,046
home.wanadoo.nl	8,624
arpitjain.in	6,054
sharepoint.microsoft.com	4,596
www.in.com	4,336
Top 5 host names of frame redirection	
Host name	**Count**
bodisparking.com	9,530
potentbusy.com	9,430
www.ndparking.com	7,192
www.sedoparking.com	1,306
searchportal.information.com	1,209
Top 5 host names of iframe redirection	
Host name	**Count**
domdex.com	59,004
www.facebook.com	14,960
ad.doubleclick.net	2,649
areasnap.com	2,219
bullishcasino.com	1,672
Top 5 host names of meta refresh redirection	
Host name	**Count**
www.facebook.com	19,931
domains.googlesyndication.com	9,875
www.lawtw.com	6,736
www2.searchresultsdirect.com	1,838
www.sedoparking.com	1,472

From Table III, we investigated all the host names and found that there are three major categories: domain parking websites, social networks websites, and advertiser websites. For example, bodisparking.com and sedoparking.com are domain parking websites. facebook.com and in.com are social networks websites. ad.doubleclick.net is advertiser websites. The first set of counts represent the view of all of the HTTP, HTML, and JavaScript redirection techniques. This list consists of 3 domain parking services, 1 advertiser, and 1 social networks. The top 5 HTTP redirect host names consist of 1 domain parking service, 3 advertisers and 1 social networks. The top 5 frame redirect host names consist of 3 domain parking services, 2 advertisers. The top 5 iframe redirect host names consist of 1 domain parking services, 3 advertisers, and 1 social networks. The top 5 meta refresh redirect host names

consist of 3 domain parking services, 1 advertiser, and 1 social networks.

Domain parking for idle domains is used to display advertisements and earn money. It is easy to understand that spammers are using these domains for monetary benefit. Advertisers are similar to domain parking services on displaying advertisements which may not be useful for users . For social networks websites, we studied in detail about Facebook URLs in Webb Spam Corpus 2011. We found that the majority of redirections from Facebook belongs to iFrame redirection, meta-refresh redirection and HTTP redirection. In iFrame redirection, there are three types of URL redirections based on the sub path of URL links: "connect", "plugins", and "widgets", which accounts for 72.6%, 24.4%, and 3% respectively. Also the "connect" URL link redirects users to the profiles hosted Facebook. In our dataset, 10,820 "connect" URL link redirects to "t35.com" profile hosting in Facebook. "t35.com" is a domain parking services website. For 3,655 "plugins" URL links, 3,379 of them are "like" box plugin and 140 of them are "activity" plugin. Normally, if you click on "like" box plugin, you will become a fan of events, products, or profiles so that you will be kept updated with news feeds and status changes. For "activity" plugin, you will join the activity if you click on it. "Widgets" URL links are similar to "plugins" URL links. 444 "widgets" URL links provide "like" button for users to click. Therefore, we can conclude that spammers are using the power of social networks to spread spam information.

B. HTTP Session Information

Webb Spam Corpus 2011 also contains the HTTP session information that was obtained from the servers that were hosting those pages. In this section, we compare two datasets focusing on the most common server IP addresses and session header values.

1) Hosting IP Addresses: Hosting IP address is the IP address that hosts a given web spam page. Figure 10 shows the distribution of all of the hosting IP addresses over network number in Webb Spam Corpus 2011. Here network number is the first 8 bits of IPV4 address. Previous study [8] said that the 63.* -69.* and 204.* -216.* IP address ranges account for 45.4% and 38.6% of the hosting IP addresses respectively in Webb Spam Corpus. While in Webb Spam Corpus 2011, the percentages of IP addresses in those two ranges change to 28.1% and 21.7% respectively. Another two IP address ranges 70.* -100.* and 170.* -203.* account for 21.3% and 14.0% of the hosting IP addresses respectively in Webb Spam Corpus 2011.

