Cutting through the Confusion: A Measurement Study of Homograph Attacks

Tobias Holgers, David E. Watson, and Steven D. Gribble
Department of Computer Science & Engineering
University of Washington

1 Introduction

Domain names are crucial to the usability of the Web, but the same characteristics that make them useful to people also make them vulnerable to attack. When a user follows a hyperlink, the domain name within the URL provides her with the first and most important indication of the identity of the organization with which she will interact. If the user is fooled into misreading a domain name, she will believe she is interacting with one organization, but she might actually be interacting with an attacker. By spoofing the content of the user's intended destination, the attacker might trick the user into revealing sensitive information. In this scenario, SSL is no help to the victim, since the attacker could obtain a valid certificate for the confused domain name.

A *homograph attack* is one technique for carrying out this scheme. A homograph is a letter or string that is visually confusable with a different letter or string. For example, using most sans-serif fonts, the Latin letter l (lower case 'el') is visually confusable with the Latin letter I (upper case 'eye'). Rendered with such a font, the following are *confusable*, if not indistinguishable:

http://www.paypal.com vs. http://www.paypal.com

An attacker who registers the confusable domain name paypai.com therefore may be able to lure victims to their site, for example by sending spam that appears to contain a hyperlink to the *authoritative* PayPal site.

Web homograph attacks have existed for some time, and the recent adoption of International Domain Names (IDNs) support by browsers and DNS registrars has exacerbated the problem [Gabr02]. Many international letters have similar glyphs, such as the Cyrillic letter р (lower case 'er,' Unicode 0x0440) and the Latin letter p. Because of the large potential for misuse of IDNs, browser vendors, policy advocates, and researchers have been exploring techniques for mitigating homograph attacks [Mozi05, Appl05, Oper05, Mark05].

There has been plenty of attention on the problem recently, but we are not aware of any data that quantifies the degree to which Web homograph attacks are currently taking place. In this paper, we use a combination of passive network tracing and active DNS probing to measure several aspects of Web homographs. Our main findings are four-fold.

First, many authoritative Web sites that users visit have several confusable domain names registered. Popular Web sites are much more likely to have such confusable domains registered. Second, registered confusable domain names tend to consist of single character substitutions from their authoritative domains, though we saw instances of five-character substitutions. Most confusables currently use Latin character homographs, but we did find a non-trivial number of IDN homographs. Third, Web sites associated with non-authoritative confusable domains most commonly show users advertisements. Less common functions include redirecting victims to competitor sites and spoofing the content of authoritative site. Fourth, during our nine-day trace, none of the 828 Web clients we observed visited a non-authoritative confusable Web site.

Overall, our measurement results suggest that homograph attacks currently are rare and not severe in nature. However, given the recent increases in phishing incidents, homograph attacks seem like an attractive future method for attackers to lure users to spoofed sites.

2 Homographs and Confusability

As previously mentioned, a homograph is a letter or string that has enough of a visual similarity to a different letter or string that the two may be confused for one another. The precise degree of similarity necessary to cause confusion is difficult to quantify, as it depends on the observer, the fonts and font sizes used, and the context in which the homograph is observed.

There are many different categories of confusable characters. They may be drawn from the same script, such as the Latin characters '-' (hyphen) and '–' (en dash). Different scripts may be involved, such as with the Latin character a and the Cyrillic character а (small letter a). Font choices can affect confusability; the Latin characters 'rn,' if rendered with a sans-serif font appear as 'rn' and can be confused with the Latin character 'm.' Two characters with very different glyphs may appear to be identical if a browser does not have support for one of them. For example, an ä ('a' with an umlaut) might be rendered without the umlaut.

Further compounding the problem is the fact that confusable characters do not need to be used when constructing confusable strings. The word *recieve* may be confused with *receive*, and even more complicated misspellings may be overlooked by a causal observer. Given all of this complexity, in this paper we do not attempt to establish perceptual thresholds of confusability and rigorously examine all possible confusable characters. Instead, we make the simplifying assumption that two characters are confusable if and only if they are listed as confusable in the Unicode Technical Report on security considerations [Davi05].

