Internationalized Domain Name Homograph Attacks

CSE 227: Computer Security - Spring 2017
University of California San Diego

Chen Lai
chl588@ucsd.edu

Zhongrong Jian
zhjian@ucsd.edu

J. Sidrach
jsidrach@ucsd.edu

Abstract

TODO

1 Introduction

TODO - Introduction to the problem - Enumeration of sections this paper talks about

Domain names were originally designed to support only ASCII characters. The proposal of Internationalized Domain Name IDN) benefits non-English speaking people, however, it also presents the potential vulnerability.

In this paper, we would analyze XXX top level .com domain by means of clustering possible homograph domains and classifying the owner of them. In section 2, we talked about the background and history of Internationalized Domain Name (IDN) and the vulnerability behind it. In section 3, several relevant researches are briefly introduced, which provide a couple of possible solutions to address IDN homograph attack. We will also present the IDN policies of current major browsers in this section. Section 4 describes the methodology of data collection and analysis we adopt in this project. In the next section, we comment on different analysis results and explain the consequences of caveats. For the last section, we briefly explain the ethical considerations of the research.

2 Background

TODO - DNS: explanation - IDN: history, explanation - Browsers: display of IDNs, algorithms

Domain name system contains the relationship between domain names with Internet Protocol(IP) address. It can can translate domain names to Internet Protocol(IP) address. At beginning, domain names support only ASCII characters. In 1996, Internationalized Domain Name (IDN) was proposed by Martin Drst for the purpose of letting non-English speaking people use Internet without additional restrictions. This extension involves representing Unicode characters in ASCII using Punycode, so that they could be then rendered back into their Unicode representation. Homograph letters (different letters whose representation is almost, if not, the same), however, present a potential vulnerability. For example, Cyrillic letter a can look identical to Latin letter a depending on the font. In other languages, like Chinese, there exists many homograph letters between traditional Chinese and simplified Chinese. Malicious attackers can then register a domain where one of the letters is actually Cyrillic but whose representation matches a Latin one. Users could be linked to this newly registered malicious page, and they may not have any visual indication (at least without further interaction) that the page is not the one they think they are visiting.

Currently, most popular website will show the hostname in Unicode depends on the language setting to defend IDN attack. For example, Chrome will display IDN in Unicode only if all characters of the domain

names belong to only one language and this language has to be the user's preferred language.

3 Related Work

TODO - Brief analysis of previous papers on the same topic

Many relevant studies have been conducted to show the vulnerability of IDN and provide possible solution to mitigate IDN homograph attack. In 2001, Evgeniy Gabrilovich and Alex Gontmakher has proved the feasibility of such attack based on the vulnerability of Unicode in *The Homograph Attack* [3]. In 2006, Vicktor Krammer proposed a defense mechanism based on address bar highlighting/indication technique and better UI of the browsers [4]. His method of defense mainly target at regular users who do not necessarily have relevant knowledge and through the better rendering of address along with explicit notification or alert on phishing URL users would be aware of malicious website. Manurer and Hfer believed that safe guard based on browsers' blacklist is far from enough to protect users from sophisticated phishing attacks [2]. They proposed an advanced method based on URL components and spell checking functionality of search engine. In this method, an URL would be divided into several components, namely, base name, sub-domains, path domain, brand name. Each of this components will be sent to search engine as a term for spelling check and in the cases that spelling suggestion is returned the URL will be counted as suspicious. The major browsers have their own IDN policies to defend against IDN homograph attack. Google Chrome, starting with Chrome 51, will display Punycode if certain tests fail on the input URL. These tests include but not limited to converting to Unicode, less than one numbering systems involved, no invisible characters, etc. IE displays URLs in IDN form if every component contains only characters from the languages configured in *Internet Options*. Firefox uses a script mixing detection algorithm based on *"Moderately Restrictive" profile of Unicode Technical Report 39* and displays Unicode when URL consists of any single script, any single script plus Latin, or it's in the white-list of other combinations. Opera shows IDN only for whitelisted TLDs. Safari has a whitelist of scripts that do not contain confusable characters and only shows the IDN form for the whitelisted scripts.

