Analysis and Detection on Abused Wildcard Domain Names Based on DNS Logs

Guangxi Yu [1, 2], Yan Zhang[1, 2*], Huajun Cui[1], Xinghua Yang[1], Yang Li[1, 2], Huiran Yang[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[*]Corresponding author, email: zhangyan80@iie.ac.cn

Abstract. Wildcard record is a type of resource records (RRs) in DNS, which can allow any domain name in the same zone to map to a single record value. Former works have made use of DNS zone file data and domain name blacklists to understand the usage of wildcard domain names. In this paper, we analyze wildcard domain names in real network DNS logs, and present some novel findings. By analyzing web contents, we found that the proportion of domain names related to pornography and online gambling contents (referred as *abused* domain names in this work) in wildcard domain names is much higher than that in non-wildcard domain names. By analyzing behaviors of registration, resolution and maliciousness, we found that abused wildcard domain names have remarkably higher risks in security than normal wildcard domain names. Then, based on the analysis, we proposed GSCS algorithm to detect abused wildcard domain names. GSCS is based on a domain graph, which can give insights on the similarities of abused wildcard domain names' resolution behaviors. By applying spectral clustering algorithm and seed domains, GSCS can distinguish abused wildcard domain names from normal ones effectively. Experiments on real datasets indicate that GSCS can achieve about 86% detection rates with 5% seed domains, performing much better than BP algorithm.

Keywords: DNS, Abused Wildcard Domain Name, Analysis, Detection

1 Introduction

The Domain Name System (DNS) is an important part of critical Internet infrastructure, which aims to translate domain names into IP addresses. In fact, the mappings are recorded in different record types, called resource records (RRs). One of these record types is *wildcard* record, which can allow any domain name in the same zone to map to a single record value (i.e. IP or domain name). Wildcard RRs are original used to forward mail to the same zone [1]. But today, with the development of Internet applications and services, wildcard RRs are used widely. Besides normal applications and services, some malicious attacks also take advantage of wildcard RRs.

To understand the use of wildcard domain names, especially the malicious usage, some works have been presented. In paper [2], based on DNS zone files, researchers found that 17.8% domain names were wildcard domain names and 19.1% of them were involved in blackhat SEO. In paper [3], based on zone file data and domain

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

name blacklists, researchers found that wildcards are popular among all types of Internet domains. And among malicious users, spammers use wildcards the most. All these works have made rich achievements, but so far, there is no comprehensive study to analyze wildcards usage based on user request data (e.g. passive DNS data or DNS logs), which can directly express real query behaviors of wildcard domains.

In this paper, we first perform an analysis on wildcard domain names from perspectives of normal domain names and abused domain names. In this study, we regard the domain names related to pornography and online gambling as ABUSED domain names, because these contents are illegal in China[1]. And we regard all wildcard domain names except for abused ones as normal wildcard domain names. Being different from prior studies constructing dataset from zone files or known malicious domain lists, we analyze wildcard domain names in real network DNS logs, which are collected from a large ISP network containing millions of hosts. In addition, we collect auxiliary data including WHOIS information and web content information to gain an insight from original DNS logs. On the whole, we analyzed 919,939 domain names. We found that 153,163(17% of all) domain names are wildcard domain names. Then by analyzing the 66.4% wildcard domain names with web contents, we found that 22.5% of them are abused domain names (related to pornography and online gambling). What's more, by analyzing wildcard domain names' behaviors of registration, resolution and maliciousness, we also found that the abused wildcard domain names have remarkably higher risks in security than normal ones.

Then, based on the analysis of wildcard domain names, we propose a machine learning based algorithm named GSCS (Graph based Spectral Clustering with Seeds) to distinguish abused wildcard domain names from normal ones. Our GSCS algorithm includes the following steps. First, to discover the similarity of resolution behaviors, we build a bipartite DNS graph and its projection graph for abused domain names. Then, by applying simple and efficient spectral clustering algorithm on the similarity matrix of the projection graph, we can divide wildcard domain names into different clusters. Finally, based on seed domain names, we can discover inherent clustered groups of abused wildcard domain names. Our experiment results based on real datasets show that GSCS can detect abused wildcard domain names more effectively than BP (belief propagation) algorithm.

