title.png

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 (IDNs) made it easier for non-English speakers to access the web. However, it also opened up the possibility of using homograph domain names maliciously, in what is called IDN homograph attacks. This kind of attack can be used to trick users into thinking they are visiting a different (but homograph) domain name. Our approach to analyze its potential impact is based on clustering IDNs with their canonical counterparts in the Alexa Top 1 million websites ranking. The results obtained indicate that there are more than a thousand IDN homographs among *.com* domains. While most of them are not being used actively, they still possess 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. 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 exist 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 scam 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 domain registrars 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 IDN homograph attacks. There are two major 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 certain conditions are met, to defend the user against a possible IDN homograph attack. The specific criteria, however, is different for each browser. Figure 1 shows a successful attack where the homograph domain name is not detected. Figure 2 shows an unsuccessful attack, where the URL is displayed in Punycode.

*Author names are ordered lexicographically, by last name.

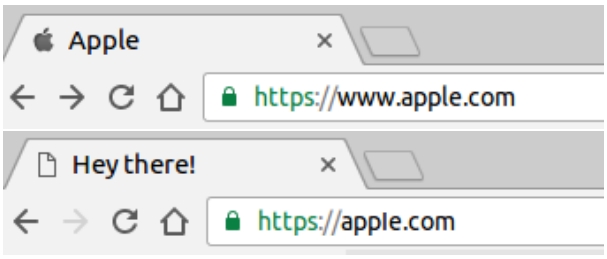

Figure 1: Apple's official website (top) and its IDN homograph (bottom), as displayed on Chrome 55.

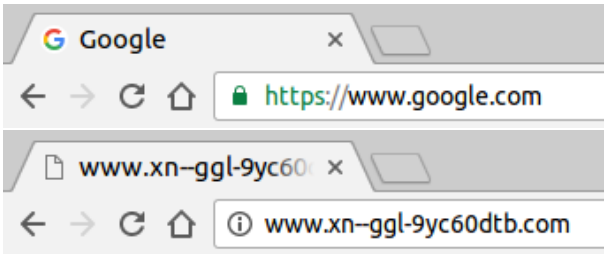

Figure 2: Google's official website (top) and its IDN homograph (bottom), as displayed on Chrome 55.

Firefox (version 54) uses a mixed script 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 white-list of other combinations. Google Chrome (version 59) performs several checks, including the search of Latin-look-alike Cyrillic characters in non-IDN TLDs. The check mentioned 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) 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) has a white-list of scripts that do not contain confusable characters, and only shows the URL in Unicode for those scripts in the white-list.

2.2 ICANN and TLD Registrars

The Internet Corporation for Assigned Names and Numbers (ICANN) is the organization responsible for the management of domain name system. In 2009, ICANN started to accept applications for internationalized country code top-level domains (IDN ccTLDs). The first IDN ccTLD was added to DNS root zone in 2010 [**?**]. ICANN is aware of the potential security risk that homograph attacks pose, and limits the use of Latin-look-alike letters. They may reject applications for IDN ccTLDs if there already exists a TLD that looks similar to the one being petitioned for. The detailed checking process is defined in their "Final Implementation Plan for IDN ccTLD Fast Track Process" [**?**].

Most second level registrars do not have an explicit public policy to prevent the registration of homograph domains. Only a few have implemented restrictions to eliminate this potential risk. For instance, `.cn`, `.biz`, `.hk`, and all Chinese TLDs will simultaneously activate both traditional and simplified Chinese IDNs once one of them is registered.

3 Related Work

Several relevant studies have been conducted to describe the potential security issues with IDNs, while providing possible solutions to mitigate the surface attack of these homograph domain names. The first study, *The Homograph Attack*, 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 [**?**]. 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 [**?**]. 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 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 spell 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, resulting in its highly dependence on the search engine and query itself.

While these last two methods were not adopted, they

have influenced other defense mechanisms 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 concluded that, while most modern browsers in fact have implemented useful measures to prevent IDN homograph attacks, email clients were not nearly as effective, specially when dealing with incoming emails [**?**]. In 2010, Johnny Al Helou and Scott Tilley argued that the introduction of ccTLDs have made this attack easier to perform again [**?**].

Our approach to analyze the potential security risk of homograph IDNs is fundamentally different from the ones evaluated. Rather than inspecting traffic or generating random homograph permutations of a string, we take a more systematic approach. We cluster all internationalized homograph domains using a snapshot of the `.com` domain zone. We then use a list of top websites and domain registrants data to classify each homograph domain name.

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 has been 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 are considered.

The second data source is the *Alexa Top 1 million sites ranking*[5], from now on referenced as *top domains snapshot*. It contains the most popular one million websites (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 is available in a public repository[6].

4.1 Data Processing

We use the *.com snapshot* to obtain all the IDNs of the `.com` domain zone. All IDNs are represented using Punycode, so they start with the `xn--` prefix. Using this information we first filter 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 to this project, so the rest of the columns of the *.com snapshot* can be discarded. We sort the domains in lexicographical order, removing duplicates in the process.

We use the *top domains snapshot* as a source for the canonical (non-international) domain name of a website. As we are only interested in non-IDNs, we can discard the ones that start with the prefix `xn--`. We also remove sub-domains, since we only have second 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 IDN homograph attacks, we cluster the IDNs in the *.com snapshot*. The representative of each cluster is a homograph domain name from the *top domains snapshot*, which we call canonical domain name. Only clusters with at least one internationalized homograph 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*[2]. 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 "." is 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 are 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 is also performed. We group the homograph domain names by its registrant organization, and rank each organization by the number of homograph domain names they have registered. 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 is the classification of the homograph domain names. This classification is done manually, to differentiate between scams and unrelated websites. To help with the classification and speed up the process, we query the *WHOIS* records[7] of every homograph domain name. Some domains are expired at the time of the classification, and are subsequently deleted from the output file.

The two high level categories 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 is made:

- *Canonical - Parking*: domain is registered but it is 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 it is not 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 obtain 458731 domain names from the *top domains snapshot* that belong to the `.com` zone. From the *.com snapshot*, we identify a total of 1044301 IDNs. However, only a low percentage of the IDNs (3.68%) have 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 IDNs 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 IDN 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 IDN. 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 scammed from users.

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.

The manual classification sheds some light on how these homograph IDNs 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

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	88.01%

Table 1: Overview of the clustering results.

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.

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 IDNs. 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 IDNs. 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 we use in this project is publicly available, and is voluntarily provided by the registrants of the domains when they register 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 is done only for the homograph IDNs, which are about one thousand. It is performed with a significant delay between petitions (3 seconds) in order to avoid saturating the servers.

7 Conclusions

We provide a systematic method to evaluate the potential threat that IDN homographs present. Our data shows that there are more than a thousand IDN homographs in the *.com* zone. Even if the vast majority these are parking domains, there are no guarantees that they will remain inactive in the future. Surprisingly, most TLDs do not have common policies to prevent the registration of homograph domains.

It is also important to remember that while the easiest solution would be displaying always URLs in Punycode instead of Unicode, that would completely defeat the purpose of IDNs. Browsers have to balance user experience and security. They need to allow users to type and read domain names in their language of choice; while at the same time they should avoid misleading users into thinking they are in a legitimate domain when they are not.

Regarding future work, the next logical steps would be to repeat this analysis in other TLDs, and to improve the homograph matching algorithm. A more challenging enhancement would be to perform the classification of homograph IDNs without manual intervention by analyzing *WHOIS* records and the *HTTP/HTML* responses.

Acknowledgments

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

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