4 Methodology

Our data collection involves two primary sources. The first one is a snapshot of the `.com` and `.net` domain zone, from now on referenced as *.com snapshot*. This snapshot was provided by Verisign[1], and it is dated on 2017/05/01. It contains, mostly, the name servers of all `.com` and `.net` domains. For the purposes of this project, only the `.com` domains were considered.

The second data source is the *Alexa Top 1 million sites ranking*[2], from now on referenced as *top domains snapshot*. It contains the most popular one million web sites, regardless whether they are `.com` domains or not. It was also retrieved on 2017/05/01.

In an effort to help with the reproducibility and replicability of this project, the original data and processing code used in this project has been made available in a public repository[3].

4.1 Data Processing

We used the *.com snapshot* to obtain all the internationalized domains of the `.com` domain zone. All internationalized domain names are represented using Punycode, so they start with the `xn--` prefix. Using this information we first filtered the *.com snapshot* to match only the `.com` domains that start with `xn--`. Since the *.com snapshot* contains name server records, a single domain may have more than one entry in the file. Only the domain name is relevant for this project, so the rest of the columns of the *.com snapshot* were discarded. We sorted the domains in lexicographic order, removing duplicates in the process.

[1] `https://www.verisign.com/en_US/channel-resources/domain-registry-products/zone-file/index.xhtml`

[2] `https://www.alexa.com/topsites`

[3] `https://github.com/jsidrach/idn-homograph-attack`

We used the *top domains snapshot* as a source for the canonical (non-international) domain name of a website, based on its ranking. As we were only interested in non-internationalized domain names, so we discarded the ones that start with the prefix `xn--`. We also removed sub-domains, as we only had top level domain names in the *.com snapshot*.

4.2 Clustering

The underlying assumptions behind the clustering process are that homograph attacks are more likely to target popular domains, and that the million domains from the *top domains snapshot* contains most of the popular domains. To detect potential internationalized domain name homograph attacks, we cluster the internationalized domain names in the *.com snapshot*. The representative of each cluster is a homograph domain name from the *top domains snapshot*. Only clusters with more than one domain name are output.

The detection of homograph domain names is done using an algorithm to check if two strings are confusable. This algorithm is described in the *Unicode Technical Standard #39*[4]. The approach followed has some caveats, that could be addressed in future work. For instance, it does not detect a homograph of `www.google.com`, where the first "." has been replaced by a similar looking Unicode character. More sophisticated homographs can also be generated by using Unicode characters similar to "/". It is also worth noting that only domain names from the `.com` zone were considered, but similar studies could be done to other top level domains following the same procedure. All things considered, we still think this approach is a valuable first step.

Another type of clustering was also performed. We grouped the homograph domain names by its registrant organization, and ranked each organization by the number of homograph domain names to their name. This could shed light on which registrant organizations are allowing homograph domain names, or even on which individuals and companies are doing it the most.

[4] `http://www.unicode.org/reports/tr39/#Confusable_Detection`

4.3 Manual Classification

The last part of the data processing was the classification of the homograph domain names. This classification was done manually, to be able to differentiate between scam and unrelated websites. Some domains expired at the time of the classification, and were subsequently deleted from the output file.

The two high level categories that were defined are *Canonical* and *Third Party*. The first one, *Canonical*, is employed when the domain is registered by the same organization as its canonical homograph domain name. The second one, *Third Party*, is used when the domain is registered by a different organization than its canonical homograph domain name. Additionally, a more detailed classification was made:

- *Canonical - Parking*: domain is registered but not accessible via HTTP.
- *Canonical - Redirect*: domain redirects (HTTP Status Code 301/307/308) to its canonical homograph domain name.
- *Third Party - Redirect to Canonical:* domain redirects (HTTP Status Code 301/307/308) to its canonical homograph domain name.
- *Third Party - Unrelated*: domain resolves, but the contents of the website are totally unrelated to its canonical homograph domain name.
- *Third Party - Parking*: domain is registered, but no accessible via HTTP, or when accessed, a default domain parking web page is displayed.
- *Third Party - Scam*: domain resolves, and the website displayed is a clear attempt (similar color, logos, etc.) to make users think they are visiting the canonical homograph domain name.