This assumption gives us only a rough approximation to the real world notion of confusability, however, as we will show in Section 3, many registered domain names do have confusable domains registered consisting of character substitutions. In most cases, these confusable domains do not have a legitimate purpose.

With this assumption in place, we can operationally define the confusability of two strings: one string is confusable with another string if and only if they are related by some number of confusable character substitutions. As an example, consider the following string, which contains four character substitutions:

Microsoft Corporation

The underlined characters are Cyrillic confusables of their Latin character counterparts. For this particular string, the set of all confusable strings related to it is enormous (32,459,975,614,080), since most of the characters in the string have at least one confusable character associated with them, and we must consider all possible one, two, three, ..., twenty-one character substitutions.

Increasing the number of confusable character substitutions in a string tends to make the string less confusable. Accordingly, in practice confusable strings tend to contain only one or two subtitutions, though as we will show in Section 3, some popular domains have registered confusables with up to five character substitutions.

In this paper, we examine a simple kind of homograph attack, in which an attacker registers a domain name that is confusable with some other domain name, presumably to lure victims to their site. In principle, two registered domains may both be associated with legitimate organizations, yet still be confusable with each other. In practice, a given set of confusable domain names tends to consist of a single *authoritative* domain, and a collection of non-authoritative, illegitimate domains. Though authoritativeness is a subjectively defined characteristic, we have found in all cases it is simple to distinguish between the authoritative domain that people intend to visit, and the non-authoritative confusables that attackers create.

3 Measurement Study

We gathered a nine-day-long trace of the Web activity generated by the population of clients in the Department of Computer Science and Engineering at the University of Washington. The department consists of approximately 40 faculty, 40 staff, 275 graduate students, and 450 undergraduate students.

There is a mixture of static IP assignment and DHCP usage in the department, but the majority of hosts that rely on DHCP receive the same IP address in practice. Accordingly, the number of IP addresses we observed in the trace, 828, is a reasonable (though not perfect) estimate of the number of hosts that were active during the trace period.

We installed a passive network tap on the router connecting the departmental subnets to the campus backbone. This tap allowed us to observe all packets flowing between department computers and external hosts. The peak traffic rate through the router was low enough that our network monitoring host dropped no packets.

Using Snort, we collected a trace consisting of all outbound HTTP GET requests. We post-processed the trace to extract the domain name associated with each request. To perform this extraction, we looked in the "Host" HTTP header field; this field is required in HTTP/1.1, and is generated by all modern browsers. Using this field saved us from having to perform reverse DNS lookups, and it also allowed us to disambiguate between multiple domains hosted on the same IP address.

Given this list of domain names, we calculated the popularity of a domain name by counting the number of GET requests directed to it. To transform our object-related popularity measure into an approximate page-relative popularity measure, we excluded requests for image data types, since otherwise a single page containing many embedded images would have a higher contribution to domain popularity than a single page containing few embedded images.

It is clear that the Web activity of a computer science department is not wholly representative of Internet-wide Web activity. However, the set of *popular* Web sites within the departmental trace has a substantial overlap with the set of top 500 global Web properties listed by Alexa Internet [Alex05]: 31 of the top 50 domains in the Alexa list appeared in our trace. As we will show in Section 3.2.2, popular Web sites are more likely to have confusable domain names registered.

3.1 Active DNS probing

Once we obtained the list of domain names from the departmental trace, our next step was to search for registered confusable domain names associated with each