It implies that spammers are comprising more various hosting IP addresses to spread web spam. The reason may be the IP blacklists used in popular anti-spam filters which force spammers to use new IP addresses for hosting web spam. To investigate most popular hosting IP addresses in Webb Spam Corpus 2011, we list top 10 hosting IP addresses based on the count of web pages. Meanwhile, through whois service, we obtain the server location and ISP (Internet service provider) for every hosting IP address.

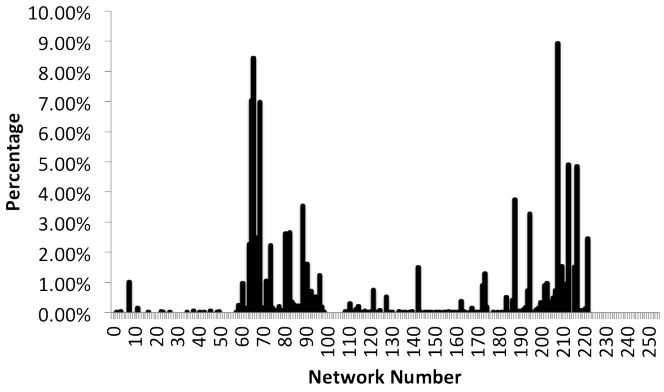

Fig. 10. Distribution of hosting IP address

Table IV shows 4 IP addresses from 63.* -69.*, 2 IP addresses from 204.* -216.*, and 4 IP addresses from other ranges. Also, it shows that 4 IP addresses from US servers, 2 IP addresses from France servers, and 4 IP addresses from other countries (Australia, Korea, Netherlands, and Germany). We can see that all servers are legitimate servers which doesn't mean those legitimate servers are the spammers. It only means the web services provided by those servers are used by the spammers for the spamming purpose.

2) HTTP Session Header: Previous study [11] has shown that HTTP session information is used for predict web spam efficiently. As the evolution of web spam, we intend to see whether HTTP session information of web spam has changed over time. To obtain most popular HTTP session information, we rank out top 10 HTTP session headers based on the count of web spam which those headers are associated with, shown in Table V.

Compared with top 10 HTTP session headers, Table V shows some changes as follows: a). new header P3P appears in top 10 list and old header PRAGMA has been removed from the list; b) the most popular values for the header SERVER and CONTENT-LENGTH have changed from "microsoft-iis/6.0" to "Apache" and from $1,470$ to 77 respectively; c). the order of the header CONTENT-LENGTH moves before X-POWERED-BY but the others keep the same relative order. Also, we find that 79.1% of the web spam pages with a SERVER header were hosted by "Apache" (60.5%) or "Microsoft IIS" (18.6%). In Webb Spam Corpus, 94.2% of the web spam pages with a SERVER header were hosted by "Apache" (63.9%) or "Microsoft IIS" (30.3%). Most popular value for the header CONTENT-LENGTH is not able to show the trend of content length so we also obtain the distribution of content length shown in Figure 11.

Figure 11 shows the average value of content-length is between $1,000$ and $1,0000$ although the most popular value is 77 bytes. As more multimedia used in web spam, the content length of web spam text gradually becomes shorter. Another thing we also need to check is whether the content of web spam also evolve over time.

TABLE IV
TOP 10 HOSTING IP ADDRESSES

Hosting IP Address	Count	Server Location	ISP
208.073.210.029	23,785	Los Angeles, CA in United States	Oversee.net
065.055.011.238	21,205	Redmond, WA in United States	Microsoft Hosting
213.186.033.019	17,543	France	Ovh Systems
066.196.085.048	16,542	Sunnyvale, CA in United States	Inktomi Corporation
069.043.160.174	13,289	Beaumaris,Victoria In Australia	Castle Access
066.045.237.214	10,834	Secaucus,NJ in United States	Interserver
222.122.053.065	9,090	Seoul, Republic of Korea	Korea Telecom
217.016.006.170	9,073	France	AB Connect
195.189.117.037	8,624	Nijmegen, Gelderland in Netherlands	Bluedome Internet Application Services BV
188.040.054.131	8,538	Germany	Hetzner Online AG