Our main contributions in this paper include:

- We found that the proportion of abused domain names (i.e., domain names related to pornography and online gambling contents) in wildcard domain names is much higher than that in non-wildcard domain names. Specifically, 22.5% versus 4.4%.
- We found that, compared with normal wildcard domain names, abused wildcard domain names have remarkably higher risks in security, including a higher proportion of domains related to malicious activities (10% versus 2.7%), a lower proportion of domains deploying SSL (2.3% versus 14%), and being more likely to avoid regulation (be registered out of China, in bulk and in recent years).

1 For illegal contents, we should note that various countries hold different laws. For example, all pornographic contents in the Internet are illegal in China, but only the contents with child pornography are illegal in U.S.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

– We propose an effective algorithm GSCS to detect abused wildcard domain names. Compared with the BP algorithm which can get only 72% detection rate, GSCS can improve the detection rate to 86%.

The rest of this paper is structured as follows. In Section 2, we provide background information on DNS and wildcard domain names. Section 3 describes the analysis of our dataset. In section 4, we make a comprehensive analysis of wildcard domain names based on web content and WHOIS information. And we propose an abused wildcard domain names detection algorithm in Section 5. Section 6 summarizes the related work. Finally, Section 7 concludes the paper's work.

2 Background

Domain name system. The domain name system is a hierarchical system, which contains local DNS servers, authority name servers and root servers. Correspondingly, a domain name is also a hierarchical string with each level related to a zone. In detail, a domain name *d* consists of a set of labels separated by dots; they are called top-level domain (TLD), second-level domain (2LD), third-level domain (3LD), etc., from right to left. TLDs are managed by registries such as CNNIC (for *cn*) and Versign (for *com* and *net*), and 2LDs are offered to public by registrars such as Alibaba and GoDaddy. Before using a domain name in the Internet, domain owner should get its 2LD from a registrar. Then, the WHOIS information of this domain name is updated to database. In general, for a domain name with a benign website, the meaning of domain name is related to content of website. However, malicious domains are usually not.

Wildcard domain names. Wildcard domain names are domain names starting with an asterisk label (*) to match non-existing subdomain names. Note that, names beginning with other labels are never wildcard domain names, and the asterisk at other places in the domain will also not work as a wildcard. As mentioned before, wildcard RRs can allow any domain name in the same zone to map to a single record value, and simple examples are shown in table 1.

Table 1. Simple examples of wildcard RR.

*.example1.com	3600	IN	MX 10	a.example1.com
*.example2.com	3600	IN	A	1.2.3.4

In the beginning, wildcard domain names are used to forward mails [1]. As shown in table 1, the MX RR would cause any MX query for any domain name ending in *example1.com* to return an MX RR pointing at *a.example1.com*. In the following, because many DNS implementations diverge from the original definition of wildcards, some other record types are extended [4]. In addition, several domain name registrars have also deployed wildcard records for TLDs to provide a platform for advertising. Some of these TLDs are country code TLDs (ccTLDs) such as *.fm*, *.la*, and there are also some Internationalized TLDs, for example the wildcard domain name **.中国* has been resolved to an IP address *218.241.116.40*, which belongs to CNNIC. Because wildcard TLD domain names are usually maintained by domain

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

name registrars, in this paper, we ignore these cases and only consider the wildcard domain names with 2LDs.

3 Data Collection

Previous works about wildcard domain names analysis collected dataset from some zone files or some malicious domain lists. In this paper, we collect data from real network DNS logs, and analyze the usages of wildcard domain names comprehensively. Additionally, we utilize auxiliary data like WHOIS and web content, etc. Below we elaborate our data.

DNS logs. We measure wildcard domain names by analyzing DNS real logs, which are generated by local DNS servers operated by a large ISP in China. These logs record the interactive information between local DNS servers and client hosts. As shown in table 2, each record in the logs consists of five fields. For the log data size, take a middle level province as an example, it is over 1.9TB per day. In this paper, we collected DNS logs over five days, from January 1 to January 5, 2018. Note that we only considered the normal queries with NOERROR response. Finally, we obtained 919,939 distinct domain names with different 2LD zones. Next, we make a comprehensive analysis based on these domain names.

Table 2. The form of a record in DNS logs

Source IP	Domain name	Query time	Destination IP	RCODE

Wildcard domain name. For each of 919,939 domain names, we queried its wildcard domain using the *dig* tool, and collected their responses together. For example, for *google.com*, we directly queried the wildcard domain name **.google.com*. In our study, we focus on A and CNAME records, because these two types of records are the main part of host queries. Finally, we collected 153,163 wildcard domain names, which accounts for about 17% of the total number of collected domain names. The result is similar to that obtained in paper [2].