rank	authoritative domain name	# possible confusables	# registered confusables	confusable names (confusable characters underlined, IDN punycode in parenthesis)
1	yahoo.com	5,202	2	yahoo.com (xn--yhoo-53d.com), yah0o.com
2	msn.com	12	1	msn.com (xn--mn-eoc.com)
3	google.com	1,156	4	g0ogle.com, go0gle.com, g0og1e.com, go0g1e.com
6	passport.net	19,584	1	passp0rt.net
8	ebay.com	252	2	ebay.com (xn--bay-qdd.com), ebay.com (xn--by-7kcs.com)
11	microsoft.com	48,552	5	microsoft.com (xn--micrsoft-qbh.com), microsoft.com (xn--microsft-sbh.com), microsoft.com (xn--micrsft-djgb.com), microsoft.com (xn--mrft-65das6nf.com), micros0ft.com
12	amazon.com	3,672	1	amazon.com (xn--amazn-mye.com)
18	fastclick.com	1,344	0	
20	aol.com	204	2	aol.com (xn--al-jbc.com), aol.com (xn--al-fmc.com)
22	go.com	17	0	
102	bankofamerica.com	25,909,632	1	bankofamerica.com (xn--bnkofamerica-x9j.com)
980	paypal.com	3,456	4	paypal.com (xn--pypal-4ve.com), paypal.com (xn--papal-fze.com), paypal.com (xn--paypl-7ve.com), paypal.com (xn--pyal-53d1h.com)

Table 1: **Registered confusables for popular domains.** This table lists the registered confusable domains for the 10 most popular English language Web sites within the Alexa 500 list, as well as two financial sites.

one. To accomplish this, for each traced name, we generated confusable names by substituting one or more characters with corresponding confusable characters. Then, we performed a DNS lookup on each generated name to test whether it was actually registered.

There is a combinatorial explosion in the number of confusable names associated with a given string when performing multiple character substitutions. Because of this, we limited our search to confusable names with at most three confusable characters. However, to explore the degree to which this caused us to miss registered confusables with a greater number of substitutions, we performed an exhaustive search of the full space for a few of the traced domains for which we found the most registered confusable names.

Since the department trace may be biased towards university and research topics, we conducted a similar evaluation using the list of the Top 500 most popular domain names, according to Alexa [Alex05]. The Alexa list contains domains ordered by a "traffic rank." This metric is the geometric mean of reach (percent of Internet users visiting the site) and page views (percentage of all daily global page views).

3.2 Results

In Table 2, we show high-level results from our study. We observed 828 clients accessing 3,425 different Web server domain names, issuing a total of 452,654 HTTP GET requests. Web sites visited in our trace were authoritative: no client ever visited a Web site with a non-authoritative, confusable domain name. However, our DNS probing found 399 registered domains whose names are confusable with authoritative Web domains visited by our users. Looking at this data another way, 298 authoritative Web domains have one or more non-authoritative, confusable, registered domains. None of our users appeared to have fallen victim to a homograph attack during our trace period, even though the potential for such an attack does exist.

trace period	June 5 -- June 13, 2005
# client IPs	828
# GET requests	452,654
# server IPs	4,991
# distinct server hostnames	3,425
# non-authoritative confusable domains visited by users	0
# non-authoritative registered confusable domains found during DNS probing	UW trace: 237
	Alexa top 500 list: 162
# authoritative sites that have non-authoritative, registered confusables	UW trace: 182
	Alexa top 500 list: 116

Table 2: **Overall results.** This table provides summary statistics describing our trace.

For those authoritative domains that had confusable domains registered, we typically found a very small number of registered confusable names. Even though a large number of confusable names are possible for a given authoritative domain name, there are usually just a handful of confusable domains registered.

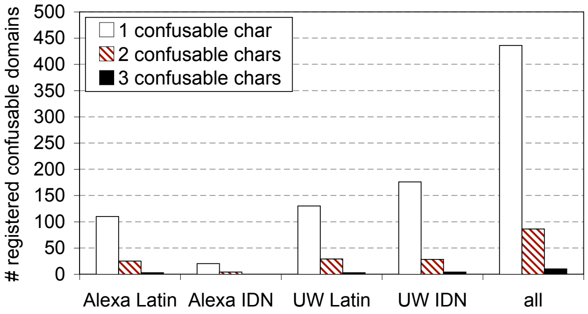

Figure 1: **# confusable character substitutions.** This graph shows how many registered confusables have one, two, or three confusable character substitutions.