TABLE V
TOP 10 HTTP SESSION HEADERS

Header	Total Count	Unique Count	Most Popular Value (Count)
CONTENT-TYPE	379,721	120	text/html(147,428)
SERVER	369,985	919	Apache(82,004)
CONNECTION	359,786	5	close(312,186)
CONTENT-LENGTH	271,654	12,004	77(22,039)
X-POWERED-BY	148,944	191	ASP.NET(70,088)
CACHE-CONTROL	141,062	585	private(70,712)
SET-COOKIE	134,063	116,522	parkinglot=1;domain=.potentbusy.com;path=/;(3931)
LINK	122,352	5,012	http://l.yimg.com/d/lib/yg/css/dynamic_200602130000.css; rel="stylesheet";type="text/css"(15,446)
P3P	92,591	248	policyref="http://www.dsnextgen.com/w3c/p3p.xml"(24,180)
EXPIRES	90,915	7,668	Mon, 26 Jul 1997 05:00:00 GMT(25,641)

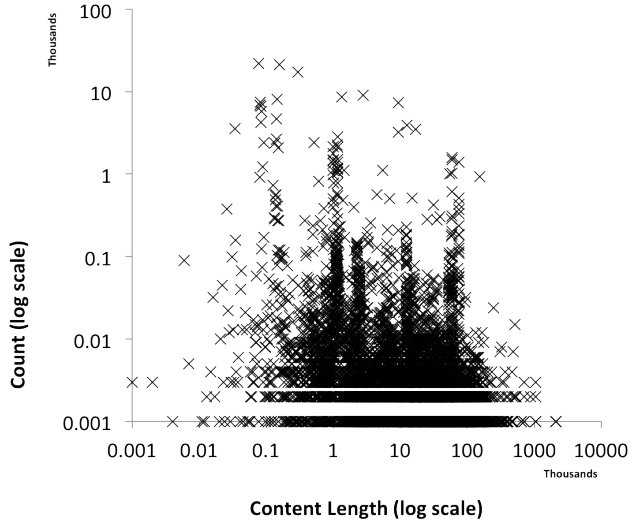

Fig. 11. Distribution of content length

C. Content

In this section, we compare two datasets on duplications and syntax changes between them. For duplications, we try to find the overlap between them based on MD5 hash values of content of web spam. For syntax changes, we intend to obtain the evolution of web spam syntax by comparing information gain of words and n-gram phrases.

1) Duplications: We compute MD5 hashes on the content of HTML web pages when we crawl the URL links. After evaluating these results, we find that there are $122,618$ unique MD5 values in Webb Spam Corpus 2011. Thus, $247,367$ of the web spam pages (66.9%) have the same HTML content as one of $122,618$ unique web spam pages. The percentage of exact content duplicates is much higher than the percentage (42%) in Webb Spam Corpus 2006 [7]. One possible reason is more URL duplications in the Webb Spam Corpus 2011.

To check the duplications between the two datasets, we iteratively compared MD5 codes of every web spam in Webb Spam Corpus 2011 and Webb Spam Corpus. The result of comparison is that $7,257$ web spam in Webb Spam Corpus 2011 are overlap with $2,834$ web spam in Webb Spam Corpus 2006. The percentages of duplications between two datasets are 2.0% and 1.3% in Webb Spam Corpus 2011 and Webb Spam Corpus 2006 respectively. Therefore, there are very few exact content duplicates existing between the two datasets.

2) Syntax Analysis: We analyze syntax of Webb Spam Corpus 2011 by computing the information gain of words in the content of web pages. Information gain, which is also called Kullback-Leibler divergence [12] in information theory, is calculated based on entropy as follows:

$$IG(T, a) = H(T) - H(T|a) \quad (1)$$

Here, T denotes a set of training examples and a presents the ath attribute of instance. $H(T)$ is the entropy of T and $H(T|a)$ is the conditional entropy of T with knowing the value of a.

Taking every web page as document, we adopt a bag of words model [13] to generate document instances in binary features. First, we need to tokenize the documents. Tokenization is the process of splitting the document up into words, phrases, symbols, or other meaningful elements called tokens. The features are the tokens in all documents and the value of feature is false if the token appears in the document or true if

not.

