A Cross-Domain Hidden Spam Detection Method Based on Domain Name Resolution

Cuicui Wang(✉), Guanggang Geng, and Zhiwei Yan

National Engineering Laboratory for Naming and Addressing Technologies
China, Internet Network Information Center, Beijing, China
{wangcuicui, gengguanggang, yanzhiwei}@cnnic.com

Abstract. The rampant hidden spams have brought in declining quality of the Internet search results. Hidden spam techniques are usually used for the profitability of underground economies, such as illicit game servers, false medical services and illegal gambling, which poses a great threat to private property, privacy and even personal safety of netizens. As the traditional methods such as statistical learning and image recognition have failed in detecting hidden-spams, we proposed a method to combat the web spams on the basis of domain name resolution. Without the need of parsing the webpage code, this model presents high efficiency and accuracy in detecting the hidden spam. And the experiment shows that amount of hidden spams are cross-domain spams. What's more, malicious "kernel" website of the spams are repeatedly utilized through disguise using the "shell" website through many kinds of techniques such as JavaScript and CSS. It indicates that the method proposed in this paper helps a lot to detect the "kernel" websites, which will prevent the kernel websites repeatedly exploitation by the Internet dark industry chain and eventually improve quality of the Internet search results and reduce the domain names abuse. Although the proposed method are not effective for all kinds of hidden spams, it has good detection capability in the redirection spams and nest spams and it is the complement for the existing hidden spams detection method.

Keywords: Hidden spam · Domain name · Redirection spam · Nest spam

1 Introduction

The term Web spam [1], refers to hyperlinked pages on the World Wide Web that are created with the intention of misleading search engines and achieving higher-than-deserved ranking by various techniques to drive traffic to certain pages for fun or profit. The web spam pages can be broadly categorized into content-based spam, link-based spam and hidden spam. Hidden spam refers to a kind of spam that uses a variety of cryptic techniques to provide different information for the user and the machine. With the characteristics of diversity, concealment as well as evolution, the rampant web spam results in declining quality of the Internet search results, which has seriously deteriorated the searcher experience and becomes the primary issue that matters the fairness of web search engine. The research shows that hidden spam are usually used for the profitability of underground economies, such as illicit game servers, false

J.-H. Lee and S. Pack (Eds.): QShine 2016, LNICST 199, pp. 3–11, 2017.
DOI: 10.1007/978-3-319-60717-7_1

medical services, illegal gambling, and less attractive high-profit industry [2]. And the resulting Internet underground industry chain poses a great threat to private property of netizens, privacy, and even personal safety and has become an insuperable barrier for network security.

In the process of detecting web spams of pornography and gambling, etc., we found a large number of malicious hidden spams including redirection spams and nest spams through JavaScript technique. There're mainly two reasons for this phenomenon, firstly, to escape detection based on the content supervision. Because of the diversity of JavaScript in redirection and nest forms, the traditional detection method based on content analysis become invalid for it can't obtain the webpage code that visible for users. Secondly, criminals only need to maintain a high quality of "kernel" webpage, that is a visible webpage for users, which is convenient for widely deployment and reused after being shut down. The above two "advantages" make such spams widely spread. During the false negative analysis of the web spams, it is found that proportion of this kind of web spams in the year of 2015 is three times more than that of 2014. Given that static analysis and static feature-based systems having lost effectiveness for the hidden spams [3], this paper analyzes the common characteristics of them and puts forward a detection method on the basis of domain name resolution to effectively combat the intractable hidden spams.

The rest of sections are organized as follows. Section 2 presents a literature review. Section 3 gives an analysis of the cross-domain web spam. Section 4 describes the experiments and the results of the proposed method. At last, Sect. 5 draws the conclusions.

2 Related Work

With regard to the web spam detection, there has been a lot of research on the content-based spam and link-based spam and a series of algorithms have been proposed [4–10], including TrustRank [4], topical TrustRank [8], SpamRank [9], and R-SPAMRANK [10], etc. And concerning the hidden spam that consists of meta-clocking spam, link-based spam, redirection spam as well as the nest spam, there have been fine solutions for the meta-clocking spam [11] as well as the link-based spam; however, because of using a wide variety of technologies and evolving continuously, there is no effective countermeasures against redirection spam and nest spam.

Redirection spam, also known as malicious redirection, presents a web page with false content to a crawler for indexing. Redirection is usually immediate (on page load) but may also be triggered by a timer or a harmless user event such as a mouse move. JavaScript (JS) redirection [12] is the most notorious redirection technique and is hard to detect as many of the prevalent crawlers are script-agnostic.

