IDN Visual Security
Deep Thinking

xisigr
Feb, 2019
About me

• Security researcher of Tencent's Xuanwu Lab
  • https://xlab.tencent.com
• Author of Web Front-End Hacker's Handbook
  • https://www.web2hack.org/
• xisigr@Xeye team
• twitter.com/xisigr ,weibo.com/xisigr
Internationalized Domain Names (IDNs)

• IDNs
  • Long time ago, domains could only consist of the Latin letters A to Z, digits, and a few other characters from the US-ASCII coded character set.
  • IDNs allow characters from the Universal Character Set Unicode since 2003.
  • In 2018, Unicode 11.0 contains a repertoire of 137374 characters covering 146 modern and historic scripts.
  • In these more than 10 years, IDNs visual confusion security issues has never been interrupted.
• What do IDNs Mean to you?
ASCII Domain names vs. Internationalized Domain Names (IDNs)

www.test.com

- Third Level Domain
- Second Level Domain
- Top Level Domain

ASCII
- Letters [a-z]
- Digits [0-9]
- Hyphen [-]
- Label length = 63

小明.我爱你

- IDN Second Level Domain
- IDN Top Level Domain

Unicode
- Valid Unicode-Label: IDNA2008
- Valid ASCII-Label: Punycode, "xn--"
- Valid Unicode-Label
- Valid ASCII-label
Homoglyph Attack blowout

- 0 and O; 1, l and I

https://en.wikipedia.org/wiki/Homoglyph
Thinking
IDN TLD

• Until late 2009, TLDs were restricted to only the Latin letters a to z. After 2009, IDN TLDs were introduced in other scripts.
  • 58 IDN ccTLDs evaluated representing 40 countries/territories
  • 56 IDN ccTLDs delegated representing 38 countries/territories
  • Requests cover 33 languages in 19 scripts

• gTLD
  • com, org, net, edu, gov, mil………
  • شبكة(网络), онлайн(在线), グーグル(谷歌), 游戏………(IDN gTLD)

• ccTLD
  • cn, jp, nz, hr, be, cc………
  • موريتانيا(毛里塔尼亚), 新加坡, 한국(韩国), السودان(苏丹)………(IDN ccTLD)
IDN Country Code Top-Level Domains

ccTLDs: 58 for 40*

* Successfully evaluated IDN ccTLDs for total countries and territories

Think about it

• IDN TLD has been around for ten years (2009-2019). With the emergence and increasing of IDN TLD, What changes to IDN visual security?

• Whether Emoji is allowed in IDN or in IDN TLD?
Domain Name Registration Process

https://whois.icann.org/en/domain-name-registration-process
Think about it

• In the whole domain name registration process, which links are relatively weak will help us to hunt IDN spoof.
IDNs Registration Rules

- **gTLD**
  - .COM .NET
  - ......
  - ......

- **ccTLD**
  - .السعودية (Saudi Arabia.)
  - Association's Hebrew IDN rules
  - ......
  - ......
.COM IDNs Registration Rules

• 1. IETF Standards
  • Compliance with all of the RFC that comprise the IDNA2008 standard.

• 2. Restrictions on Specific Languages
  • All IDN registrations require a 3 letter Language Tag.

• 3. Restrictions On Commingling Of Scripts
  • All code points within an IDN must come from the same Unicode script
  • …..

• 4. ICANN’s Restricted Unicode Points
  • …..

