A Reexamination of Internationalized Domain Names: the Good, the Bad and the Ugly

Baojun Liu\textsuperscript{1}, Chaoyi Lu\textsuperscript{1}, Zhou Li\textsuperscript{2}, Ying Liu\textsuperscript{1}, Haixin Duan\textsuperscript{1}, Shuang Hao\textsuperscript{3} and Zaifeng Zhang\textsuperscript{4}

\textsuperscript{1}Tsinghua University, \textsuperscript{2}IEEE Member, \textsuperscript{3}University of Texas at Dallas, \textsuperscript{4}Netlab of 360
Spot The Difference!

Real Apple

xn--80ak6aa92e.com
The Party Going on...

• Can we believe what we see?

facebook.com  facebook.com  facebook.com  facebook.com
facebook.com  facebook.com  facebook.com  facebook.com
facebook.com  facebook.com  facebook.com  facebook.com

Bookmark or type your own URL Kids!

Share  Save

Programmer  1 point · 4 months ago

totally need to check it beforehand next time.

Share  Save

2 points · 4 months ago
Internationalized Domain Names

• To build a multilingual Internet
  • Standardized by RFC3490 (IDNA, 2003)
  • Registration authorized by ICANN in 2003

• Allowed at different domain levels
  • 151 IDN TLDs until May 2018 (e.g., 中国, xn--fiqs8s)
  • Offered under TLDs (e.g.,テスト.com)

例子.测试
예제들.테스트
例.テスト
(example.test in different languages)
Encoding of IDN

- **Punycode**
  - Defined by RFC3492 for IDNA
  - Converting Unicode strings to ACE strings
  - For backward compatibility in DNS

他们为什么不说中文
(Why don’t they speak Chinese)

\[
\text{Punycode} \quad \& \quad \text{prefixing}
\]

\[\text{xn--ihqwcrb4cv8a8dqqg056pqjye}\]

Can be used as domain labels
A Reexamination

• 15+ years since the first installation
  • Greatly promoted by ICANN and several registries
  • Volumes are increasing over the years
  • Controversial: homograph attack, IDN deception, ...
  • Not yet comprehensively studied

• Revisiting the IDN initiative
  • IDN development / characteristics
  • Kind / scale of abuse
Dataset Collection

Zone Files

com
net
org
iTLD

“xn--”

IDN

non-IDN

Sample

Domain lists

WHOIS
Passive DNS
URL Blacklists
SSL Certificates
### Dataset Collection

- **Collected dataset**

<table>
<thead>
<tr>
<th>TLD</th>
<th>Snapshot on</th>
<th># IDN (SLD)</th>
<th>WHOIS</th>
<th>Blacklisted</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>Sept 21, 2017</td>
<td>1,007,148</td>
<td>590,542</td>
<td>5,284</td>
</tr>
<tr>
<td>net</td>
<td>Sept 21, 2017</td>
<td>231,896</td>
<td>131,573</td>
<td>746</td>
</tr>
<tr>
<td>org</td>
<td>Oct 5, 2017</td>
<td>25,629</td>
<td>19,271</td>
<td>59</td>
</tr>
<tr>
<td>iTLD (53)</td>
<td>Oct 5, 2017</td>
<td>208,163</td>
<td>2,226</td>
<td>152</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>-</td>
<td><strong>1,472,836</strong></td>
<td><strong>739,160</strong></td>
<td><strong>6,241</strong></td>
</tr>
</tbody>
</table>
• **A. Language**
  - Using LangID* for language identification
  - **75%+** IDN are in languages of east Asian countries

<table>
<thead>
<tr>
<th>Language</th>
<th>Total IDN</th>
<th>Blacklisted IDN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>766,735 (52.03%)</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>3,495 (56.02%)</td>
<td>238 (3.81%)</td>
</tr>
<tr>
<td>Korean</td>
<td>128,291 (8.71%)</td>
<td>902 (14.46%)</td>
</tr>
<tr>
<td>German</td>
<td>72,110 (4.90%)</td>
<td>119 (1.91%)</td>
</tr>
</tbody>
</table>

[*] langid.py: An off-the-shelf language identification tool. ACL 2012
IDN Characteristics

• B. Registration
  • Correlating with WHOIS data
  • Creation date

![Graph showing IDN Characteristics]
IDN Characteristics

- **B. Registration**
  - Correlating with WHOIS data
  - Creation date
  - Registrant