In Table 1, we show a list of registered confusable domains found for the top 10 most popular English language Web sites within the Alexa 500 list, as well as two financial sites. Note that this table only reports on registered DNS names with three or fewer confusable character substitutions, as previously described in Section 3.1.

3.2.1 Number of character substitutions

Intuitively, one should expect that registered confusable domain names will tend to consist of a small number of confusable character substitutions. Each confusable character may not always render identically to the intended character. Accordingly, while one confusable character in a confusable domain name may escape notice, two or three such characters may not.

Figure 1 shows that most registered confusable domain names only contain a single confusable character, suggesting this intuition is correct. As well, this data validates our choice of limiting the search space of our DNS probes to names with no more than three character substitutions: less than 3% of confusable names we found had three substitutions.

To further validate this choice, we performed an exhaustive search for confusables using the two domain names with the most registered confusables, microsoft.com and paypal.com. This full search of all 48,552 possible microsoft confusables and 3,456 paypal confusables found only one confusable domain that our limited search missed: a microsoft.com confusable with five confusable character substitutions.

3.2.2 Popularity and registered confusables

Figure 2 shows, for an authoritative site of a given popularity rank, the fraction of all registered confusable names found that are associated with authoritative sites of equal or greater popularity. As well, the figure includes a logarithmic curve fit for the "UW IDN" data series. The graphs show that popular authoritative sites have more registered confusable names than unpopular authoritative sites.

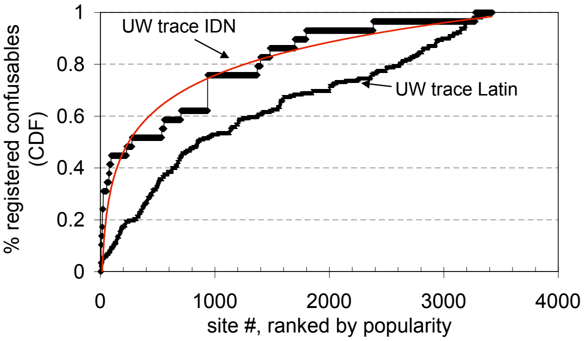

Figure 2: **Popularity vs. registered confusables.** This CDF shows, for a site of a given popularity, the fraction of registered confusable names found that are associated with authoritative sites of equal or greater popularity. Popular sites have more registered confusable names.

If registered confusable domain names were uniformly distributed across authoritative sites, these lines would have a constant slope. Instead, we see that for both UW IDN and UW Latin confusables, popular authoritative sites have more confusable names registered for them than unpopular authoritative sites. This effect is most striking for IDN confusables; 80% of registered IDN confusables found are associated with the top 30% of authoritative sites. The effect is less striking for Latin confusables, but we hypothesize that the effect would reveal itself more prominently with a longer trace that would include additional unpopular domains.

3.2.3 Latin vs. IDN names

Our search for registered confusable domain names included domains consisting entirely of Latin character substitutions, and IDN domains that included some Unicode character substitutions. In Table 3, we show how many of each of these exist for both the Alexa 500 list and domains visited in the UW trace.

Our results show that most registered confusable domains consist entirely of Latin characters: IDN confusable domains containing Unicode characters account for only 15% and 12% of the Alexa and UW lists, respectively. While a relatively small fraction, IDN confusable domains do have a noticeable presence, and they can be expected to grow as browser support for IDN increases. For example, the upcoming Microsoft Internet Explorer version 7 browser is expected to have IDN support, making confusable Unicode domain names potentially more attractive to attackers.

	Latin	IDN	total
Alexa Top 500	138 (85%)	24 (15%)	162
UW Trace	208 (88%)	29 (12%)	237

Table 3: **Latin vs. Unicode confusables.** This table shows the number of registered confusable domains found that contain only Latin confusable characters, and the number of IDN domains that contain some Unicode confusable characters.

