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

The introduction of Internationalized Domain Names made access the web easier to non-English speakers. However, it also opened the possibility of using homograph domain names maliciously. We analyzed the potential impact of this type of attack, commonly referred to as IDN homograph attack. Our approach was based on clustering internationalized domain names with their counterparts in the Alexa Top 1m web sites ranking. The results obtained indicate that there are more than a thousand (at least) IDN homographs among *.com* domains. Most of them are not being used actively, but they still posses a potential security risk as there are no guarantees they will remain inactive in the future.

Keywords: Internationalized Domain Names, Homograph Attack, Phishing, Unicode, Punycode, Browser, Computer Security.

1 Introduction

Domain names were originally designed to only support ASCII characters. Internationalized Domain Names (IDNs) were proposed back in December 1996 by Martin Dürst[1] for the purpose of letting non-English speakers 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 security 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 could then register a domain where one of the letters is actually Cyrillic but whose representation matched 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.

In this paper, we analyze the potential impact of this type of attack, commonly referred to as IDN homograph attack. Section 2 explains past and present policies of major browsers and top-level domain (TLD) registrants in an attempt to prevent the attack. Section 3 explores relevant work in the literature related to homograph attacks. Section 4 describes the methodology of data collection and analysis adopted in this project. Section 5 comments the different results obtained. Section 6 addresses the ethical concerns regarding the data collection. Section 7 presents the conclusions obtained from this project.

2 Background

This section summarizes different mechanisms used to defend against internationalized domain name homograph attack. There are two mayor players trying to prevent this kind of attack, browsers and TLDs. Browsers, on the client side, display the non-Unicode version of the domain name based on different white-lists and restrictions. TLDs, on the other side, have different policies on domain name registrations to try to restrict the use of homographs. A more comprehensive study would be required for other possible attack vectors, such as email clients.

2.1 Browsers

All major browsers display the URL in Punycode instead of Unicode if certains conditions are met, to defend the user against a possible IDN homograph attack. The specific conditions, however, are different for each browser. Figure 1 shows a successful attack where the homograph domain name was not detected. Figure 2 show an unsuccessful attack, where the URL is displayed in Punycode.

*Author names are ordered lexicographically, by last name.

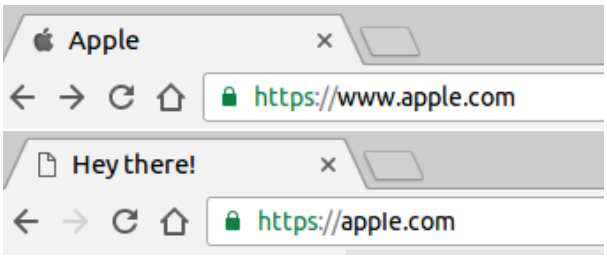

Figure 1: Apple's official website (top) and its IDN homograph (bottom), as displayed on Chrome 55.

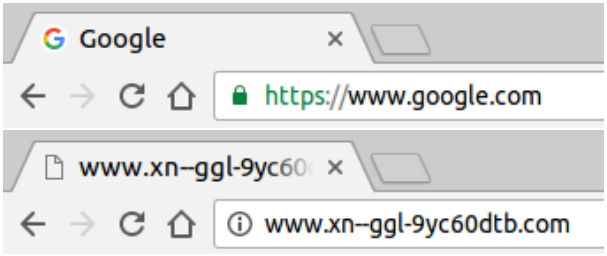

Figure 2: Google's official website (top) and its IDN homograph (bottom), as displayed on Chrome 55.

Firefox (version 54 at the time of writing) uses a script mixing detection algorithm based on the "Moderately Restrictive" profile defined in the *Unicode Technical Standard #39*[2]. It displays the URL in Unicode when all characters belong to a single script, to a single script plus Latin, or to a the white-list of other combinations. Google Chrome (version 59 at the time of writing), among other things, checks Latin-look-alike Cyrillic characters in non-IDN TLDs. This change was introduced recently, after a proof-of-concept IDN homograph attack working on Chrome was submitted to their issue tracker[3]. Internet Explorer and Edge (version 11 and 40 at the of writing, respectively) display the URL in Unicode if every component of the URL contains only characters from the languages configured in the operating system's *Internet Options*. Safari (version 10 at the time of writing) has a white-list of scripts that do not contain confusable characters, and only shows the URL in Unicode for those white-listed scripts.

2.2 Top-Level Domains

TODO - What does different TLDs do - ccTLDs

The Internet Corporation for Assigned Names and Numbers(ICANN) is responsible for the management of domain name system. In 2009, ICANN started to accept the application for internationalized country code top-level domain(IDN ccTLD) for the benefits of non-English speaking people and the first IDN ccTLD was added to DNS root zone in 2010**??**.

ICANN is aware of the potential vulnerability of homograph attacks and limits the usage of letter that is similar to Latin. For any application of potential IDN ccTLD, if there exists letter that is similar to an existing TLD, ICANN may reject them, especially for two-letter TLD. Three-letter TLD have higher chance to be accepted, as they are considered to be safer than two-letter TLD. ICANN listed detailed checking process in the Final Implementation Plan**??**.

For the second level domains, some registrant registers similar IDNs for brand protection, like Google. Google has registered most IDN in 2005.

3 Related Work

Several relevant studies have been conducted to describe the potential security issues with internationalized domain names, while providing possible solutions to mitigate the surface attack of these homograph domain names. The first of these studies, *The Homograph Attack* [3], was written by Evgeniy Gabrilovich and Alex Gontmakher in 2001, and proved the feasability of the attack.