WHOIS information. To obtain the registration information of the collected wildcard domain names, we leverage the WHOIS records published by registrars. We used the *Ruby whois*[2] tool to obtain WHOIS information of 135,785 (88.7% of all) wildcard domain names and used a python script to sparse them. For the missing of remaining domain names, the major reasons are request block and incomplete information provided by registrars.

Web content. We implemented a crawler to visit the websites of the collected wildcard domain names. Meanwhile, we also recorded the HTMLs and URLs for further analysis. We finally extracted 101,763 (66.4% of all the wildcard domain names) HTMLs. The two major reasons for missing web contents of the remaining domain names are the request timed out and websites lacking (i.e., Websites not exist).

2 https://whoisrb.org/

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

4 Analysis on Abused Wildcard Domain Names

Wildcard domain names offer DNS administrators the convenience of changing host names. However, problems do exist. In this section, we analyzed the usages of wildcard domain names through a series of automated and manual experiments, and then gave quantitative analysis based on these experiments. In detail, first, we grouped domain names into several categories according to text of HTMLs crawled in Section 3. Second, we analyzed the registration characteristics based on WHOIS data of the collected wildcard domain names. Next, we analyzed the resolution behaviors of these domain names. Finally, we checked the maliciousness and SSL deployment of the collected wildcard domain names.

4.1 Content Categories

Based on web content data from 101,763 (66.4% of all) wildcard domain names, we grouped these collected wildcard domain names into several categories using a semi-automatic method. We first manually looked into the title, page text of a few HTMLs and summarized seven main categories according to the key words of websites. For example, adult websites usually contain some descriptive words, such as porn, sex, gay, etc. Descriptions of seven main categories are as follows:

1) Porn. We define domain names linked to adult content like pornographic pictures, videos and novels as porn domain names;
2) Gambling. It refer to domains related to online gambling;
3) Parking. It refer to domains linked to ads constructed by domain-parking agency, based on the words included in a domain name;
4) Sale. Domain names sold over the Internet by domain agency are regarded as domains for sale;
5) Business & Gov. Domain names serve as normal business and government;
6) Entertainment. Domain names serve as entertainment content like games;
7) Error. Web pages of domain show an error caused by web servers.

Then, based on these key words belonging to different categories, we created generic content-signatures to automatically categorize the remaining pages into each category. Finally, we automatically classified all crawled webpages, and the results are shown in table 3.

Table 3. Content Categories

Categories	Number	Proportion	Description
Porn	19028	18.7%	Adult / Pornographic domain
Gambling	3888	3.8%	Online gambling domain
Parking	2032	2%	Parking Domain
Sale	5860	5.8%	Domain for sale
Business & Gov	28303	27.8%	Business/Government related domain
Entertainment	7206	7.1%	Entertainment/Game/Lottery, etc.
Error	9268	9.1%	Server error
Unclassified	26178	25.7%	Unclassified domain

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

Finding 1. Looking over the website categories, pornographic and online gambling websites take up a remarkable proportion of the wildcard domain names. Notably, about 22.5% crawled webpages contain adult and gambling contents, which are referred as ABUSED domain names. As a comparison, using the same method, we analyzed 100K non-wildcard domain names (i.e., domains without wildcard RRs). Finally we found that only about 4.4% crawled webpages contain pornographic or online gambling information. In China, pornographic and online gambling websites are banned by the Internet regulators. In other words, the relatively large proportion of websites of abused domain names suggests that wildcard domain names used to spread illegal information have not been regulated efficiently.

4.2 Registration Characteristics

As mentioned before, we obtained WHOIS information of 135,785 (88.7% of all) wildcard domain names. Based on the WHOIS data, in this subsection, we made a comprehensive analysis of wildcard domain name registration from perspectives of registrars, registrants and registration time windows. Specially, we studied registration behaviors by correlating domain names with their content categories. In the following, we summarize our findings.

Finding 2. Compared with normal wildcard domain names, abused wildcard domain names were much more likely to be registered out of China. We identified more than 2,100 registrars. For abused wildcard domain names and normal ones, the detailed distributions of the top 5 registrars are shown in table 4. Especially, Godaddy and Alibaba are two dominant registrars in domain market, and Alibaba plays an important role in China domain market. From table 4 we can see that, for abused wildcard domain names, only one registrar (Alibaba) in top 5 is from China, while for normal wildcard domain names, only one registrar (GoDaddy) in top 5 is not from China. This suggests that registrars out of China may hold loose conditions for registration, thus abused wildcard domain name owners like to register illegal domain names from them to avoid regulation.