For the words in web pages, we first list top 20 most popular words in Webb Spam Corpus and their appearance as a percentage of documents that contain them, shown in Figure 12.

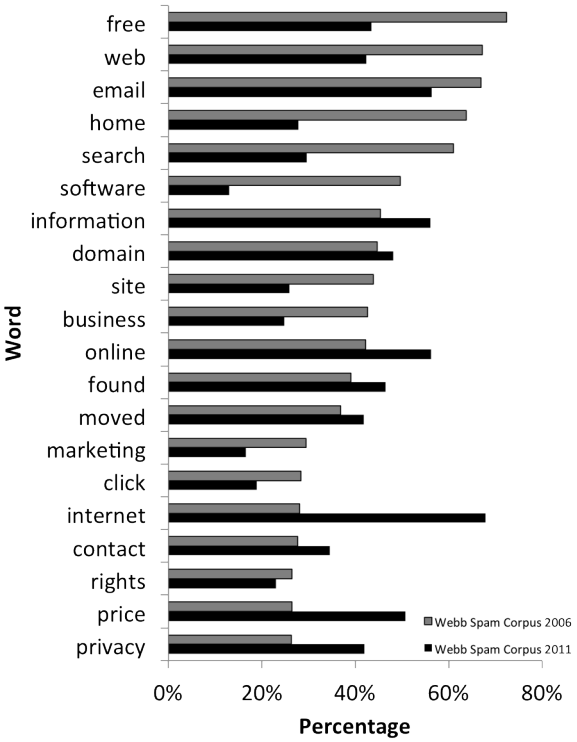

Fig. 12. Top 20 most popular words in Webb Spam Corpus [2006/2011] vs. percentage of documents that contain them in two datasets

Figure 12 shows that some words in the top 20 list appear less than in Webb Spam Corpus 2011 such as "free", "web", "home", "search", and "software". Some words appear more frequently than in Webb Spam Corpus 2011 such as "information", "online", "internet" and "price". It indicates the trend of spammy words and changes over time.

Besides most popular words, we also look into the discriminative words which distinguish two datasets. We ranked them by the value of their information gain according to the formula and used different labels to mark the instances in two datasets. The result of top 10 words based on information gain is shown in Figure 13.

Figure 13 shows top 10 words based on information gain. We further found that all words except "playlist" appear in Webb Spam Corpus 2006 while only four words including "playlist" , "vault", "cio", and "advertisement" present in Webb Spam Corpus 2011. Since we transformed all words into lower case format, words such as "cio" and "itworld" should be "CIO" and "ITworld". Word "playlist" normally appears in multimedia section of social media. For example, user profile has the embedded radio player which has a playlist for visitors.

Moreover, we compared n-gram (n is from 2 to 3) sequences in the two datasets. After using Perl's Text::Ngrams

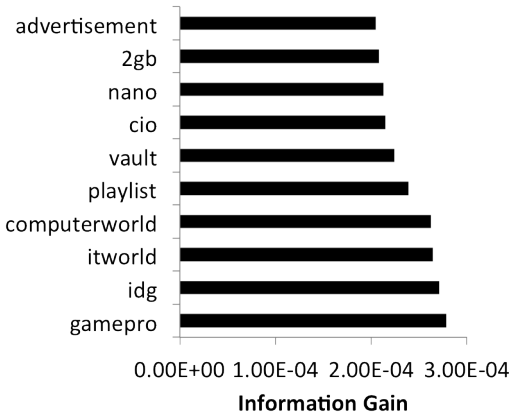

Fig. 13. Top 10 words based on information gain

module[4], we list top 20 n-gram (n is in the range of from 2 to 3) in two datasets based on frequency shown in Table VI.

