You Have More Abbreviations than You Know: A Study of AbbrevSquatting Abuse

Pin Lv[1,2], Jing Ya[1,2⋆], Tingwen Liu[1,2], Jinqiao Shi[1,2], Binxing Fang[1,2], and Zhaojun Gu[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing
[3] Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China, Tianjin
{lvpin, yajing, liutingwen, shijinqiao, fangbx}@iie.ac.cn
15620968007@163.com

Abstract. Domain squatting is a speculative behavior involving the registration of domain names that are trademarks belonging to popular companies, important organizations or other individuals, before the latters have a chance to register. This paper presents a specific and unconcerned type of domain squatting called "AbbrevSquatting", the phenomena that mainly happens on institutional websites. As institutional domain names are usually named with abbreviations (*i.e.*, short forms) of the full names or official titles of institutes, attackers can mine abbreviation patterns from existed pairs of abbreviations and full names, and register forged domain names with unofficial but meaningful abbreviations for a given institute. To measure the abuse of AbbrevSquatting, we first mine the common abbreviation patterns used in institutional domain names, and generate potential AbbrevSquatting domain names with a data set of authoritative domains. Then, we check the maliciousness of generated domains with a public API and seven different blacklists, and group the domains into several categories with crawled data. Through a series of manual and automated experiments, we discover that attackers have already been aware of the principles of AbbrevSquatting and are monetizing them in various unethical and illegal ways. Our results suggest that AbbrevSquatting is a real problem that requires more attentions from security communities and institutions' registrars.

Keywords: Domain Squatting, AbbrevSquatting, Institutional Domain Names, Abbreviations

1 Introduction

The Domain Name System (DNS) plays a critical role in supporting the Internet infrastructure by providing a distributed and fairly robust mechanism that resolves Internet host names into IP addresses. The reliability and agility

⋆ Corresponding author: Jing Ya (yajing@iie.ac.cn).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

that DNS offers has been fundamental to the effort for institutions, companies and organizations to scale information, business and service across the Internet. However, because of this, many attackers heavily rely on DNS to implement and scale their malicious operations.

In fact, domain squatting is a very common tactic used to facilitate DNS abuse by registering domains that are confusingly similar [1] to those belonging to popular companies, important organizations or other individuals. Domain squatting is hard to be eliminated. Because it involves the education of users' DNS interaction, rather than the technical correction of a protocol shortcoming, or a software vulnerability. There are several types of domain squatting techniques proposed in past researches. Typosquatting takes advantage of typographical errors [2–4]. Bit squatting utilizes accidental bit flips [5, 6]. Homograph-based squatting domains abuse the visual similarity of different characters [7, 8]. Homophone-based squatting domains abuse the pronunciation similarity of different words [9]. And, combosquatting combines a recognizable brand name with other common keywords [10].

In this paper, we present a specific and unconcerned type of domain squatting called "AbbrevSquatting", the phenomena that mainly happens on institutional websites. Institutional websites are created by associations, organizations or public institutes which aim to release official information and provide online services. In order to make users memorize them easily, such websites usually are bound to domains of abbreviated names that correspond to their full names or official titles (*i.e.*, using abbreviations of the names or titles). For example, the domain name '`cocc[.]net.cn`' is named after its official title '`China Ocean and Climate Change Information Network`'. And, the '`cocc`' in the domain name is the combination of the first letter of '`China Ocean and Climate Change`', which is part of the official title. While, we can also name it with '`coaccin`', which is the combination of the first letter of the official title. Obviously, there are other patterns of abbreviation. AbbrevSquatting takes advantage of the variety of abbreviations for a full name or official title and the users' confusion of which abbreviation represents the institute. They mine abbreviation patterns from existed pairs of abbreviations and full names, and register forged domain names with unofficial but meaningful abbreviations for a given institute. AbbrevSquatting is quite different from known domain squatting techniques. First, for a given institute, the Abbrevsquatting domain names are generated with its full name or official title, but not its official domain names. Second, the AbbrevSquatting domain names are generated with different types of abbreviation patterns but not slight changes on the input domain names.