Table 4. Categories of registrars (Top 5)

Abused Wildcard Domain Names	Ratio	Normal Wildcard Domain Names	Ratio
GoDaddy.com, LLC	18.3%	Alibaba Cloud Computing Co., Ltd.	25.5%
NameCheap Inc.	9.6%	GoDaddy.com, LLC	10.1%
Alibaba Cloud Computing Co., Ltd.	9.3%	Xin Net Technology Corporation	6.1%
NameSilo, LLC	6.6%	Chengdu West Dimension Digital Technology Co., Ltd.	5.7%
Danesco Trading Ltd.	3.9%	eName Technology Co.,Ltd.	4.3%

Finding 3. Compared with normal wildcard domain names, abused wildcard domain names were registered more recently. About 40% abused wildcard domain names were registered in recent one year, and about 53% were registered in recent two years. However, registration dates of domain names in other categories were scattered. Totally, about 26% normal domain names were registered in recent one year, and 35% were registered in recent two years. The cause of this phenomenon may be that in recent years more and more people try to register abused domain

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

names for great economic benefit, and they also try to avoid regulation by using a large number of new domain names.

Finding 4. Compared with normal wildcard domain names, abused wildcard domain names were more likely to be registered in bulk. As shown in figure 1, based on the data of created date, we compared the differences of registration characteristics between abused wildcard domain names and normal ones. Finally, by counting the days that have more than 20 registered domain names, we found that there are 8 days for normal wildcard domain names while 103 days for abused ones. Correspondingly, only 175 (0.2% of all) normal wildcard domain names were registered in these 8 days, and 4,239 (18.5% of all) abused wildcard domain names were registered in those 103 days. As a case, registrant *Li xiaoyu* registered 203 domain names in June 14, 2017, and resolved all of them to pornographic websites.

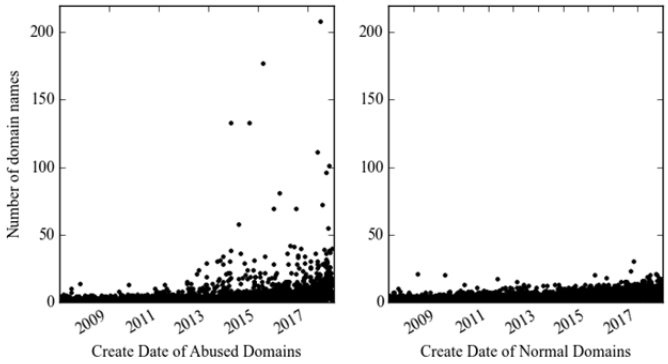

Fig. 1. Comparisons of abused and normal domains based on their created date data

4.3 Resolution Behaviors

To understand the resolution behaviors of wildcard domain names, we analyzed wildcard records of destination IPs and name servers. In this paper, we obtained destination IP and name server for each wildcard domain name by using the *dig* tool. We summarize our findings as follows.

Finding 5. For abused wildcard domain names, their destination IP addresses are relatively concentrated than those of normal wildcard domain names. We collected 90,897 IP addresses used by wildcard domain names and analyzed the IP distributions of abused wildcard domain names and normal ones. By analyzing /24 IP addresses, we found that the IP addresses of abused wildcard domain names are relatively concentrated than the normal ones. As shown in figure 2 (a), we could find that about 20% abused wildcard domain names were resolved to top 10 IP addresses, and top 100 IP addresses held about 50% abused wildcard domain names. We also collected 87,546 IP addresses used by non-wildcard domain names and compare their distributions with those of wildcard domain names. Results are shown in figure 2 (b), which

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

indicates that the IP addresses of wildcard domain names are relatively concentrated than those of non-wildcard ones.