In Table VI, $\langle N \rangle$ denotes any number sequence. Also we have removed redirections and the grams which only contain number sequences. Webb Spam Corpus 2011 has $22,894,416$ 2-gram sequences and $14,223,621$ 3-gram sequences, compared with $17,049,809$ 2-gram sequences and $6,488,343$ 3-gram sequences in Webb Spam Corpus 2006. Table VI shows that there are more numeric sequences appearing in 2-gram sequences in Webb Spam Corpus 2006 than in Webb Spam Corpus 2011. 3-gram sequences in Webb Spam Corpus 2006 are more related to links and search while those in Webb Spam Corpus 2011 are more related to price and money.

V. Related Work

Webb et al. [7] introduced the first large-scale dataset - the Webb Spam Corpus which is a collection of approximately 330,000 web spam pages. It addressed the challenge of the lack of publicly available corpora in previous Web spam research [14], [15], [16], [17], [18]. Further, they conducted intensive experimental study of web spam through content and HTTP session analysis on it [8]. They categorized Web spam into five groups: Ad Farms, Parked Domains, Advertisements, Pornography, and Redirection. Besides, they performed HTTP session analysis and obtained several interesting findings. After that, Webb et al. [11] presented a predicative approach to Web spam classification using HTTP session information (i.e., hosting IP addresses and HTTP session headers). They found that HTTP session classifier effectively detected 88.2% of the Web spam pages with low a false positive rate 0.4%. Our work is to further experimental study on evolution of web spam through content and HTTP session analysis on new Webb Spam Corpus. By comparing the two large-scale datasets in different time ranges, we obtained the trend of Web spam and behavior changes of spammers.

Fetterly et al. [19] presented their work on a large-scale study of the evolution of web pages through measuring the rate and degree of web page changes over a significant period

[4]http://search.cpan.org/dist/Text-Ngrams/Ngrams.pm

TABLE VI
TOP 20 n-GRAM (n IS FROM 2 TO 3) SEQUENCES BASED ON FREQUENCY IN THE TWO DATASETS

Webb Spam Corpus 2006				Webb Spam Corpus 2011			
2-gram	**Frequency**	**3-gram**	**Frequency**	**2-gram**	**Frequency**	**3-gram**	**Frequency**
of the	149,029	just a few	26,585	of the	212,626	w ⟨N⟩ org	138,162
in the	88,505	⟨N⟩ x ⟨N⟩	26,488	http www	169,180	http www w	126,770
V ⟨N⟩	77,254	is just a	26,016	w ⟨N⟩	140,524	www w ⟨N⟩	126,770
to the	77,050	the links below	25,910	⟨N⟩ org	138,247	⟨N⟩ org ⟨N⟩	92,935
on the	72,948	links below to	25,834	Price ⟨N⟩	127,091	org ⟨N⟩ ⟨N⟩	91,219
⟨N⟩ A	71,207	for your favorite	25,801	www w	126,770	mg x ⟨N⟩	73,110
v ⟨N⟩	66,725	a few clicks	25,799	in the	126,273	⟨N⟩ mg x	73,110
X ⟨N⟩	64,701	the search box	25,750	USD ⟨N⟩	117,108	⟨N⟩ ⟨N⟩ xmlenc	69,898
a ⟨N⟩	63,490	the Web for	25,723	Related Searches	103,259	⟨N⟩ USD ⟨N⟩	63,904
⟨N⟩ x	63,019	looking for is	25,705	Save ⟨N⟩	100,710	Found The doument	58,506
⟨N⟩ D	60,603	to search the	25,689	x ⟨N⟩	99,327	Found Found The	58,424
B ⟨N⟩	59,164	search the Web	25,658	org ⟨N⟩	93,803	⟨N⟩ Found Found	58,424
x ⟨N⟩	58,455	below to search	25,636	Privacy Policy	93,328	You Save ⟨N⟩	58,127
A ⟨N⟩	57,568	few clicks away	25,633	to the	91,951	⟨N⟩ You Save	58,103
may be	57,522	Use the search	25,632	hair loss	77,774	Price ⟨N⟩ You	56,065
⟨N⟩ GB	56,328	search box above	25,632	Internet Bellen	77,544	Admin Page Insights	54,726
⟨N⟩ a	55,437	above or the	25,625	⟨N⟩ mg	74,103	Retail Price ⟨N⟩	54,400
Price ⟨N⟩	55,424	Whatever you re	25,623	mg x	73,110	⟨N⟩ Retail Price	54,058
⟨N⟩ B	55,153	or hte links	25,622	on the	71,891	Download Price ⟨N⟩	54,049
⟨N⟩ s	53,330	Web for your	25,619	for the	70,177	Price ⟨N⟩ Retail	53,986