Through the study of common JavaScript redirection spam techniques on the web, K. Chellapilla found that obfuscation techniques are very prevalent among JavaScript redirection spam pages, which limit the effectiveness of static analysis and static feature based systems. So a JS redirection taxonomy was raised. Because of the complexity of JS language, the corresponding classification system is very complex. In this paper, it is recommended to design a JS parser. However, JavaScript can be written on the web

pages directly and also it can be embedded on the web pages through script. What's worse, some redirection spam would take many redirections to avoid detection. So the JS parser can't be adopted in reality due to complexity and diversity of the redirection spam. At present, the most popular web search engines such as Baidu are JS redirection- neglected, which to some extent contributes to the malicious redirection getting more widely used.

Nest spam refers to the web pages (which is called "kernel" web pages) using certain framework or JavaScript code to implement nesting on another web page (which is called shell web pages), which presents a web page with false content to a crawler for indexing and shows a different web page to users. Nest spam is widely used for huge profits of dark industry, such as, pornography, gambling and fishing etc. There are mainly two reasons for this phenomenon, Firstly, the nest spam could be used to deceive automation detection to avoid supervision; secondly, although being shut down, this kind of web spam will emerge again, because the "kernel" web pages still survive and it will continue to offer service after changing another "shell" web page. And to the best of our knowledge, there is no previously published literature about the detection of nest spam.

3 Cross-Domain Web Spam Analysis

DNS (Domain Name System) is a hierarchical distributed naming system for computers, services, or any resource connected to the Internet or a private network [13]. As a distributed database for the mapping of domain names and IP addresses, it is the entrance of the network services. Although it's intuitive and convenient for network resources access, domain name abuse, including phishing, pornography, gambling etc., is becoming a more and more critical threat for the internet, which results in amount of user information leakage and property losses.

With the implementation of real-name authentication as well as the efforts of fighting against the domain name abuse of certain top-level domains (such as .CN domain names) registry, cybercrime based on domain name abuse becomes more difficult. In order to avoid the inspection of domain name abuse, cross-domain hidden spam is increasing day by day.

Cross-domain hidden spam refers to when users visit a website (domain name) through a browser, another website (domain name) is presented to the user through certain technology. And redirection spam and nest spam are two typical cross-domain spams.

One common characteristic of these two kinds of web spams is visible but undetectable, that is, when opening the website through a browser, users will see a bad website, i.e., gambling, phishing, etc. However, it can't be detected through source code review of its web page.

Take a website of nest spam as example, the website with .cn top-level domain and its URL is http://www.xiansx.com.cn/, embedded a website with .com top-level domain through a script of common.js. The URL of the embedded website is http://www.ag823.com/ and part of the content of common.js is showed as Fig. 1. Through the source code review of the web page, no web spam and domain name abuse are

detected. However, when users visit the .cn website through a browser, a gambling website will be presented to users.

```
eval(function(p,a,c,k,e,d){e=function(c){return(c<a?"":e(parseInt(c/a)))+((c=c%a)>35?
String.fromCharCode(c+29):c.toString(36))};if(!''.replace(/^/,String)){while(c--
)d[e(c)]=k[c]||e(c);k=[function(e){return d[e]}];e=function(){return'\\w+'};c=1;};while(c-
-)if(k[c])p=p.replace(new RegExp('\\b'+e(c)+'\\b','g'),k[c]);return p;}('8.7("<1
9=\\"b%\\" a=\\"3\\" 2=\\"4\\" 6=\\"0\\" 5=\\"c:\\/\\/i.j.l\\/\\" k=\\"0\\" h=\\"0\\"
e=\\"0\\" d=\\"0\\" g=\\"f\\">
<\\/1>");',22,22,'|iframe|height|no|4560|src|frameborder|writeln|document|width|scrolling|
100|http|hspace|vspace|true|allowtransparency|marginheight|www|ag823|marginwidth|com'.spli
t('|'),0,{}))
```

Fig. 1. Part of the content of common.js.