• 5. Special Characters
  • …..
<table>
<thead>
<tr>
<th>Arabic</th>
<th>Georgian</th>
<th>Latin</th>
<th>Rejang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenian</td>
<td>Glagolitic</td>
<td>Lepcha</td>
<td>Runic</td>
</tr>
<tr>
<td>Avestan</td>
<td>Greek</td>
<td>Limbu</td>
<td>Samaritan</td>
</tr>
<tr>
<td>Balinese</td>
<td>Gujarati</td>
<td>Lisu</td>
<td>Saurashtra</td>
</tr>
<tr>
<td>Bamum</td>
<td>Gurmukhi</td>
<td>Lycian</td>
<td>Sinhala</td>
</tr>
<tr>
<td>Batak</td>
<td>Han</td>
<td>Lydian</td>
<td>Sundanese</td>
</tr>
<tr>
<td>Bengali</td>
<td>Hangul</td>
<td>Malayalam</td>
<td>Syloti Nagri</td>
</tr>
<tr>
<td>Bopomofo</td>
<td>Hanunoo</td>
<td>Mandaic</td>
<td>Syriac</td>
</tr>
<tr>
<td>Brahmi</td>
<td>Hebrew</td>
<td>Meetei Mayek</td>
<td>Tagalog</td>
</tr>
<tr>
<td>Buginese</td>
<td>Hiragana</td>
<td>Mongolian</td>
<td>Tagbanwa</td>
</tr>
<tr>
<td>Buhid</td>
<td>Imperial Aramaic</td>
<td>Myanmar</td>
<td>Tai Le</td>
</tr>
<tr>
<td>Canadian Aboriginal</td>
<td>Inscriptional Pahlavi</td>
<td>New Tai Lue</td>
<td>Tai Tham</td>
</tr>
<tr>
<td>Carian</td>
<td>Inscriptional Parthian</td>
<td>Nko</td>
<td>Tai Viet</td>
</tr>
<tr>
<td>Cham</td>
<td>Javanese</td>
<td>Ogham</td>
<td>Tamil</td>
</tr>
<tr>
<td>Cherokee</td>
<td>Kaithi</td>
<td>Ol Chiki</td>
<td>Telugu</td>
</tr>
<tr>
<td>Coptic</td>
<td>Kannada</td>
<td>Old Persian</td>
<td>Thaana</td>
</tr>
<tr>
<td>Cuneiform</td>
<td>Katakana</td>
<td>Old South Arabian</td>
<td>Thai</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>Kayah Li</td>
<td>Old Turkic</td>
<td>Tibetan</td>
</tr>
<tr>
<td>Devanagari</td>
<td>Kharoshthi</td>
<td>Oriya</td>
<td>Tifinagh</td>
</tr>
<tr>
<td>Egyptian Hieroglyphs</td>
<td>Khmer</td>
<td>Phags Pa</td>
<td>Vai</td>
</tr>
<tr>
<td>Ethiopic</td>
<td>Lao</td>
<td>Phoenician</td>
<td>Yi</td>
</tr>
</tbody>
</table>
Think about it

• How many IDN rules are there in the world?
• How do we find IDN Spoof under these rules?
Continue Thinking
Unicode Visual Security Issues

Grapheme Rendering

Combining Mark

Mixed Script

Bidirectional

Punycode
Grapheme Rendering

• Inadequate Rendering
  • An additional problem arises when a font or rendering engine has inadequate support for characters or sequences of characters.

• Quick Example

<table>
<thead>
<tr>
<th>String</th>
<th>UTF-16</th>
<th>Punycode</th>
</tr>
</thead>
<tbody>
<tr>
<td>eɋ.com</td>
<td>0065 006C 0323 002E 0063 006F 006D</td>
<td>xn--e-zom.com</td>
</tr>
<tr>
<td>eɋ.com</td>
<td>0065 0323 006C 002E 0063 006F 006D</td>
<td>xn--l-ewm.com</td>
</tr>
<tr>
<td>eɋ.com</td>
<td>1EB9 006C 002E 0063 006F 006D</td>
<td>xn--l-ewm.com</td>
</tr>
</tbody>
</table>

http://www.unicode.org/reports/tr36/#Inadequate_Rendering_Support
CVE-2018-4124

① telugu sign virama (U+0C1C)
② telugu sign virama (U+0C4D)
③ telugu letter nya (U+0C1E)
④ zero width non-joiner (U+200C)
⑤ telugu vowel sign aa (U+0C3E)
Glyphs in Complex Scripts

1. Tamil letter sha (U+0BB6)
2. Tamil sign virama (U+0BCD)
3. Tamil letter ra (U+0BB0)
4. Tamil vowel sign ii (U+0BC0)

1. Tamil letter sa (U+0BB8)
2. Tamil sign virama (U+0BCD)
3. Tamil letter ra (U+0BB0)
4. Tamil vowel sign ii (U+0BC0)
Mixed-Script

• Mixed-Script Spoofing
  • The characters in some scripts, though different in meaning, are usually identical or nearly identical in appearance. However, the existence of visually confusable characters across scripts offers numerous opportunities for spoofing.

• Quick Example

<table>
<thead>
<tr>
<th>String</th>
<th>UTF-16</th>
<th>Punycode</th>
</tr>
</thead>
<tbody>
<tr>
<td>top.com</td>
<td>0074 [03BF] 0070 002E 0063 006F 006D</td>
<td>xn--tp-jbc.com</td>
</tr>
<tr>
<td>top.com</td>
<td>0074 [006F] 0070 002E 0063 006F 006D</td>
<td>top.com</td>
</tr>
</tbody>
</table>

http://www.unicode.org/reports/tr36/#Mixed_Script_Spoofing
Punycode

• PunyCode Spoofing
  • Punycode is a special encoding used to convert Unicode characters to ASCII, which is used to encode internationalized domain names (IDN). The Punycode transformation is relatively dense. That means that it is fairly likely that arbitrary words after the "xn--" will result in valid labels.