of time. They focused on statistical analysis on the degree of change of different classes of pages. Youngjoo Chung [6] studied the evolution and emergence of web spam in three-yearly large-scale of Japanese Web archives which contains 83 million links. His work focus on the evolution of web spam based on sizes, topics and hostnames of link farms, including hijacked sites which are continuously attacked by spammers and spam link generators which will generate link to spam pages in the future. Irani et al. [20] studied the evolution of phishing email messages and they classified phishing messages into flash attacks and non-flash attacks and analyzed transitory features and pervasive features. In our paper, we also studied the evolution of web spam but there are two important ways which are different from his work: First, we focus on redirection techniques, HTTP session information and content not link farms. Second, the majority of the datasets we study on is in English language not in Japanese. It may have common features between them but our datasets are more representative than his dataset in terms of the popularity of web spam in English language.

In previous research, we proposed a social spam detection framework for social networks [21]. We studied three popular objects in social networks including profile, message, and web page objects. The classification of web page model shows promising results for associative learning.

We collected new web spam corpus and studied the evolution of web spam. Our work addresses the lack of publicly available dataset for research and also shows the trend of web spam in social media.

VI. CONCLUSIONS

We introduced new large-scale web spam corpus – Webb Spam Corpus 2011 which is a collection of approximately 330,000 web spam pages. Adopting the automatic web spam collection method [7], we crawled the Internet through more than one million URL links in spam email messages during the time range between May 2010 and November 2010. In data cleansing of new dataset, we found that legitimate URL attacks by spammers are using more URLs in social media and search engine domains.

In addition to introducing new dataset, we also performed intensive study on Webb Spam Corpus 2011 through redirection, HTTP session analysis, and content. In redirection analysis, we found that less redirections in Webb Spam Corpus 2011 (about 70% source URLs returning no redirection). Another observation is Webb Spam Corpus 2011 has less 302 "Found" redirections and location redirection but it has more iframe redirections. Also Webb Spam Corpus has 4% redirections which are other types of HTTP redirections. For most common host names in redirection chains, we obtained a interesting finding that social networks are used for hosting web profile spam and the widgets and plugins of social networks become convenient spamming traps to attract click traffic. Furthermore, we investigated the HTTP session information of Web spam in Webb Spam Corpus 2011. For hosting IP addresses, the percentages of IP addresses in ranges 63.* -69.* and 204.* -216.* have been reduced from 45.4% to 28.1% and from 38.6% to 21.7% respectively. For HTTP session headers, new header P3P appears in top 10 list and old header PRAGMA has been removed from the list. The most popular values for the header SERVER and CONTENT-LENGTH have changed from "microsoft-iis/6.0" to "Apache" and from $1,470$ to 77 respectively. Also we generated the distribution of content length of Web spam and found the content length of web spam text gradually becomes shorter. Moreover, we analyzed duplications and syntax changes in Webb Spam Corpus 2011. 66.9% web spam pages in Webb Spam Corpus 2011 have the same HTML content as one of $122,618$ unique web spam pages, which is much higher than the percentage (42%) in Webb Spam Corpus. Two datasets have very few percentage of exact content duplicates in common (2.0% for Webb Spam Corpus 2011 and 1.3% for Webb Spam Corpus). For content

analysis, we listed the trend of top 20 most popular words in Webb Spam Corpus and top 10 words based on information gain to distinguish the two datasets. Also we compared n-gram (2-3) based on frequency in the two datasets.