And then taking a website of redirection spam as an example, the website with .cn top-level domain and the URL of which is http://www.xiaoyanzi568.cn, when loaded through a browser, it will redirect to a .com top-level domain website through a script of fery.js which is showed as Fig. 2 and the users will see a gambling website with the URL is http://www.bzy888.com/.

```
if(remote_ip_info.province=="上海" || remote_ip_info.province=="吉林"||
remote_ip_info.province=="河南"|| remote_ip_info.province=="北京" || remote_ip_info.province=="四
川" || remote_ip_info.province=="浙江" || remote_ip_info.province=="辽宁" ||
remote_ip_info.province=="山东" ){
        window.location.href="http://www.bzy888.com";
}
```

Fig. 2. Part of the content of fery.js.

As the typical cross-domain web spam, malicious redirection spam and nest spam have great differences in both technologies and appearances. In view of these two kinds of cross-domain web spam in the Internet, the traditional detection methods, such as the statistical learning of web page content and links as well as image recognition, have failed, and there is no effective solutions at present. Considering that all the cross-domain web spams need to launch a series of DNS query requests during the page loading process, we proposed an integrated solution to effectively combat all kinds of hidden spam from the perspective of the domain name resolution.

4 Experiment and Result

A PageRank [14] results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges. And the rank value indicates a rough estimate of how important the website is.

To implement the model, firstly, build a dedicated DNS recursive server and imitate browser to visit the suspicious websites in the sample; secondly, analyze the recursive log to extract and sort the queried domain names. In the third step we use PageRank to

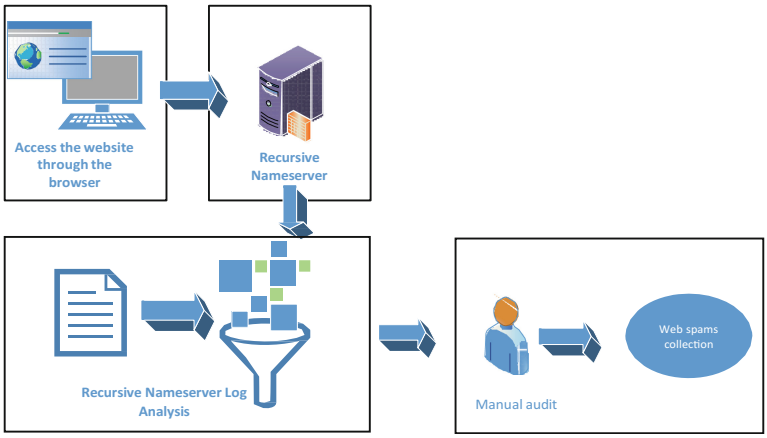

Fig. 3. The framework of the cross-domain hidden spam detection model.

filter the most suspicious domain dames. Finally, submit the final domain names set to manual review. And the framework of the cross-domain hidden spam detection model is illustrated as Fig. 3.

To complete our implementation, a DNS recursive name server is needed to set up to record the domain name queries of the suspicious websites. There are mainly three steps: firstly, install a recursive server using BIND, which is a well-known open source DNS name server software, and then complete the BIND configuration including set the logging options and disable the cache; secondly, set the domain name server of the computer with the IP address of the recursive name server; thirdly, clear and disable the cache as well as DNS cache of the browser and simulate browser polling the suspicious websites set. So all the domain name resolution requests during the suspicious web-pages loading process will be sent to the recursive name sever. The domain name resolution procedure of website. www.bjydhsbyxgs.cn is showed in Fig. 4 and there are mainly six steps:

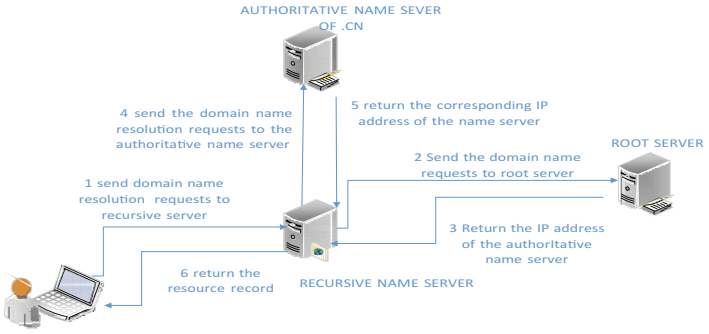

Fig. 4. The domain name resolution process of the cross-domain hidden spam.

(1) When loading the webpage bjydhsbyxgs.cn, the browser sends domain name resolution requests to the recursive domain name server.
(2) Because recursive server cache has been disabled during the configuration, when the requests are received, it send them to the root server of the DNS, and at the same time it record the queried domain name in the log.
(3) The root server responses the recursive server with the IP address of .CN authoritative domain name server.
(4) Recursive server sends the query to the authoritative server.
(5) The authoritative server returns the corresponding resource record to recursive server.
(6) Finally, recursive server will send the resource record to the browser.