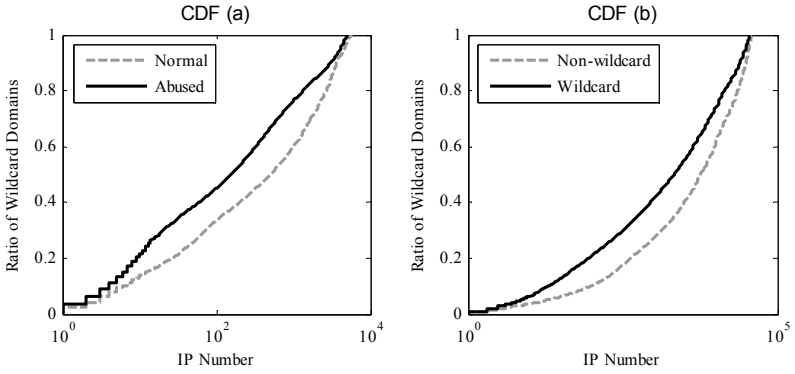

Fig. 2. CDF of IP addresses (IP numbers are shown in logarithmic coordinate)

Finding 6. For wildcard domain names resolved to the same IP, the name servers of abused ones are more concentrated than those of normal ones. To analyze the usage of name servers, we first grouped the wildcard domain names by each destination IP, and then we analyzed the number of distinct name servers used in each group.

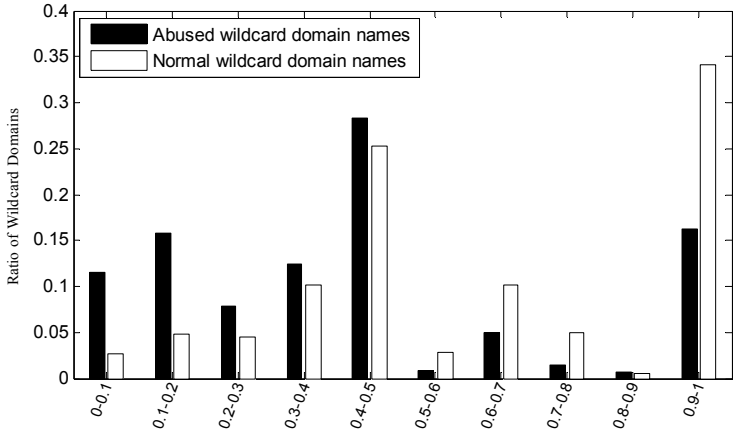

Fig. 3. Usage of name servers. The black bars describe the case of name servers used by abused domain names, and the white bars describe the case of name servers used by normal ones.

As shown in figure 3, the X-axis is the ratio of the number of name servers over the number of wildcard domain names in each group. A lower value of this ratio means more wildcard domain names are resolved by the same name server. And the Y-axis is the relative number of wildcard domain names within different ratio ranges. The results show that the proportions of abused wildcard domain names are always higher than those of normal ones when the value of X-axis is lower than 0.5, suggesting the name servers of abused wildcard domain names are more concentrated than those of normal ones.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

4.4 Maliciousness and Security of Domain Names

To analyze the malicious use of collected wildcard domain names, we checked these domain names with VirusTotal[3] and malicious domain lists, including DNS-BH[4], Malware Domain List[5], etc. Besides the malicious use, to analyze security of the collected wildcard domain names, we also checked SSL configuration of them.

Finding 7. The proportion of malicious domain names in abused wildcard domain list is apparently higher than that in normal wildcard domain list. Totally, we found 4,155 domain names were involved in malicious activities. When we looked into the categories of these malicious domain names, we found that about half of them were abused wildcard domain names. In other words, about 10% abused wildcard domain names were involved in malicious activities, however, only 2.7% normal wildcard domain names were related to malicious activities. This finding suggests a higher risk of being compromised for users when accessing websites with abused wildcard domain names. Obviously, pornography and online gambling contents provided by abused wildcard domain names are easy to allure victims.

Here, we also made a blackhat SEO analysis of wildcard domains based on our DNS logs. For blackhat SEO testing, we use the method proposed in paper [2], which only considers the difference of hyperlinks in webpages between two visits for a domain name. We randomly selected 5K domain names and we finally found that 77(1.5%) domain names are suspicious blackhat SEO domain names, which is much lower than 19.1% mentioned in paper [2]. In detail, 24.7% of these SEO domain names are abused domain names and 34.5% domain names are parking related domain names.

Finding 8. Only 2.3% abused wildcard domain names have adopted wildcard certificates to secure Internet traffic between users and web servers. As a contrast, about 14% normal wildcard domain names have adopted wildcard certificates. In today's Internet, HTTPS is a popular and effective information security protection method. Usually, web administrators adopt HTTPS only for several detailed domain names. For wildcard domain names, wildcard certificates can secure entire domains under the same zone with a single, flexible certificate. To analyze the SSL deployment of wildcard domain names, we extracted URLs of these domain names. Finally, discarding redirection, we found 11,306 URLs among all 101,763 domains with web contents adopted wildcard certifications. In detail, only 527 URLs belonged to abused domain names. To make a comparison, we also analyzed the SSL deployment of normal wildcard domain names. Finally, we found the application rate of SSL deployment in these domain names is higher than that in abused ones. This finding suggests that owners of abused wildcard domains rarely concern the security of transportation between their websites and users.