In 2006, Vicktor Krammer proposed a defense mechanism based on address bar highlighting on the browsers and a better user interface [5]. This method of defense mainly targets regular users, by making them explicitly aware of the URL differences. It checks each character in a URL separately, highlighting digits or characters of various scripts in different background colors, so that users can easily tell the difference. It also suggests using larger font sizes in the address bar, to help distinguish possible homograph characters from their Latin counterparts.

In 2012, Maurer and Höfer argued that safe-guards based on browsers' blacklists was far from enough to protect users from sophisticated phishing attacks [1]. They proposed a more complex method based on URL components and the spell checking functionality of any search engine. In this method, a URL is be divided into several parts, namely, base name (the primary domain name as registered at the registrar), sub-domain (part before the primary domain, i.e., before the last identifier and TLD), path domain (sub-folder of the URL), and brand name (any popular brand name that is present in the URL). Their algorithm tries to prevent the attack by analyzing each of these URL parts. Each one is sent to the search engine for a spelling check, and in the cases that a spelling suggestion is returned, the URL will be marked as suspicious. By splitting the URL into these parts, this method aims to cover almost all possible cases of confusable URLs or domain names. However, this mechanism relies exclusively on the search engine to return spelling suggestions, and as a result, it is highly de-

pendent on the search engine and query itself.

While these last two methods were not adopted, they have influenced other possible mitigations currently implemented in most modern browsers. For instance, the base name of the URL is displayed in a darker color in Google Chrome and Mozilla Firefox. Safari goes a step further and only shows the base name of the URL.

Other attempts to analyze the impact of homograph attacks have also been made. In 2009, Peter Hanay and Christopher Bolan [4] concluded that, while most modern browsers in fact have implemented useful measures to prevent internationalized domain name homograph attacks, email clients where not nearly as effective, specially when dealing with incoming emails. In 2010, Johnny Al Helou and Scott Tilley [2] argued that the introduction of country-code Top Level Domains (ccTLDs) have made this attack easier to perform again.

Our approach to analyze the potential security risk of homograph internationalized domain names was fundamentally different the ones evaluated. Rather than inspecting traffic or generating random homograph permutations of a string, we took a more systematic approach. We clustered all internationalized homograph domains using a snapshot of the `.com` domain zone. We then used a list of top web sites and domain registrants data to classify each homograph domain name.

4 Methodology

Our data collection involved two primary sources. The first one was a snapshot of the `.com` and `.net` domain zone, from now on referenced as *.com snapshot*. This snapshot was provided by Verisign[4], and it is dated on 2017/05/01. It contains, for the most part, 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 was the *Alexa Top 1 million sites ranking*[5], from now on referenced as *top domains snapshot*. It contains the most popular one million web sites (as ranked by Alexa), 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[6].

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 lexicographical order, removing duplicates in the process.

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, we discarded the ones that start with the prefix `xn--`. We also removed subdomains, since 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 clustered the internationalized domain names in the *.com snapshot*. The representative of each cluster is a homograph domain name from the *top domains snapshot*, which we called canonical domain name. Only clusters with at least one internationalized homograph domain name were output.

The detection of homograph domain names was done using an algorithm to check if two strings are confusable. This algorithm is described in the *Unicode Technical Standard #39*. 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.

4.3 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. To help with the classification and speed up the process, we queried the *WHOIS* records[7] of every homograph domain name. Some domains had 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*, was employed when the domain is registered by the same organization as its canonical homograph domain name. The second one, *Third Party*, was 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 was registered but it was not accessible via HTTP.
- *Canonical - Redirect*: domain redirected (HTTP Status Code 301/307/308) to its canonical homograph domain name.
- *Third Party - Redirect to Canonical:* domain redirected (HTTP Status Code 301/307/308) to its canonical homograph domain name.
- *Third Party - Unrelated*: domain resolved, but the contents of the website were totally unrelated to its canonical homograph domain name.
- *Third Party - Parking*: domain was registered, but it was not accessible via HTTP, or when accessed, a default domain parking web page was displayed.
- *Third Party - Scam*: domain resolved, and the website displayed was a clear attempt (similar color, logos, etc.) to make users think they were 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 matching homograph canonical domain name in the *top domains snapshot*. 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.

The manual classification sheds some light on how these homograph internationalized domain names are currently 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 matching canonical homograph domain. However, the potential security risk is still present as there are no guarantees that the parked domains will remain inactive in the future. A detailed breakdown of this classification can be found in Table 3.

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”.

6 Ethical Considerations

The domain information used in this project is publicly available, and was voluntarily provided by the registrants of the domains when they registered them. The full `.com` domain zone snapshot will not be publicly disclosed, as its access requires a paid subscription. The retrieval of the *WHOIS* information was performed only for the homograph internationalized domain names, which were about one thousand. It was performed with a significant delay between petitions (3 seconds) in order to avoid saturating the servers.

7 Conclusions

TODO - Conclusions of our work - Possible future work - TLDs - Hard to balance (rendering idn useless vs preventing the attack)

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.

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.

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.

Acknowledgments

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

References

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

[2] AL HELOU, J., AND TILLEY, S. Multilingual web sites: Internationalized domain name homograph attacks. In *Web Systems Evolution (WSE), 2010 12th IEEE International Symposium on* (2010), IEEE, pp. 89–92.

[3] GABRILOVICH, EVGENIY. GONTMAKHER, A. The homograph attack. *Communications of the ACM 45*, 2 (2002), 128.

[4] HANNAY, P., AND BOLAN, C. Assessment of internationalised domain name homograph attack mitigation. In *Australian Information Security Management Conference* (2009), p. 13.

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

Notes

[1] `https://tools.ietf.org/html/draft-duerst-dns-i18n-00`

[2] `http://www.unicode.org/reports/tr39`

[3] `https://bugs.chromium.org/p/chromium/issues/detail?id=683314`

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

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

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

[7] `https://whois.icann.org`

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.