To measure AbbrevSquatting abuse, we first analyse common abbreviation patterns used in institutional websites with a data set of one hundred thousands of institutional domains, and eight abbreviation patterns are minded. We generate 6,219,924 potential AbbrevSquatting domains with three popular abbreviation patterns, and find 1,370,014 (22.03%) of which are already registered. Then, we check the maliciousness of registered AbbrevSquatting domains with VirusTotal API and seven different blacklists, and group the domains into several

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

categories with crawled webpages and final links. Through a series of manual and automated experiments, we find that attackers have already been aware of the principles of AbbrevSquatting and are monetizing them in various unethical and illegal ways. AbbrevSquatting abuse is a real problem that security communities and institutions registrars should pay more attentions to.

Our main contributions in this paper are:

- In this paper, we present a specific and unconcerned type of domain squatting called "AbbrevSquatting". It mainly happens on institutional websites. Attackers mine the abbreviation patterns from existed pairs of abbreviations and full names, and register forged domain names with unofficial but meaningful abbreviations for a given institute.
- We analyze a data set of one hundred thousands of institutional domains, and mine eight abbreviation patterns (can cover up 89.27% of data set). We generate 6,219,924 potential AbbrevSquatting domains with three popular abbreviation patterns, and find 1,370,014 (22.03%) of which are already registered.
- Through a series of manual and automated experiments, we find that attackers have already been aware of the principles of AbbrevSquatting. Most of the generated domains are used to be parked, and some are listed in public blacklists. Our findings show that AbbrevSquatting is a real problem that requires more attentions from security communities and institutions' registrars.

The rest of this paper is structured as follows. In Section 2, we provide background information on institutional domain names and definition of AbbrevSquatting in general. Section 3 describes the analysis of our dataset and the way we generate potential AbbrevSquatting domains. We measure the abuse of AbbrevSquatting domain names in Section 4. Section 5 summarizes the related work. Finally, Section 6 concludes the paper's work.

2 Background

2.1 Institutional Domain Names

A domain name is a unique and easy-to-remember name that identifies and links to the address of a website on the internet. Domain names can generally be divided into two parts: second level domain and top level domain. Second level domain is the customisable part of the domain name that individuals, organisations or companies register to represent them on the internet. Top-level domains (also known as TLDs) are the next level of organisation on the internet. There are typically two kinds of TLDs, including Generic TLDs (gTLDs, *e.g.* '.com', '.net', '.org', '.edu', '.gov', *etc.*) and Country-code TLDs (ccTLDs, *e.g.* '.uk', '.cn', '.com.cn', '.net.cn', '.org.cn', '.edu.cn', '.gov.cn', *etc.*).

Institutional domain names are created and registered by associations, organizations or public institutes to release official information and provide online

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

services. They provide varieties of comprehensive and convenient platforms for institution administrators and Internet users to deal with public affairs online. In order to make Internet users remember them easily, the customisable parts (*i.e.*, second level domains) of such domain names are usually created and registered which correspond to their full names or official titles (*i.e.*, using abbreviations of the corresponding names).

For instance, the domain name '`cocc[.]net.cn`' links to the institutional website with official title of '`China Ocean and Climate Change Information Network`'. And, the second level domain '`cocc`' of the domain name is named with the combination of the first letter of '`China Ocean and Climate Change`', which is part of the official title.

2.2 AbbrevSquatting

For a given institute, we can create multiple abbreviations with its full name or official title. As for '`China Ocean and Climate Change Information Network`', the official domain name is '`cocc[.]net.cn`'. We can replace the '`cocc`' in the domain name with '`coaccin`', which is the first letter of all the words in the corresponding name. AbbrevSquatting takes advantage of the variety of abbreviation patterns for an institutional name and the users' confusion of which abbreviation represents the official website. The attack is based on abbreviations of domain names, *i.e.*, sets of abbreviations that are all coming from the same institute, but are named in different patterns.

AbbrevSquatting is quite different from other kinds of known domain squatting techniques mainly in two aspects. Firstly, for a given institute, the Abbrevsquatting domain names are generated with its full name or official title, but not its official domain names. Secondly, the AbbrevSquatting domain names are generated with different types of abbreviation patterns but not slight changes on the input domain names. Theoretically, AbbrevSquatting is much more difficult for Internet users to distinguish.