5 Results

We obtained 458731 domain names from the *top domains snapshot* that belonged to the `.com` zone. From the *.com snapshot*, we identified a total of 1044301 internationalized domains. However, only a low percentage of the internationalized domain names (3.68%) had a homograph canonical domain name in the *top domains snap-*

Domains	#	%
Canonical domain names	*458731*	*8.31%*
With IDN homographs	825	6.04%
Without IDN homographs	457906	2.27%
Internationalized Domain Names	*1045400*	*91.69%*
With canonical homograph	1099	3.68%
Without canonical homograph	1044301	2.74%

Table 1: Overview of the clustering results.

shot. Even if we take into account that our homograph detection algorithm is rather limited (as described before), these results still indicate that the majority of internationalized domain names are not being used in homograph attacks. An overview of these clustering results can be found in Table 1.

Table 2 shows the number of internationalized domain name homographs for the top ten `.com` domains. It is worth noting that most of the homograph domains of *google.com* are actually registered by Google, possibly as a defensive measure. Another surprising result of the homograph clustering is that 81.87% of the groups have just one homograph internationalized domain name. This suggests that this kind of attack may not be targeted only to top websites, but rather to any domain where valuable information could be obtained from the users by a scam.

Domain	# of IDN homographs
google.com	24
youtube.com	3
facebook.com	9
baidu.com	3
yahoo.com	4
reddit.com	1
qq.com	2
taobao.com	1
live.com	1
vk.com	6

Table 2: Top ten .com domains in the Alexa ranking with IDN homographs.

The manual classification sheds some light on how these homograph internationalized domain names are being used. For the most part (82.34%), the domains are registered by a third party, and not being used actively (parking). One possible explanation for this is that the domain was bought in an attempt to be re-sold later at a higher price to the owner of the canonical homograph domain. A detailed breakdown of this classification can be found in Table 3.

Status	#	%
Canonical	*88*	*8.31%*
Parking	64	6.04%
Redirect	24	2.27%
Third Party	*971*	*91.69%*
Redirect to Canonical	39	3.68%
Unrelated	29	2.74%
Parking	872	82.34%
Scam	31	2.93%

Table 3: Breakdown of the manually classified homograph IDNs.

Table 4 contains the top ten registrants with the most homograph internationalized domain names. Most of these registrants are actually domain privacy companies, so it is impossible to know if the domains belong to the same person or not. There are also several individuals with a high number of homograph internationalized domain names. It is also worth mentioning that some of these individuals seem to be targeting specific industries. For instance, a single person has registered homographs of “audi”, “citroen” and “diesel”.

Registrant organization	Registrant email	# of homograph IDNs
Domains By Proxy, LLC	–	89
Super Privacy Service c/o Dynadot	privacy@dynadot.com	23
Domain Registries Foundation	–	22
Duong Thien	thiendv@outlook.com	18
Syngenuity Limited	manager@syngenuity.com	12
Helpnet: Brand Development & Sales	help@strongestbrands.com	12
ONUNO L.L.C.	corucas@gmail.com	11
Privacy Protection Service INC d/b/a	contact@privacyprotect.org	10
Hubertus Henz	hu_h5@yahoo.de	9
wuyu	wy65535@126.com	7

Table 4: Top ten registrants with the most homograph IDNs.

6 Ethical Considerations

TODO - Publicly available information - Delay between whois to not saturate server, also only in small subset (about 1k domains) - Brief explanation why this research is ethical

7 Conclusions

TODO - Conclusions of our work - Possible future work - TLDs

TODO DELETE [1]

Acknowledgments

We would like to thank Louis DeKoven and Stefan Savage for their help and support throughout this project.

References

[1] Internet prototocol, 1981.

[2] *Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity against Phishing* (Heidelberg, Berlin, 2012), vol. 7672 of *Lecture Notes in Computer Science*, Springer.

[3] Gabrilovich, Evgeniy. Gontmakher, A. The homograph attack. *Communications of the ACM 45*, 2 (2002), 128.

[4] Krammer, V. Phishing defense against idn address spoofing attacks. *PST '06*, 32 (2006).