To sum up, we collected a new Webb Spam Corpus of approximately 330,000 web pages. We derive insights from this dataset as well as do an evolutionary study by intensive analysis and comparison between Webb Spam Corpus 2011 and Webb Spam Corpus 2006. Also we obtained lots of interesting findings between them. The future work we plan on is the evaluation of classification comparison on new Webb Spam Corpus.

ACKNOWLEDGEMENTS

This research has been partially funded by National Science Foundation by IUCRC/FRP (1127904) , CISE/CNS (1138666), RAPID (1138666), CISE/CRI (0855180), NetSE (0905493) programs, and gifts, grants, or contracts from DARPA/I2O, Singapore Government, Fujitsu Labs, Wipro Applied Research, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

REFERENCES

[1] The harris interactive survey–trends and tudes 2009. [Online]. Available: http://www.harrisinteractive.com/vault/HI_TrendsTudes_2009_v08_i04.pdf

[2] J. Alpert and N. Hajaj. Google blog: We know the web was big. [Online]. Available: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[3] Z. Gyöngyi and H. Garcia-Molina, "Web spam taxonomy," in *Proceedings of 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, May 2005.

[4] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *Proceedings of the 30th International Conference on Very Large Databases (VLDB 04)*, Toronto, Canada, August 2004.

[5] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis, "we.b: the web of short urls," in *Proceedings of the 20th international conference on World wide web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 715–724.

[6] Y. Chung, "A study on the evolution and emergence of web spam," Ph.D. dissertation, Univ. of Tokyo, Tokyo, Japan, 2011. [Online]. Available: http://jairo.nii.ac.jp/0021/00024500/en

[7] S. Webb, J. Caverlee, and C. Pu, "Introducing the webb spam corpus: Using email spam to identify web spam automatically," in *Proceedings of the Third Conference on Email and Anti-Spam (CEAS 2006)*, Mountain View, CA, USA, July 2006.

[8] ——, "Characterizing web spam using content and http session analysis," in *Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS 2007)*, Mountain View, CA, USA, August 2007, pp. 84–89.

[9] M. Cutts. Google blog: Using data to fight webspam. [Online]. Available: http://googleblog.blogspot.com/2008/06/using-data-to-fight-webspam.html

[10] Twitter. Twitter blog: Shutting down spammers. [Online]. Available: http://blog.twitter.com/2012/04/shutting-down-spammers.html

[11] S. Webb, J. Caverlee, and C. Pu, "Predicting web spam with http session information," in *Proceedings of the Seventeenth Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, CA, USA, October 2008.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[13] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Proceedings of 10th European Conference on Machine Learning (ECML-98)*, Springer Verlag, Heidelberg, DE, August 1998, pp. 4–15.

[14] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The connectivity sonar: detecting site functionality by structural patterns," in *In Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*. ACM Press, 2003, pp. 38–47.

[15] A. A. Benczur, K. Csalogany, T. Sarlos, M. Uher, and M. Uher, "Spamrank - fully automatic link spam detection," in *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[16] K. Chandrinos, I. Androutsopoulos, G. Paliouras, and C. D. Spyropoulos, "Automatic web rating: Filtering obscene content on the web," in *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ser. ECDL '00. London, UK, UK: Springer-Verlag, 2000, pp. 403–406.

[17] B. D. Davison, "Recognizing nepotistic links on the web," in *In AAAI-2000 Workshop on Artificial Intelligence for Web Search*. AAAI Press, 2000, pp. 23–28.

[18] I. Drost and T. Scheffer, "Thwarting the nigritude ultramarine: learning to identify link spam," in *In Proceedings of the 16th European Conference on Machine Learning (ECML*, 2005, pp. 233–243.

[19] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th international conference on World Wide Web*, ser. WWW '03, New York, NY, USA, 2003, pp. 669–678.

[20] D. Irani, S. Webb, J. Giffin, and C. Pu, "Evolutionary study of phishing," *eCrime Researchers Summit, 2008*, pp. 1–10, 2008.

[21] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 11)*, Perth, Australia, September 2011, pp. 46–54.