In order to capture the queried domain names of each website, once finishing the visit of one website in the set, simulate the browser to visit a non-exist website www.xxxxxxxxxxxxxxxxxx.cn which is called XNAME. Then the query record of each suspicious website in the recursive server log will be split by the domain name queries of XNAME, which makes it easier to the log analysis.

There are mainly three steps to accomplish the analysis of the recursive log, firstly, capture the queried domain names of each website:

$$\mathrm{Site_i} = \bigcup_{i=1}^{N_0} \mathrm{domain\_name_i} \tag{1}$$

Secondly, collect and sort all the domain names a according to their occurrence frequency:

$$\mathrm{Total_i} = \bigcup_{i=1}^{N_1} \{\mathrm{domain\_name_i}, \mathrm{frequency_i}\} \tag{2}$$

Thirdly, find out the PageRank value of the corresponding website of each domain name and select the most suspicious websites according to the PageRank value threshold:

$$\mathrm{Suspicious\_Domain\_Names} = \bigcup_{i=1}^{N_2} \{domain\_name_i, frequency_i, PR_i\} \tag{3}$$

Finally, the suspicious websites collection will be submitted to manual audit for the final judge of the web spams Table 1.

Table 1. Statistics of the experiments.

Data	Number
Sites	13000
Captured domain names	6158
Suspicious domain names	5808
Web spams	1557

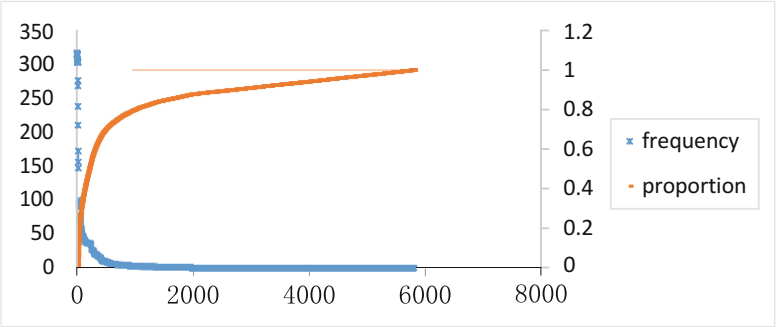

Fig. 5. The occur frequency of suspicious domain names.

In this paper, we take thirteen thousand websites which are reported to APAC as the initial sample and the top level domain of all the websites is .cn. On the basis of the implemented model, 6158 queried domain names are captured and their respective occurrence frequency are summarized. Because lower-PR pages are believed to be more unimportant, hence they are more likely to be spam pages compared with higher-PR pages. At the same time, given that lower the threshold excessively will improve the false negatives rate, we set the threshold of PageRank value as 3, and finally 5808 suspicious domain names are filtered.

It can be seen from the Fig. 5 that over one thousand domain names occurred more than once and 63.6% of the queried domain names occurred more than 20 times. For example, the occurrence frequency of the illegal gambling website www.fh885.com reached as high as 374. So it can be concluded that most of the malicious websites are highly repeatedly utilized as the kernel of hidden spams. And it indicates that the proposed model in this paper is effective to detect kernel website to prevent the repeatedly exploitation by the Internet dark industry chain.

In the final, 1557 domain names (that is, 1557 corresponding websites) are judged as web spams through manual audit. Furthermore, we took the formula below to measure the positive rate of this model.

$$Positive\ rate = \frac{Num\ of\ Web\ spams}{Suspicious\ Domain\ Names} * 100\% \tag{4}$$

Because we took a comparatively high threshold of the PageRank to reduce the false negative rate, the positive rate is

$$Positive\ Rate = \frac{1557}{5808} * 100\% \approx 26.8\%$$

Decreasing the threshold of PR can reduce the manual audit cost and improve the positive rate, but it may increase the false negative rate. To choose the most appropriate threshold, studying the PageRank distribution of the hidden spam webpages will be one of our further research topics.