3 Measurement Methodology

Given the definition of AbbrevSquatting in Section 2.2, we provide a methodical way to measure AbbrevSquatting abuse using a dataset of one hundred thousands of institutional domains as the authoritative domains. First, we give a description of our data set, and mine the common abbreviation patterns they usually use. Then, we generate potential AbbrevSquatting domain names with three popular abbreviation patterns which are different from the official domains.

3.1 Data Set

The discovery of domain squatting activity requires a set of authoritative domains as targets. We obtain 134,806 Chinese institutional domain names from

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

Table 1. An example of data, 'CP' means 'Chinese Pinyin', 'EN' is 'English Words'.

Domain Name	`cocc[.]net.cn`
Full Name_CP	Guo Jia Hai Yang Xin Xi Zhong Xin
Full Name_EN	National Marine Information Center
OfficialTitle_CP	Zhong Guo Hai Yang Yu Qi Hou Bian Hua Xin Xi Wang
OfficialTitle_EN	China Ocean and Climate Change Information Network

Table 2. Percentages of TLDs used in authoritative domain list

TLD	**Percent (%)**	**TLD**	**Percent (%)**	**TLD**	**Percent (%)**
gov.cn	36.44	net	4.24	edu.cn	0.57
com	34.69	org	2.81	ac.cn	0.28
cn	12.50	org.cn	2.44	sh.cn	0.21
com.cn	4.48	net.cn	0.69	others	0.66

our cooperative partner as the authoritative domains. In our dataset, each domain name has a full name and an official title both in Chinese language. The full name is the name of a association, organization or institute, and the official title is the title of its institutional website. The two names may be the same. We use a Python package named Pinyin [4] and Baidu translation API [5] to extract the Chinese Pinyin and English words of each name or title. Table 1 shows an example item of our dataset used in this paper.

We further analyse the Top-Level Domains (TLDs) used in our dataset, as shown in Table 2. From Table 2, we can observe that TLDs used by institutional domain names are various and the common ones are '`.gov.cn`', '`.com`', '`.cn`', '`.com.cn`', '`.net`', '`.org`' and '`.org.cn`', which are more than one percent of all the domains. In the later generation process, we choose the seven most commonly used TLDs as the suffix of domain names.

3.2 Abbreviation Patterns Mining

To generate the potential AbbrevSquatting domain names, we also need a list of rules and models in addition to the authoritative domains. In this section, we mine the common abbreviation patterns used in the institutional domains.

Specifically, we mine the association relationships between the second level domains and full names (including four phrases as shown in Table 1) with strong rules. We finally extract eight rules (*i.e.*, abbreviation patterns) in the institutional domain names of our data set. The eight abbreviation patterns can cover up 89.27% of all the domains. The distribution of each pattern is shown in Table 3. We also give a manual analysis for the remained 10.73% domain names with unknown pattern, and find that they are not related to the corresponding full names or official titles at all.

The eight abbreviation patterns are defined as follows:

[4] https://pypi.python.org/pypi/pinyin
[5] http://fanyi-api.baidu.com/api/trans/product/index

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

Table 3. Abbreviation patterns used in the 134,786 Chinese institutional domain names

Pattern	Comment	Count	Percent
AFL	The first letter of all the words in a name	9366	6.95
PFL	The first letter of parts of the words in a name	56470	41.89
FLS	First Letters of several words in a name	15838	11.75
PWS	Parts of the words in a name	6378	4.73
CEC	Combination of English and Chinese Pinyin	8295	6.15
CSL	Contain sign '-' in the domain name	2612	1.94
CIR	Contain integers in the domain name	6045	4.48
SDN	Sub domains of the superior websites	15343	11.38
UNK	Unknown patterns	14459	10.73

Table 4. Common abbreviations of English words used in our dataset

Word	Abbreviate	Count	Word	Abbreviate	Count
education	edu	408	technology	te	65
school	sc	220	science	sc	63
chinese	chin	176	information	info	56
library	lib	170	agriculture	agri	48
small	sm	167	statistical	stat	47
agricultural	agri	126	taxatio	tax	46
tourism	tour	113	technology	tech	39
network	ne	109	network	net	30
center	ce	106	commerce	com	30
investment	invest	97	photography	photo	18
statistics	stat	86	company	co	15
cooperative	coop	68	geological	geo	14
institute	in	65	standardization	standard	9

AFL Pattern. In this pattern, a domain name is named with the first letter of all the words in a full name or official title. For example, '`tpeh`' in '`tpeh[.]net`' is named after the full name '`Tianjin Planning Exhibition Hall`'.