According to the above findings, we can see that abused wildcard domain names not only are related to illegal contents but also have higher risks in security than normal wildcard domain names. So it's necessary to distinguish wildcard domain names

3 https://www.virustotal.com

4 http://www.malwaredomains.com

5 https://www.malwaredomainlist.com

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

from normal ones. In the next section, we propose the GSCS algorithm to detect abused wildcard domain names.

5 Abused Domain Detection Based on DNS Graph

5.1 The GSCS Algorithm

In this section, to mine the relationships among abused wildcard domain names and detect them, we propose a graph-based method. In fact, graph-based method has already been used in malicious domain names detection [5-7]. Different from the former works, we exploit the inner relationships among abused wildcard domain names based on information of name servers and WHOIS. In addition, we avoid using traditional classification algorithms, which will be heavily influenced by unbalanced dataset of abused wildcard domain names.

We first describe the DNS graph model. Given a bipartite DNS graph $G = (D, I, E)$, the vertex set D and I consists of wildcard domain names and destination IPs, and the edge set E represents the connections between domain names and IPs. We then build a projection graph (named P) of bipartite G to extract hidden information between nodes in vertex set D. Here, we show the GSCS algorithm and introduce the detailed steps as follows:

Algorithm GSCS (Graph based Spectral Clustering with Seeds)

```
Input:   Wildcard domains
         IPs: destination IPs of wildcard domains
         Seeds: Abused domain name seeds
Output: Abused Clusters: Clusters of abused domain names
1.  G = Bulid_graph(Wildcard domains, IPs)
2.  P = Projection(G)
3.  for each connected component of P do
4.        if ns_consistency_score < confidence_threshold then
5.              sub_component = Spectral_cluster(component)
6.              for each element of sub_component do
7.                    if seed in element then
8.                          Move element to Abused Clusters
9.                    end
10.             end
11.       else
12.             Move component to Abused Clusters
13.       end
14. end
15. return Abused Clusters
```

- Using step 1 and 2, we transform records of wildcard domain names into a graph. In our analysis, we use /24 IPs to construct the bipartite graph, because IP addresses of collected wildcard domain names are relatively concentrated and /24 is a common block size of BGP routing prefixes [8].
- Through systematic analysis of the graph, we find that the projection graph consists of a large number of isolated components, so we analyze each of them separately (step 3-14).
- Based on *finding 6*, we use a consistency score to filter out components. The score is defined by

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

$$ns\_consistency\_score = \frac{max_{ns}}{num_{ns}}$$

Where, max_{ns} refers to the number of name server, which is used by the largest number of domain names in one component, and num_{ns} is the total number of name servers in one component. For example, if all domain names in one component are resolved by the same name server, the score is equal to 1.

— For other components, we first apply a spectral clustering algorithm to decompose each of them into sub-components. Then, using seeds of abused domain names, we filter out all sub-components with seeded domain names.

Next, we simply describe the spectral clustering algorithm used in algorithm GSCS. The key step in the spectral clustering algorithm is computing similarity matrix. In this paper, we construct similarity matrix using weight of edge, and the weight is defined by

$$weight = \begin{cases} 1 - \frac{|D_1 - D_2|}{T} & |D_1 - D_2| < T \\ \frac{1}{|D_1 - D_2|} & |D_1 - D_2| \geq T \end{cases}$$

Where, $|D_1 - D_2|$ is the interval of registration date between every two domain names. In detail, we extract the information of registration data from WHOIS data. Additionally, we set $T = 30$ based on results of several experiments and use X-means to cluster nodes of domain names.

