Internationalized Domain Name Homograph Attacks

Chen Lai, Zhongrong Jian, J. Sidrach

University of California San Diego

CSE 227: Computer Security - Spring 2017

Internationalized Domain Name Homograph Attacks

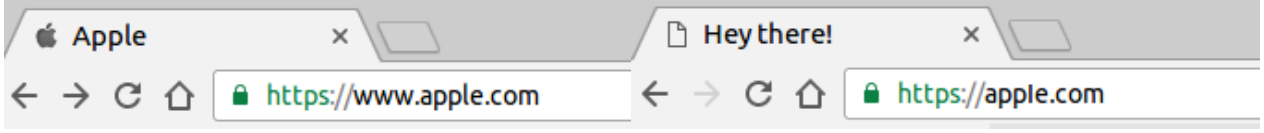
- Domains registered with Punycode (xn– prefix)
- Displayed in Unicode in browsers' navigation bar
- What could go wrong?
- *Some* of the Unicode homoglyphs of a, b, c:

a	ɑ	α	а	𝛼	𝐚	𝑎	𝒂
LATIN SMALL LETTER A	LATIN SMALL LETTER ALPHA	GREEK SMALL LETTER ALPHA	CYRILLIC SMALL LETTER A	APL FUNCTIONAL SYMBOL ALPHA	MATHEMATICAL BOLD SMALL A	MATHEMATICAL ITALIC SMALL A	MATHEMATICAL BOLD ITALIC SMALL A
b	Ь	Ь	Ь	b	𝐛	𝑏	𝒃
LATIN SMALL LETTER B	LATIN CAPITAL LETTER TONE SIX	CYRILLIC CAPITAL LETTER SOFT SIGN	CHEROKEE LETTER SI	CANADIAN SYLLABICS AIVILIK B	MATHEMATICAL BOLD SMALL B	MATHEMATICAL ITALIC SMALL B	MATHEMATICAL BOLD ITALIC SMALL B
c	ϲ	с	⊂	ⅽ	𐐽	𝐜	𝑐
LATIN SMALL LETTER C	GREEK LUNATE SIGMA SYMBOL	CYRILLIC SMALL LETTER ES	LATIN LETTER SMALL CAPITAL C	SMALL ROMAN NUMERAL ONE HUNDRED	DESERET SMALL LETTER CHEE	MATHEMATICAL BOLD SMALL C	MATHEMATICAL ITALIC SMALL C

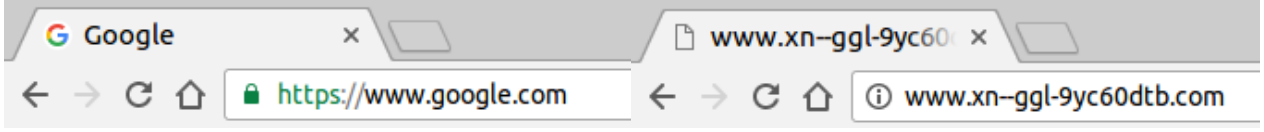
(Image provided by Google)

Defense Mechanisms - Browsers

- URLs displayed in Punycode if certain checks fail
- No common policy across major browsers
- Successful attack



- Unsuccessful attack



Defense Mechanisms - ICANN and TLD Registrars

ICANN

- Rejects ccTLD applications that look similar to existing ones
- Does not enforce restrictions to second-level domains

TLD Registrars

- No common public policy to deal with homograph domains
- Notable exception: Chinese TLDs (simplified/traditional)

Methodology - Clustering

Data sources

- *.com* zone snapshot
- Alexa Top 1 million web sites ranking

Clustering Process

- Filter IDNs from snapshot
- Filter non-IDNs from Alexa ranking
- Cluster all homograph IDNs, using a non-IDN homograph domain from the Alexa ranking as the representative

Methodology - Classification

Manual classification using

- WHOIS records
- HTTP/HTML responses

Categories

- Canonical - Parking
- Canonical - Redirect
- Third Party - Redirect to Canonical
- Third Party - Unrelated
- Third Party - Parking
- Third Party - Scam

Results (1)

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.

Results (2)

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.

Results (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.

Results (4)

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.

Conclusions

Current state

- No common TLD policies in place
- 1000+ homograph IDNs detected
- Most domains inactives (parking)
- No guarantees they will stay parked in the future

Future work

- Analyze other TLDs
- Improve homograph matching algorithm (OCR?)
- Automatic classification of homograph IDNs

Q & A