• Quick Example
  • URL: http://䕮䕵䕶䕱.com
  • PunyCode URL: http://xn--google.com

http://www.unicode.org/reports/tr36/#TablePunycodeSpoofing
Bidirectional Text

• Bidirectional Text Spoofing
  • When characters are mixed with left-to-right text (LTR) and right-to-left text (RTL), Unicode Bidirectional Algorithm will use a precise set of rules to determine the final visual rendering. However, presented with arbitrary sequences of text, this may lead to text sequences which may be impossible to read intelligibly, or which may be visually confusable.

• Quick Example
  • Access to http://127.0.0.1/%D8%A7/example.org
  • In address bar, maybe visual rendering http://example.org/\1/127.0.0.1

http://www.unicode.org/reports/tr36/#Bidirectional_Text_Spoofing
LTR vs RTL

• URL - LTR
  • Subdomain: hi
  • Domain: google
  • TLD: com
  • Path: search
  • http://hi.google.com/search

• URL - RTL (hebrew)
  • Subdomain: ף
  • Domain: ר
  • TLD: ל
  • Path: מ
  • 渲染：http://ף.ר.ל/מ
Combining Mark

• Combing mark spoofing
  • Combining mark are characters that are intended to modify other characters. Combining character sequence maybe become visually indistinguishable with other characters.

• Quick Example
  • google♀.com
  • ♀ = U+18A9

http://www.unicode.org/reports/tr36/#TableCombiningMarkOrderSpoofing
THINK OUTSIDE THE BOX
Safari Address Bar Spoof Using Latin-d
CVE-2018-4277
<table>
<thead>
<tr>
<th>Latin Extended-D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0</strong></td>
<td></td>
</tr>
<tr>
<td><strong>1</strong></td>
<td></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td></td>
</tr>
<tr>
<td><strong>3</strong></td>
<td></td>
</tr>
<tr>
<td><strong>4</strong></td>
<td></td>
</tr>
<tr>
<td><strong>5</strong></td>
<td></td>
</tr>
<tr>
<td><strong>6</strong></td>
<td></td>
</tr>
<tr>
<td><strong>7</strong></td>
<td></td>
</tr>
<tr>
<td><strong>8</strong></td>
<td></td>
</tr>
<tr>
<td><strong>9</strong></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong></td>
<td></td>
</tr>
<tr>
<td><strong>B</strong></td>
<td></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td></td>
</tr>
<tr>
<td><strong>E</strong></td>
<td></td>
</tr>
</tbody>
</table>

**latin small letter dum (U+A771)**

In Safari address bar
Latin: icloud.com VS Latin Extended-D: icloud.com
xn--iclou-rl3s.com

Address bar says icloud.com - this is not icloud.com

(by xisigr)
Domain include d can be spoof

- Top 10k, domain include d > 25%
  - linkedin.com
  - baidu.com
  - jd.com
  - adobe.com
  - wordpress.com
  - dropbox.com
  - godaddy.com
  - reddit.com
  - ......
Address Bar Spoof using Cyrillic
CVE-2017-5060

https://www.xudongz.com/blog/2017/idn-phishing
Cyrillic

cyrillic small letter a (U+0430)
cyrillic small letter er (U+0440)
cyrillic small letter palochka (U+04CF)
cyrillic small letter ie (U+0435)
Latin: apple.com VS Cyrillic: apple.com

apple.com

Verisign

xn--80ak6aa92e.com

Verisign
Latin: apple.com VS Cyrillic: apple.com

Hey there!

This site is obviously not affiliated with Apple, but rather a demonstration of a flaw in the way Unicode is handled in browsers. **It is very possible that your browser isn’t affected.**

Check out the [complete blog post by Xudong Zheng](https://www.apple.com) for more details.
Firefox URL spoof using RTL
CVE-2018-5117