5.2 Evaluation

Based on the information of wildcard domain names, we first built a domain resolution graph and its corresponding projection graph. Totally, we found 29,492 connected components in the projection graph. Next, to evaluate the effectiveness of our GSCS algorithm, we varied confidence threshold of the consistency score and seed size. In experiments, we set the threshold to 0.6, 0.7 and 0.8 respectively. Under the condition of each value, we randomly selected seeds from the abused wildcard domain name list, and set the seed size range from 1% to 10% with a step length of 1%. Then we calculated the true positive rate (TPR) and the false positive rate (FPR) based on different groups of seeds that are arranged in order of size. Finally, we found that both TPR and FPR increase with the size of seeds, and we drew the ROC of GSCS with different thresholds in figure 4. We can see that the performance when threshold is set to 0.8 is better than the performance when threshold is set to the other two values. Especially, when setting threshold to 0.8 with 5% seed domain names, we can get 86% TPR with 3% FPR. We can also see that, when we set a small seed size, we get low TPR and FPR. Because the graph is composed of many isolated components, the smaller seed size we set, the more components are discarded. Conversely, the smaller threshold we set, the more components are considered, so FPR will go higher. However, when the seed size goes higher than 5%, TPR will nearly not increase while FPR will still get higher. So, threshold 0.8 and seed size 5% should be appropriate choices for our GSCS algorithm.

For false positives, we found that most of them belonged to *Error* or *Unclassified* categories mentioned in subsection 4.1. For example, when setting threshold to 0.8

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

with 5% seeds, there were 2,171 false positives. However, we found about 70% of them belonged to *Error* or *Unclassified* categories, which suggests that these domain names would have been used as abused domains before they were discarded.

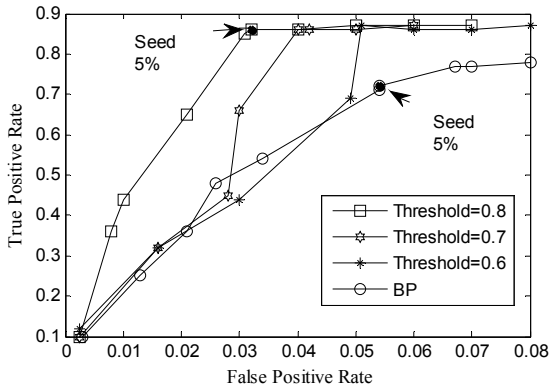

Fig. 4. ROC of various confidence thresholds

To further analyze the effectiveness of our detection method, we make a comparison with BP algorithm, which is often used in the field of graph analysis and has also been used to detect malicious domain names [6, 9]. In this paper, we used the abused wildcard domain names and normal wildcard domain names collected from our data as ground truth, and we assigned priors to graph nodes and edge potential matrices according to [6], which are shown in table 5 and table 6.

Table 5. Priors assigned to a domain node

Node	P(Abused)	P(Normal)
Abused	0.99	0.01
Normal	0.01	0.99
Unknown	0.5	0.5

Table 6. Edge potential matrices

	Abused	Normal
Abused	0.51	0.49
Normal	0.49	0.51

In our comparison experiments, we also set the seed size range from 1% to 10% with a step length of 1%, and we found both TPR and FPR increase with the size of seeds, as those in GSCS do. For the convenience of study, we show the results of BP in figure 4. We can see that our proposed method outperforms BP for the task of abused wildcard domain names detection. In detail, when using the same 5% seed domains used in GSCS, BP only obtained 72% TPR with 5.3% FPR. After analyzing, we found the factor of isolated components is a key reason leading to this inferior performance. Because we use a small number of seeds lying in several isolated components, other components without seeds cannot get information of propagation from these components with seeds.

From the above results and analysis, we can conclude that our GSCS algorithm is an effective solution to detect abused wildcard domain names.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

6 Related Work

Wildcard domain names. Wildcard record is a type of RRs, which has been widely used in the Internet. Now, several studies have been proposed and focused on security implications of wildcard domain names. Du *et al*. [2] conducted the first comprehensive investigation on wildcard domain names used for blackhat SEO technique called "spider pool". Based on DNS zone files, their research shows 17.8% of all domain names are wildcard domains and 19.1% wildcard domain names are used for spider pool. Similarly, based on DNS zone files and known malicious domain lists, Kalafut *et al.* [3] studied the prevalence of wildcard DNS configuration and showed that it is broadly involved in malicious behaviors. In addition, several studies also briefly mentioned usage of wildcard domain names. Liu *et al.* [10] mentioned that wildcard records were also used to spawn shadowed domains, which are malicious subdomains under legitimate domains compromised by miscreants. Sharifnya *et al*. [11] found wildcard records used in botnet to resolve to a C&C server.

Although wildcard domain name has been widely used, now the study has still not paid much attention to it. And there is no comprehensive study to analyze abused wildcards usage based on user request data. So our study can be regarded as a complementary research on wildcard domain names.