3.2.4 The intent behind confusable domains

Our data shows that many non-authoritative, confusable domain names have been registered. We now turn our attention to understanding what goal attackers had when registering them. Homographs can be used to construct elaborate Web spoofing or phishing attacks, in which the victim is fooled into revealing sensitive information. However, attackers may have other less dangerous goals in mind, such as attracting victims to a site in order to display advertisements.

To understand the attacker's intent behind a confusable domain, and to gauge the current risk that homograph attacks pose, we manually examined all non-authoritative confusable domains that we found registered. Based on our examination, we categorized each site into one of the following seven categories in decreasing order of (subjectively assigned) risk to the victim:

- **Web spoofing:** the confusable site spoofs the content of the authoritative site.
- **Redirect to competitor:** the victim is redirected to a commercial competitor of the authoritative site.
- **Advertisement:** ads are shown to the victim.
- **For sale:** the registered confusable domain name is advertised as being for sale.
- **Unrelated:** the site has content which is unrelated to the authoritative site.
- **No content:** the registered confusable domain name does not have an active Web server, or the server returns blank pages.
- **Redirect to authoritative:** the victim is redirected to a the authoritative site, perhaps as a defensive measure put in place by the authoritative site itself.

A given site may belong in more than one category, such as a site that is for sale and also shows ads. We attempted to emphasize the more subtle, and thus potentially more dangerous, uses of homographs and thus categorized each site only in its highest risk category.

Intent	% of registered confusable domains		
	Alexa 500 list	UW trace	union
Web spoofing	0.6%	1.7%	1.3%
Redirect to competitor	2.5%	2.2%	2.3%
Advertisement	43.2%	45.5%	44.5%
For sale	16.7%	14.3%	13.0%
Unrelated	13.6%	10.8%	12.0%
No content	19.1%	18.2%	20.6%
Redirect to authoritative	4.3%	7.4%	6.3%

Table 4: **Intent of registered confusables.** This table shows the fraction of registered confusable domains that were observed to have the listed intent.

Table 4 summarizes the results. Advertising, a relatively benign function, was overwhelmingly the most popular use for confusable domain names. There were very few spoofed sites among registered domains we observed. Additionally, we verified that none of these spoofed sites attempted to trick the user into submitting sensitive information. Instead, these spoofed sites either consisted of parodies of the authoritative site, or they served to warn potential victims about the dangers of homograph attacks.

4 Related work

Web spoofing attacks were first considered by [Felt97]. [Gabr02] first discussed using homographs as a part of a web spoofing attack. Early versions of the attack relied on similarities between Latin letters and numbers. For example, an attacker could register an address where `o` is replaced by `0` (zero), or `l` with `1` (one).

With the introduction of International Domain Names (IDN) the number of visually confusable characters has increased dramatically. IDN attacks have been possible in Mozilla [Mozi05], Safari [Appl05] and Opera [Oper05] for at least one publicly available release, though the latest versions have adopted some defensive mechanisms. Browser-based solutions to the homograph problem are currently incomplete, however, as they either rely on trusted registrars or disable significant portions of the IDN namespace.

Registrars issuing IDN domains have been asked to put in place policies to prevent two homographic domains from being registered to different sites [Mark05]. Relying on registrars to help solve the problem has disadvantages, since registrars must contend with multiple jurisdictions and potentially conflicting regulatory restrictions. However, this approach is compatible with other solutions to the Web spoofing problem. For example, trust bars [Herz04], the eBay Toolbar [eBay], and Spoof-

Guard [Chou04] give users immediate and unforgeable security context information.

[Goth05] evaluates the current rate and cost of phishing scams, and concludes that while the cost has been reduced in recent years, it is still costing billions of dollars. [Weny05] discusses using Web crawlers to look for visually similar Web pages. Others researchers in the usability, cryptography, and anti-phishing communities have proposed several mechanisms to defend against phishing attacks. For example, Jakobsson [Jako05] proposes an economic analysis to quantify the risks of an attack and to develop methods for defending against them. As another example, Adida et al. propose the adoption of identity-based ring signatures to provide digitally signed email to eliminate spam-based phishing attacks [Adid05]. Dhamija and Tygar propose the concept of "security skins," a browser extension that allows remote sites to prove its identity to users in a way that is usable but hard for attackers to spoof.

