Assessing Browser-level Defense against IDN-based Phishing

Hang Hu\textsuperscript{1}\textsuperscript{*}  Steve T.K. Jan\textsuperscript{1,2}\textsuperscript{*}  Yang Wang\textsuperscript{1}  Gang Wang\textsuperscript{1}

\textsuperscript{1}University of Illinois at Urbana and Champaign
\textsuperscript{2}Virginia Tech

* Equal contribution
Imagine Visiting This Website...

SOMETHING SMELLS....

PHISHY.............
Internationalized Domain Names (IDN)

- Cyrillic Small Letter A
  - Block: Cyrillic
  - Script: Cyrillic

- Latin Small Letter A
  - Block: Basic Latin
  - Script: Latin
IDN Homography

• **IDN** allows people around the world to use their own language for domain names
  – Support **Unicode** characters
  – Use **Punycode** to work with legacy systems such as DNS

  bücher.de
  books
  🇩🇪 🇬🇧 🇺🇸

  **Unicode:** “bücher.de”
  **Punycode:** “xn--bcher-kva.de”

• IDN homograph enable highly deceptive phishing
  – Exploits the fact that different Unicode characters look alike
Browser Defense

- Displaying Punycode as a defense

- But we observe inconsistent reactions sometimes
  - Punycode not shown when a phishing site mimics a popular domain name
This Paper: Research Questions

• What policies do major browser vendors implement to prevent IDN homographs, and how well are they enforced?

• Are there ways to systematically bypass existing policies to create homograph IDNs?

• How well can end users recognize homograph IDNs?

Black-box measurements across browser vendors and versions (2015-2020)

User study
Blackbox Testing (1): Claimed Policies

- Claimed policies vary across browsers

- Unicode script mixing (blocked)
- Unicode script mixing (allowed)
- Skeleton rule (top domains)
- Whole-script confusable + TLD
- Confusable characters (blocked)
- Unicode scripts (allowed)

Publicly available Documentations/code

- Domain “skeleton” matches with top domain names (5000 popular sites)
- All characters in the domain name are Cyrillic (no-mixing). But TLD is not Cyrillic!
Blackbox Testing (1): Claimed Policies

- Claimed policies vary across browsers

5,673 Testing IDNs
To test the claimed policies
Blackbox Testing (2): Evasion

• Construct potentially evasive testing cases

IDNs with more extended Unicode confusible characters

Important target domain names beyond the “popular list” (e.g., regional hospitals)

Whole-script confusible + allowed TLD

Within the prohibited scripts, certain Unicode blocks can be mixed

Implementing the Test Framework

• Testing browsers across platforms and versions

<table>
<thead>
<tr>
<th>Browsers</th>
<th>Versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chrome (21)</td>
<td>51.0-81.0</td>
</tr>
<tr>
<td>Firefox (15)</td>
<td>61.0-75.0</td>
</tr>
<tr>
<td>Microsoft Edge (6)</td>
<td>15.0-18.0</td>
</tr>
<tr>
<td>Safari (4)</td>
<td>10.0-13.0</td>
</tr>
<tr>
<td>IE (4)</td>
<td>8.0-11.0</td>
</tr>
<tr>
<td>Android Chrome (7)</td>
<td>5.0-9.0</td>
</tr>
<tr>
<td>iOS Safari (13)</td>
<td>10.2-13.2</td>
</tr>
</tbody>
</table>

Testing IDNs

Browser automation + Screen recording

Video frame analysis
OCR (image → text)
Classify Punycode
# Result Analysis (on 9K Testing IDNs)

<table>
<thead>
<tr>
<th>Browsers</th>
<th>Chrome</th>
<th>Firefox</th>
<th>Safari</th>
<th>Edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>1,963</td>
<td>4,233</td>
<td>4,085</td>
<td>1,963</td>
</tr>
<tr>
<td>Failure Rates</td>
<td>20.62%</td>
<td>44.46%</td>
<td>42.91%</td>
<td>20.62%</td>
</tr>
</tbody>
</table>

- Latest versions of browsers (as of May 2020)
  - All browsers failed on certain testing cases
  - Chrome is stricter compared with others, with lowest failure rates

Defense Failed
Result Analysis (Evasion Tests)

Chrome supported IDNs first (immediately vulnerable)

Firefox/Safari policies did not have major updates for 2+ Yrs

Major updates on IDN polices

Revoked IDN policies (reallowed Unicode blocks Such as “Latin Extended-A”)

Not yet supporting IDNs
Homograph IDNs in Practice

• Are there IDNs impersonating real-world websites?

<table>
<thead>
<tr>
<th>.com Zone file</th>
<th>IDNs</th>
<th>Homograph IDNs impersonating top 10K sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>400 million</td>
<td>916, 805</td>
<td>1,855</td>
</tr>
</tbody>
</table>

...google.com, microsoft.com, ąsos.com, spóttify.com, wellsfärgo.com, amazon.com, coinbasē.com, göögle.com, bitçoin.com, bitcoin.com ...  

35.9% bypassed **Chrome v81.0**  
90.3% bypassed **Safari v13.0**  
93.9% bypassed **Firefox v75.0**
User Study Results

Q: Would users fall for homograph IDNs?

Homograph IDNs that bypassed Chrome defense are still deceptive to users (about 45% of error rates)

Assessing Browser-level Defense against IDN-based Phishing

Hang Hu\textsuperscript{2}, Steve T.K. Jan\textsuperscript{*1,2}, Yang Wang\textsuperscript{1}, Gang Wang\textsuperscript{1}
\textsuperscript{1}University of Illinois at Urbana-Champaign \textsuperscript{2}Virginia Tech
\{hanghu, tekang\}@vt.edu, \{yyw, gangw\}@illinois.edu

Abstract

Internationalized Domain Names (IDN) allow people around the world to use their native languages for domain names. Unfortunately, because characters from different languages were introduced and standardized in 2003 [28], which support Unicode characters from a variety of languages.

As more IDNs are registered, a growing concern is that IDN can be used to impersonate other domain names for phishing purposes. This is done by different characters from the same language...
Countermeasures

• Add new rules to address failed cases
  – Difficult to guarantee completeness

• Use visual similarity metrics (e.g., perceptual hashing) to detect impersonation against a wide range of domains
  – Scalability issues, may have false positives

• Disabling IDNs by default
  – Only shows Unicode when the IDNs match users’ browser language(s)
Conclusions

• Empirical tests on major browser vendors on their IDN homograph defense schemes
  – All tested browsers have weaknesses in their defense policies
  – Not all the browsers improve their defense overtime

• User study shows homograph IDNs are deceptive to users

• Reported results to Chrome, Firefox, and Safari
Thank You!

https://gangw.cs.illinois.edu/
gangw@Illinois.edu