Graph-based detection method. As DNS data has the characteristics of graph, graph-based methods have been proposed to detect malicious or abused domain names. In general, graph-based methods can be divided into two categories, including hosts-domains graph and domain resolution graph. For hosts-domains graph, Lee *et al*. [5] proposed *GMAD*, a graph expressing DNS query sequences, to detect infected clients and malicious domain names. Using event logs collected by enterprises, Manadhata *et al*. [6] constructed a host-domain graph to detect malicious domain names. For domain resolution graph, Berger *et al*. [7] proposed a detection system called *DNSMap* to detect malicious website using dynamic FQDN-to-IP address mappings. In addition, to infer the maliciousness of unknown node in graph, some researchers chose to use BP algorithm [6, 9, 12]. To distinguish an abused wildcard use from a benign one, we proposed a detection method referencing graph idea.

7 Conclusion

In this paper, we first performed comprehensive analysis on wildcard domain names. Being different with former works that use DNS zone file data and domain name blacklists, our work is based on real DNS query logs and information of web content and WHOIS. We found that 153,163(17%) domain names in our dataset were wildcard domain names. Our important findings from the analysis include: 1) the proportion of abused domain names (i.e., domain names related to pornography and online gambling contents) in wildcard domain names is much higher than that in non-wildcard domain names (22.5% versus 4.4%); 2) abused wildcard domain names have remarkably higher risks in security than normal wildcard domain names. Then, based on the analysis, we proposed an effective algorithm named GSCS to detect abused

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16

wildcard domain names. GSCS first uses a domain graph to study the similarities of abused wildcard domain names' resolution behaviors, and then applies spectral clustering algorithm and seed domains to detect abused wildcard domain names. Experiments on real datasets indicate that GSCS can get about 86% detection rates with 5% seed domains, performing much better than BP algorithm. Future work will focus on further improving the detection rate by applying more machine learning methods with several datasets and more entries.

Acknowledgments

The work was supported in part by Scientific Research Foundation of the Institute of Information Engineering, Chinese Academy of Sciences (Grant No. Y6Z0011105 and J810091105).

References

1. Mockapetris, P.V.: Domain names-concepts and facilities. RFC 1034. (1987)
2. Du, K., Yang, H., Li, Z., Duan, H.-X., Zhang, K.: The Ever-Changing Labyrinth: A Large-Scale Analysis of Wildcard DNS Powered Blackhat SEO. In: USENIX Security Symposium, pp. 245-262. (2016)
3. Kalafut, A., Gupta, M., Rattadilok, P., Patel, P.: Surveying dns wildcard usage among the good, the bad, and the ugly. In: International Conference on Security and Privacy in Communication Systems, pp. 448-465. Springer, (2010)
4. Lewis, E.: The role of wildcards in the domain name system. RFC 4592. (2006)
5. Lee, J., Lee, H.: GMAD: Graph-based Malware Activity Detection by DNS traffic analysis. Computer Communications 49, 33-47 (2014)
6. Manadhata, P.K., Yadav, S., Rao, P.: Detecting malicious domains via graph inference. In: European Symposium on Research in Computer Security, pp. 1-18. Springer, (2014)
7. Berger, A., D'Alconzo, A., Gansterer, W.N., Pescapé, A.: Mining agile DNS traffic using graph analysis for cybercrime detection. Computer Networks 100, 28-44 (2016)
8. Xu, K., Wang, F., Gu, L.: Behavior analysis of internet traffic via bipartite graphs and one-mode projections. IEEE/ACM Transactions on Networking (TON) 22, 931-942 (2014)
9. Zou, F., Zhang, S., Rao, W., Yi, P.: Detecting malware based on DNS graph mining. International Journal of Distributed Sensor Networks 11, 102687 (2015)
10. Liu, D., Li, Z., Du, K., Wang, H., Liu, B., Duan, H.: Don't Let One Rotten Apple Spoil the Whole Barrel: Towards Automated Detection of Shadowed Domains. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 537-552. ACM, (2017)
11. Sharifnya, R., Abadi, M.: DFBotKiller: domain-flux botnet detection based on the history of group activities and failures in DNS traffic. Digital Investigation 12, 15-26 (2015)
12. Oprea, A., Li, Z., Yen, T.-F., Chin, S.H., Alrwais, S.: Detection of early-stage enterprise infection by mining large-scale log data. In: Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on, pp. 45-56. IEEE, (2015)

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22734-0_16