PFL Pattern. In this pattern, a domain name is named with the first letter of part of the words in a name. For example, '`cocc`' in '`cocc[.]net`' is named after the official title '`China Ocean and Climate Change Information Network`'.

FLS Pattern. In this pattern, a domain name uses first letters of several words in a full name or official title. For example, '`tianjinswim`' in '`tianjinswim[.]com`' is named after the full name '`Tianjin Swimming Center`'.

We further analyse the `FLS` abbreviation pattern in depth, and find that the condition that first few letters of a word used in Chinaes Pinyin usually happens in initial consonants, *i.e.*, 'zh', 'sh', 'ch'. As for the English words, we analyse some abbreviations for English words. The most commonly used abbreviations are as shown in Table 4.

PWS Pattern. In this pattern, a domain name is named with parts of the words in a full name or official title. For example, '`hanbofood[.]com`' is named after the full name '`Taiyuan Hanbo Food Industry Co Ltd`'.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

CEC Pattern. In this pattern, a domain name is named with the combination of English words and Chinese Pinyin. For example, '`nxzwnews`' in domain name '`nxzwnews[.]net`' is named after the Chinese name '`Ning Xia Zhong Wei Xin Xi Wang`' and English name '`Zhongwei News Network`'.

CSL Pattern and **CIR Pattern.** The two patterns contain sign '-' or integers in domain names. The details of the two patterns are complex. We will discuss them in our future work.

SDN Pattern. In this pattern, an institute uses a sub domain of its superior institute, such as '`czj.xlgl.gov.cn`', '`tjj.xlgl.gov.cn`'. As sub domain names are administrated by the main registered domains (*i.e.*, second level domains), we consider that AbbrevSquatting only exists in the second level domains.

3.3 Generating Domains

As we discuss in Section 2.1, a registered domain name includes two parts, *i.e.*, second level domain and top level domain. The top level domains we use in this paper are '`.gov.cn`', '`.com`', '`.cn`', '`.com.cn`', '`.net`', '`.org`' and '`.org.cn`', which are most commonly used in the institutional domain names of our data set. The second level domains are customisable, and generated by the abbreviation patterns of the full names or official titles.

In order to generate a controlled number of domain names and simultaneously measure AbbrevSquatting abuse effectively, we implement three generation methods with the most popular abbreviation patterns. The three methods are used to generate the customisable parts of the domains (*i.e.*, second level domains). And, the generation process is based on the four phrases of each institute as shown in Table 1.

Next, we give a detailed description of each generation method with the data in Table 1 as an example. From Table 1, we can observe that '`cocc`' in the domain name '`cocc[.]net.cn`' is named after the English official title '`China Ocean and Climate Change Information Network`' with the PFL pattern.

The first method is called "**ComAllMethod**". In this method, we generate the customisable parts of the potential AbbrevSquatting domains with a combination of the first letter of all the words in a phrase. For '`cocc`' in '`cocc[.]net.cn`', we can also name it with '`gjhyxxzx`', '`nmic`', '`zghyyqhbhxxw`', and '`coaccin`'.

The second method is called "**ComTopMethod**". In this method, we generate the customisable parts of the potential AbbrevSquatting domains with a combination of the first letter of the top n (*e.g.*, $n = 4, 5, 6$) words in each phrase. The length of the second level domain is limited between 4 and 6. The range is decided from the statistics of our data set. If the length of a phrase is less than 4, we handle it with the first method. For '`cocc`' in '`cocc[.]net.cn`', we can also name it with '`gjgy`', '`gjhyx`', and '`gjhyxx`' after the Chinese full name with this method.