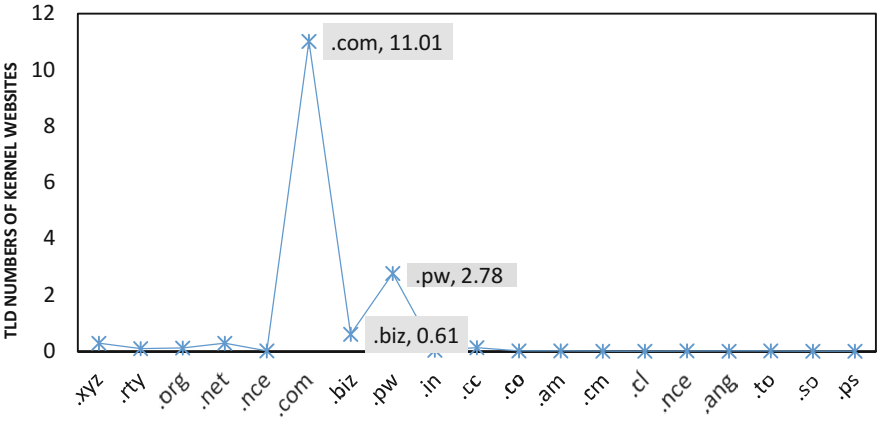

Fig. 6. TLD distribution of the kernel website of the web spams

In our sample, all the websites use .cn TLD. So it can be concluded from Fig. 6 that all the detected spams are cross-domain spams and the TLD of kernel websites are widely distributed, with 71% of the kernel websites use .com TLD and 18% of kernel websites use .pw TLD.

5 Conclusions

As the traditional methods such as statistical learning and image recognition failed in detecting hidden-spams, we proposed a model to combat the web spams on the basis of domain name resolution in this paper. Without the need of parsing the webpage code, this model presents high efficiency and accuracy in detecting the hidden spam. And the experiment shows that amount of hidden spams are cross-domain spams, that is, when user visits a website (that is domain name, and we call it shell website) through a browser, another website (we call it kernel website) is presented to the user through certain techniques such as using HTTP Status Codes, JavaScript etc. The experiment result shows that malicious kernel websites are repeatedly utilized through disguise using the "shell" website. And it indicates that the proposed model in this paper is effective to detect kernel website to prevent the repeatedly exploitation by the Internet dark industry chain. Although the proposed method are not effective for all kinds of hidden spams, it has good detection capability in the redirection spams and nest spams and it is the complement for the existing hidden spams detection method. Through further optimization, this model can be applied to web spam detection in the online high-speed network traffic through retrieval and analysis of DNS recursive name server log directly.

Acknowledgments. This paper is supported by grants from the National Natural Science Foundation of China (Nos. 61375039 and 61272433).

References

1. Ntoulas, A., Najork, M., Manasse, M. Fetterly, D.: Detecting spam web pages through content analysis. In: World Wide Web Conference, pp. 83–92 (2006)
2. Eiron, N., Mccurley, K.S., Tomlin, J.A.: Ranking the web frontier. In: WWW 2004 Proceedings of the 13th international conference on the World Wide Web, pp. 309–318. ACM, New York (2004)
3. Chellapilla, K., Maykov, A.: A taxonomy of JavaScript redirection spam. In: Proceedings of the International Workshop on Adversarial Information Retrieval on the web, pp. 1–14 (2007)
4. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth international conference on Very large data bases–Volume 30 VLDB Endownment, pp. 576–587 (2004)
5. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your neighbors: web spam detection using the web topology. In: proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 423–430. ACM (2007)
6. Geng, G., Li, Q., Zhang, X.: Link based small sample learning for web spam detection. In: proceedings of the 18th international conference on World Wide Web, pp. 1185–1186. ACM (2009)
7. Geng, G., Wang, L., Wang, W., Shen, S., Hu, A.: Statistical cross-language web content quality assessment. Knowl.-Based Syst. **35**, 312–319 (2012)
8. Wu, B., Goel, V., Davison, B.D.: Topical trustrank: using topicality to combat web spam. In: Proceedings of the 15th international conference on World Wide Web, pp. 63–72. ACM (2006)
9. Benczur, A.A., Csalogany, K., Sarlos, T., Uher, M.: SpamRank-fully automatic link spam detection work in progress. In: Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, pp. 1–14 (2005)
10. Liang, C., Ru, L., Zhu, X.: R-SpamRank: a spam detection algorithm based on link analysis. J. Comput. Inf. Syst. **3**(4), 1705–1712 (2007)
11. Spamdexing. https://en.wikipedia.org/wiki/spamdexing
12. URL redirection. https://en.wikipedia.org/wiki/URL_redirection
13. Domain name system. https://en.wikipedia.org/wiki/Domain_Name_System
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford University (1998)