5 Conclusions

While visually confusable, non-authoritative domains have been registered in practice, the threat actually posed by these domains currently does not live up to the potential feared by the community [Oper05, Mozi05, Appl05]. Many popular Web sites do have associated confusable domains registered, but the most common functions of these confusable domains are benign, such as serving advertisements. However, as support for IDN names grows, homograph attacks do have the potential to become more common and malicious.

Overall, our results show that: (1) users often visited sites that have confusable domains registered, but no user visited one of these non-authoritative domains during our trace; (2) popular sites are much more likely to have registered non-authoritative confusable domains than unpopular sites; (3) confusable domains tend to have a single confusable character within them, and currently only 12-15% of confusable domains rely on Unicode confusable characters; and (4) most confusable domains have relatively benign intent, such as showing advertisements. Though a small fraction do spoof the authoritative site, even these spoofed sites appear to have relatively benign intent, such as parody.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-0430477 and ANI-0132817, by an Alfred P. Sloan Foundation Fellowship, and by gifts from Intel Corporation and Nortel Networks.

References

[Adid05] Ben Adida, Susan Hohenberger and Ronald L. Rivest, *Separable Identity-based Ring Signatures: Theoretical Foundations for Fighting Phishing Attacks.* DIMACS Workshop on Theft in E-Commerce, Piscataway, New Jersey, April 2005.

[Alex05] Alexa Web Search, Alexa Internet Inc., *Global Top 500 Sites*, June 7 2005.

[Appl05] Anonymous, *About Safari International Domain Name support*, Apple Computer Inc., March 2005, http://docs.info.apple.com/article.html?artnum=301116

[Chou04] Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh and John C. Mitchell, *Client-side defense against web-based identity theft*. Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS '04), San Diego, CA, February 2004.

[Davi05] Mark Davis, Draft Unicode Technical Report #36, *Security Considerations in the Implementation of Unicode and Related Technology*, February 20, 2005.

[Dham05] Rachna Dhamija and J.D. Tygar, *The Battle Against Phishing: Dynamic Security Skins*. Proceedings of the 2005 ACM Symposium on Usable Security and Privacy, July 2005.

[eBay] eBay Toolbar, available at http://pages.ebay.com/ebay_toolbar.

[Felt97] Edward W. Felten, Dirk Balfanz, Drew Dean, and Dan S. Wallach, *Web Spoofing: An Internet Con Game*, Technical Report 540-96, Department of Computer Science, Princeton University, February 1997.

[Gabr02] E. Gabrilovich, A. Gontmakher, *The Homograph Attack described*, Communications of the ACM, 45(2):128, February 2002

[Goth05] Greg Goth, Phishing Attacks Rising, But Dollar Losses Down, IEEE Security and Privacy, Volume 3 Issue 1, January 2005

[Herz04] A. Herzberg, A. Gbara, *TrustBar: Protecting (even Naïve) Web Users from Spoofing and Phishing Attacks*, Bar Ilan University, 2004

[Jako05] Markus Jakobsson, *Modeling and Preventing Phishing Attacks*, Financial Cryptography 2005

[Mark05] Gervase Markham, *IDN Update*, March 24, 2005, http://weblogs.mozillazine.org/gerv/archives/007785.html

[Mozi05] Anonymous, *Mozilla Foundation Security Advisory 2005-29*, Mozilla Organization, February 17th, 2005.

[Oper05] Anonymous, *Advisory: Internationalized domain names (IDN) can be used for spoofing.*, Opera Software ASA, February 25th, 2005.

[Veri05] VeriSign, Inc., *i-Nav Internationalized Domain Name browser plug-in*, http://www.idnnow.com/index.jsp

[Weny05] Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Zhang Min, Xiaotie Deng, Detection of phishing webpages based on visual similarity, Special interest tracks and posters of the 14th International conference on World Wide Web, May 2005.