The third method is called "**ComSegMethod**". The customisable parts of the potential AbbrevSquatting domains are generated based on word segmenta-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

Table 5. Profiles of the Generated Domain Names

Method	Generated	Registered	Percent(%)	HTMLs	Percent(%)
ComAll	1,858,230	179,591	9.66	96,135	53.53
ComTop	1,725,810	570,892	33.08	339,527	59.47
ComSeg	2,635,884	619,531	23.50	376,074	60.70
Total	**6,219,924**	**1,370,014**	**22.03**	**811,736**	**59.25**

tion. For the two Chinese phrases, we use a Python package named Jieba [6] to segment each phrase. For the two English phrases, we use the prepositions (*e.g.*, 'in', 'on', 'of', 'at' *etc.*) as delimiters to segment each phrase. For instance, the official title '`China Ocean and Climate Change Information Network`' can be segmented into '`China Ocean`', '`Climate Change Information Network`'. So, we can name it with '`co`', '`ccin`' and '`coccin`'. We set the length of the second level domain is less than 7 according to statistics.

We generate the customisable parts of domains with the above three methods. A potential AbbrevSquatting domain name is the combination of the customisable part and a suffix (*i.e.*, top level domain).

The profiles of our generated domain names are shown in Table 5. We totally generate 6,219,924 potential AbbrevSquatting domain names, targeting the 134,806 Chinese institutional domains in our data set.

In order to identify registered domain names, we perform a `whois` lookup for each domains. Then, we implement a crawler to visit the websites of the registered domain names to extract those provide web services. We also record the HTMLs and final URLs for further analysis. As shown in Table 5, we finally identify 1,370,014 domain names (22.03% of all the generated domain names) are already registered, and extract 811,736 (59.25% of all the registered domains) HTMLs. This paper focuses on the analysis of the domains which are registered and provide web services.

4 Measuring Results

In this section, we measure the AbbrevSquatting abuse through a series of automated and manual experiments. First, we check the maliciousness of the registered potential AbbrevSquatting domains with a public scanning API and seven different domain name blacklists. Second, we group the domain names into several categories according to the HTMLs and final URLs we crawled in Section 3.3.

4.1 Checking Maliciousness

To shed light on the malicious use of the registered potential AbbrevSquatting domain names, we check the generated domain names with a public scanning API and seven different domain name blacklists.

Firstly, we check the domains with a public API provided by VirusTotal [11]. VirusTotal is a website which aggregates many antivirus products and online

[6] https://pypi.python.org/pypi/jieba/

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

Table 6. Descriptions of categories

Category	Description
`Redirection`	Pages redirecting to another link
`Parked/For Sale`	Pages that have no content other than being advertised as for sale
`Entertainment`	Pages showing entertainment/gambling/lottery content
`Server Error`	Pages displaying an error, which caused by a server-side problem
`Adult Content`	Pages showing adult/pornographic content
`No Content`	Pages that have no content (*e.g.*, blank pages)
`Containing`	Pages containing legitimate content that happen to reside on a squatting variant of an authoritative domain
`Other`	Unclassified pages that do not fall into any of the above categories

scan engines, in addition to a myriad of tools to extract malicious signals from the input domains/urls/files. VirusTotal provides a public API that allows for automation of some of its online features. We get the scanned results of each domain through the public API. And, 2769 domains are found to be involved with virus or malicious activities.

Secondly, we check the generated domain names against seven different domain name blacklists [12–18]. The seven domain name blacklists come from malwaredomainlist.com, Ransomware Tracker, urlvir.com, abuse.ch's list of Zeus Tracker, nothink.org, joewein.de LLC, and malware domain blocklist by RiskAnalytics. The check is performed on the the second level domains, as AbbrevSquatting domains may choose different top level parts. We find that 2087 domain names have been public in the seven blacklists.

4.2 Categorization Results

With crawled data, we group the generated domain names into several categories. The crawled data includes a HTML and a final URLs for each domain. The final URL is used to detect redirection from the visited domain name to another different domain name. The HTML is a web page and contains the content of the website. We categorize each domain according to a full text analysis.

Specially, we follow a semi-automatic approach to implement the categorization. Firstly, we manually skim over the contents of a few pages and group together pages that with similar contents. The majority of these are parked pages, *i.e.*, pages that show ads, somewhat relevant to the domain name and usually also advertise that the domain may be for sale. Other groups are pages with little content, stating that the site is 'under construction', placeholder pages by popular registrars informing their clients how to setup a website on their registered domain, and pages containing generic errors, such as '404 Forbidden'. There are also websites with some normal content.

We summarize seven main categories according to the content of the websites. The descriptions of all the categories are shown in Table 6.