CVE-2018-5117

https://xn--gbla1c4e.xn--ngbc5azd/

**TLD**: شبكة
punycode=xn--ngbc5azd
Direction: RTL

**SLD**: اسماء
punycode= xn--gbla1c4e
Direction: RTL
https://xn--ggbla1c4e.xn--ngbc5azd
CVE-2018-5117

TLD: شبكة
punycode=xn--ngbc5azd
Direction: RTL

Pathname: #%20%20%20no.io
Direction: RTL

https://xn--ggbla1c4e.xn--ngbc5azd/#%20%20%20no.io

SLD: اسماء
Punycode= xn--ggbla1c4e
Direction: RTL

A lot of %20
%20=Blank character

Hebrew
Direction: RTL
Bidirectional Text

URL Spoof Using RTL IDN TLD
CVE-2018-4205

https://support.apple.com/en-us/HT208854
POC-1

TLD: شبكة
punycode=xn--ngbc5azd
Direction: RTL


(345)-LD: www.apple.com
Direction: LTR

SLD: اماء
Punycode: xn--ggb-la3j
Direction: RTL
POC-1
(by xisigr)
POC-2


TLD: شبكة
punycode=xn--ngbc5azd
Direction: RTL

(345)-LD: www.apple.com
Direction: LTR

SLD: اسماء
Punycode: xn--ggbia3j
Direction: RTL

Pathname: 9999.html
Direction: LTR
POC-2
(by xisigr)
POC-3

http://www.apple.com.xn--ggbla3j.xn--ngbc5azd/999...html

(345)-LD: www.apple.com
Direction: LTR

SLD: اماء
Punycode: xn--ggbla3j
Direction: RTL

Pathname: 99999999999999999999999999999999999999999999999999999999999.html
Direction: LTR

TLD: شبكة
punycode=xn--ngbc5azd
Direction: RTL
POC-3
(by xisigr)
Safari show website address

- show full website address
- only show domain
only show domain

http://www.apple.com.xn--ggbla3j.xn--ngbc5azd/999...html

POC-4

http://www.apple.com.xn--ggbla3j.xn--ngbc5azd/%DB%B0.html

TLD: شبكة
punycode=xn--ngbc5azd
Direction: RTL

(345)-LD: www.apple.com
Direction: LTR

SLD: اماء
Punycode: xn--ggbla3j
Direction: RTL

Pathname: ۰.html

U+06F0: extended arabic-indic
digit zero
Direction: LTR??
POC-4
(by xisigr)
Origin out-of-order

www.gmail.com.xn--mgb.999.xn--ggbla3j.xn--ngbc5azd
Firefox/Chrome

spoof by xisigr

spoof by xisigr
Safari/Edge

spoof by xisigr

spoof by xisigr
Think about it

• How to resolve RTL Spoof in address bar?
• Only show origin?
  • showing only the origin reduces utility, though. That's for the UX designers to decide.
  • Either way, it doesn't fully resolve these issues. You can still do RTL spoofs with just the origin (like having the labels shown out-of-order).
• Some error logic.
  • Eg: "www" or "m" subdomain is removed
  • https://bugs.chromium.org/p/chromium/issues/detail?id=881694

• Spoofable RTL URLs in the UI
  • https://bugs.chromium.org/p/chromium/issues/detail?id=351639
Combining character

Safari Address Bar Spoof Using Combining character
CVE-2018-4260

<table>
<thead>
<tr>
<th>Hebrew</th>
<th>Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>עברית</td>
<td>י (Vav)</td>
<td>Hebrew letter vav (U+05D5)</td>
</tr>
<tr>
<td></td>
<td>◯ (Holam)</td>
<td>Hebrew point holam (U+05B9)</td>
</tr>
<tr>
<td></td>
<td>צ (Samekh)</td>
<td>Hebrew letter samekh (U+05E1)</td>
</tr>
</tbody>
</table>
combining character sequence

• Character + combining mark = *combining character sequence*

hebrew point holam (U+05B9)

hebrew letter vav (U+05D5)
Latin: io.com VS Hebrew: יז.com

Latin: io.com

Hebrew: יז.com

combining character sequence

hebrew letter vav (U+05D5)
hebrew point holam (U+05B9)
hebrew letter samekh (U+05E1)
Latin: io.com VS Hebrew: ישראל.com

Engineered for a new generation of needs and demands, IO delivers the data center as a service @scale™.

Fake io.com

V6yADoAo/1Hm78m5/E/I7EgHJNefB1cimz1ZRntjoMCOAFc+8a (Spoof by xisigr)
Hebrew alphabet: וס.com VS Hebrew: וס.com

- Hebrew letter vav with holam (U+FB4B)
- Hebrew letter samekh (U+05E1)

combining character sequence

xn--meb6676j.com

hebrew letter vav (U+05D5)
hebrew point holam (U+05B9)
hebrew letter samekh (U+05E1)

xn--hdb9c9a.com

io.com
grapheme is same but meaning is difference

Alphabetic Presentation Forms

combining character sequence

(U+FB4B) → 0x0 FB 4B

(U+05D5) + (U+05B9) → 0x0 05 D5 05 B9
Think about it

• Should combination character sequence be supported to register?
• Is it possible to strictly check and prevent such malicious intentions at the Registrar stage?
• Think about these questions in the URL, you may find something new.
  • http://unicode.org/faq/char_combmark.html.
IDN Challenging

- Browser
- Rules
- ICANN
- Registrar
- Unicode

IDN
thinking of you