Next, we create generic content-signatures that could automatically categorize the remaining pages into each category. With this method, we can eventually

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

Table 7. Results of the categorization

Generated domains			Redirection domains		
Category	**Count**	**Percent**	**Category**	**Count**	**Percent**
Redirection	203751	**25.10%**	**Parked**	106479	52.26%
Parked	472163	58.17%	**AdultContent**	3617	1.78%
Entertainment	42995	5.30%	**Entertainment**	15572	7.64%
ServerError	114725	14.13%	**ServerError**	29382	14.42%
AdultContent	19427	2.39%	**Others**	48701	23.90%

automatically classify 85.98% of all the crawled webpages. The remaining unclassified domains are classified manually by a random sampling analysis.

By combining the results of the automatic classification and those of our manual investigation, we categorize all the potential AbbrevSquatting domains. The results of the categorization are shown in Table 7.

Parked/For Sale domains: Parked domains are the preferred monetizing way for domain squatters [19–21]. As we mentioned earlier, these domains contain no real content, except ads which are constructed on demand, usually by a domain-parking agency, based on the words included in a domain name and preferences by the owner of the domain. In total, parked/for sale domains represent the largest chunk of existing potential AbbrevSquatting domain names, with 471,526 cases (58.17% of all the webpages).

Redirection domains: While examining the AbbrevSquatting domains that redirect users to other different domains, we find that most of them are redirected to parked domains. We totally detect 203,751 redirection domains by checking the final URLs of each domain. While, 106,479 (52.26% of all the redirection domains) cases are parked domains. These domains are mainly redirected to large parked service agency websites, *e.g.*, `sedoparking.com`, `www.buydomains.com`, `cashparking.com` and so on. Redirection domains are also used in other categories, such as **Entertainment**, **Server Error**, **Adult Content**, and the distributions of each category are shown in Table 7. The left column shows the categories distribution for all the webpages. The right column shows the distribution of each category for all the redirection domain names.

We also find 152 websites with blank pages, which have no content. For the remaining unclassified pages, we randomly select 100 samples to analyze manually. We find that most of them contain legitimate content that happen to reside on a squatting variant of an authoritative domain.

5 Related Work

Domain squatting is a type of cybersquatting involving the registration of domain names that are trademarks belonging to other companies, institutions or individuals, before the latter have a chance to register [22, 23]. Several studies have been proposed and focused on domain squatting abuse in general.

Wang *et al.* [19] proposed models for the generation of typosquatting domains from authoritative ones. Janos *et al.* [2, 4] proposed techniques for identifying typosquatting. Agten *et al.* [3] studied typosquatting using crawled data over

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

a period of seven months and found out that few trademark owners protect themselves by defensively registering typosquatting domains. Apart from typosquatting, Nikiforakis *et al.* [6] quantified the extent to which attackers are leveraging bitsquatting, where random bit-errors occurring in the memory of commodity hardware can redirect Internet traffic to attacker-controlled domains. Their experiments show that new bitsquatting domains are registered daily and monetized through ads, affiliate programs and even malware installations. They later performed a measurement of another type of domain squatting called 'soundsquatting', where attackers abuse homophones to attract users and confuse text-to-speech systems [9].

As for AbbrevSquatting, the Chinese website '`xinhuanet.com`' ever reported some similar illegal behaviors [24]. But, to the best of our knowledge, this paper is the first one which deeply analyze the principles and measure the abuse of AbbrevSquatting. We mine abbreviation patterns from a data set of authoritative domains, and generate a large number of potential AbbrevSquatting domains. We measure the AbbrevSquatting abuse through a series of experiments.

6 Conclusion and Future Work

In this paper, we present a specific and unconcerned type of domain squatting technique, which is called "AbbrevSquatting". It mainly happens on institutional websites. Attackers mine the abbreviation patterns from existed pairs of abbreviations and full names, and register forged domain names with unofficial but meaningful abbreviations for a given institute. We analyze a data set of institutional domains, and mine eight abbreviation patterns (can cover up 89.27% of data set). We generate 6,219,924 potential AbbrevSquatting domains with three popular abbreviation patterns, and find 1,370,014 (22.03%) of which are already registered. Through a series of manual and automated experiments, we find that attackers have already been aware of the principles of AbbrevSquatting. Most of the generated domains are used to be parked domains, and some are listed in public blacklists. Our findings show that AbbrevSquatting is a real problem that requires more attentions from security communities and institutions' registrars.

We measure the abuse of the registered potential AbbrevSquatting domains which provide web services in this paper. In our future work, we would like to analyze the abuse of the potential AbbrevSquatting domains which do not provide web services. And, we also will analyze the changes of AbbrevSquatting domains with time.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No.2016YFB0801003 and the Open Project Foundation of Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China No.CAAC-ISECCA-201801.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17

References

1. Anticybersquatting Consumer Protection Act - Wikipedia . https://en.wikipedia.org/wiki/Anticybersquatting_Consumer_Protection_Act.
2. Szurdi Janos, Kocso Balazs, Cseh Gabor, Spring Jonathan, Felegyhazi Mark, and Kanich Chris. The Long "Taile" of Typosquatting Domain Names. In *Proc. of USENIX Security Symposium (USENIXSecurity)*, pages 191–206, 2014.
3. Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proc. of Network and Distributed System Security Symposium(NDSS)*, 2015.
4. Taha K. Mohammad, Xiang Huo, Zhou Li, and Chris Kanich. Every Second Counts: Quantifying the Negative Externalities of Cybercrime via Typosquatting. In *Proc. of IEEE Symposium on Security and Privacy*, 2015.
5. Artem Dinaburg. Bitsquatting: DNS Hijacking without Exploitation. In *Proc. of BlackHat Security*, 2011.
6. Nikiforakis Nick, Acker Steven, Van, Meert Wannes, Desmet Lieven, Piessens Frank, and Joosen Wouter. Bitsquatting: Exploiting bit-flips for fun, or profit? In *Proc. of International Conference on World Wide Web*, pages 989–998, 2013.
7. Gabrilovich Evgeniy and Gontmakher Alex. The homograph attack. *Communucations of the ACM*, 45(2)(128), 2002.
8. Tobias Holgers, David E. Watson, and Steven D Gribble. Cutting through the confusion: a measurement study of homograph attacks. In *Proc. of USENIX Annual Technical Conference*, pages 261–266, 2006.
9. Nick Nikiforakis, Marco Balduzzi, Lieven Desmet, Frank Piessens, and Wouter Joosen. Soundsquatting: Uncovering the use of homophones in domain squatting. In *Information Security*, pages 291–308, 2014.
10. Kintis Panagiotis, Miramirkhani Najmeh, Lever Charles, and Chen. et al Yizheng. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *Proc. of CCS*, pages 569–586, 2017.
11. VirusTotal. https://www.virustotal.com.
12. Malware Domain List. https://www.malwaredomainlist.com.
13. Ransomware Domain Blocklist. https://ransomwaretracker.abuse.ch.
14. Monitor Malicious Executable Urls. http://http://www.urlvir.com/export-hosts/.
15. ZeuS Tracker :: ZeuS blocklist. https://zeustracker.abuse.ch/blocklist.php?download=domainblocklist.
16. NoThink! http://www.nothink.org/blacklist/blacklist_malware_dns.txt.
17. joewein.de LLC. http://www.joewein.net/dl/bl/dom-bl.txt.
18. DNS-BH. http://www.malwaredomains.com.
19. Wang Yi-Min, Beck Doug, Wang Jeffrey, Verbowski Chad, and Daniels Brad. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. In *Proc. of SRUTI*, pages 31–36, 2006.
20. Moore Tyler and Edelman Benjamin. Measuring the perpetrators and funders of typosquatting. In *Proc. of Financial Cryptography and Data Security*, pages 175–190, 2010.
21. Vissers Thomas, Joosen Wouter, and Nikiforakis Nick. Parking sensors: Analyzing and detecting parked domains. In *Proc. of NDSS*, 2015.
22. B. Edelman. Large-scale registration of domains with typographical errors, 2003. http://cyber.harvard.edu/archived_content/people/edelman/typo-domains/.
23. S. E. Coull, A. M. White, T. f. Yen, F. Monrose, and M. K. Reiter. Understanding domain registration abuses. In *Proc. of IFIP SEC*, 2010.
24. Report. http://news.xinhuanet.com/politics/2015-07/23/c_1116010850.htm.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_17