<table>
<thead>
<tr>
<th>Email</th>
<th># IDN</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:776053229@qq.com">776053229@qq.com</a></td>
<td>2,609</td>
<td>All are southwest city names in China.</td>
</tr>
<tr>
<td><a href="mailto:daidesheng88@gmail.com">daidesheng88@gmail.com</a></td>
<td>1,562</td>
<td>All are about online gambling.</td>
</tr>
<tr>
<td><a href="mailto:tetetw@gmail.com">tetetw@gmail.com</a></td>
<td>1,453</td>
<td>All are short words in Chinese.</td>
</tr>
</tbody>
</table>

**Large-scale opportunistic registrations, of specific pattern / topic**
IDN Characteristics

• B. Registration
  • Correlating with WHOIS data
  • Creation date
  • Registrant
  • Registrar

22.99% (JP) 4.02% (KR)
10.86% (CN) 1.88%

East Asian markets are more active.
IDN Characteristics

• C. DNS statistics

  • **Active time & query volume** (IDN vs. non-IDN)
  • IDNs have **shorter active time**, except malicious ones
IDN Characteristics

- **C. DNS statistics**
  - **Active time & query volume** (IDN vs. non-IDN)
  - IDNs have **shorter active time**, except malicious ones
  - IDNs are **visited less frequently**, except malicious ones

Malicious IDNs are effective at trapping users.
IDN Characteristics

• D. Content & intention
  • Manual classification of 500 webpages

More likely leading to errors or meaningless content, for IDNs.
IDN Characteristics

• E. SSL certificate
  • 4.5%+ (65K+) IDN install invalid certificates, which is similar to prior study on all domains*.
  • Most certificates are shared among domains.

<table>
<thead>
<tr>
<th>Category</th>
<th># IDN (% certificates)</th>
<th># non-IDN (% certificates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expired</td>
<td>8,411 (12.5%)</td>
<td>8,730 (24.9%)</td>
</tr>
<tr>
<td>Invalid Authority</td>
<td>12,169 (18.1%)</td>
<td>5,801 (16.7%)</td>
</tr>
<tr>
<td>Invalid Common Name</td>
<td>45,133 (67.3%)</td>
<td>19,527 (45.5%)</td>
</tr>
</tbody>
</table>

[*] Analysis of the HTTPS certificate ecosystem. IMC 2013
IDN Characteristics

• To sum up
  • Volume: 1.4M IDNs account for 1% domains
  • Language: east Asian countries are at the front line
  • Registration: long-term & opportunistic both exist
  • Visits: IDNs are less active than non-IDNs
  • Content: less IDNs are with meaningful content
  • SSL certificate: certificate sharing is prevalent
IDN Abuse in Blacklists

• Homograph attack
  • Exploits visual resemblance among domains

• Semantic attack
  • Type-1: brand name + keyword
    icloud登录.com  apple邮箱.com
  • Type-2: translating English keywords
    mercedes-benz.com  奔驰汽车.com
## Homograph Attack

### A. Browser policies

- RFC3490 (IDNA): avoid exposing raw ACE encoding
- Firefox & Chrome: display based on **character sets**

<table>
<thead>
<tr>
<th>Platform</th>
<th>PC</th>
<th>iOS</th>
<th>Android</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browser</td>
<td>Ver.</td>
<td>iTLD IDN Supported</td>
<td>Homograph Attack</td>
</tr>
<tr>
<td>Chrome</td>
<td>62.0</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Firefox</td>
<td>57.0</td>
<td>Need prefix</td>
<td>Bypassed</td>
</tr>
<tr>
<td>Opera</td>
<td>49.0</td>
<td>/</td>
<td>Bypassed</td>
</tr>
<tr>
<td>Safari</td>
<td>11.0</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>IE</td>
<td>11.0</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>QQ</td>
<td>9.7</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Baidu</td>
<td>8.7</td>
<td>/</td>
<td>Bypassed</td>
</tr>
<tr>
<td>Qihoo 360</td>
<td>9.1</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Sogou</td>
<td>7.1</td>
<td>/</td>
<td>Vulnerable</td>
</tr>
<tr>
<td>Liebao</td>
<td>6.5</td>
<td>/</td>
<td>Bypassed</td>
</tr>
</tbody>
</table>

- **Bypassed**: Browser behavior is as expected, Homograph Attack is not triggered.
- **Vulnerable**: Browser behavior is not as expected, Homograph Attack is triggered.
- **Need prefix**: Browser requires a prefix to display the URL correctly.
- **Unicode only**: URL is displayed in Unicode format.
- **Punycode only**: URL is displayed in Punycode format.
- **Not supported**: Browser does not support iTLD IDN.
- **Title**: Browser uses title attribute as fallback.
Homograph Attack

• A. Browser policies
  • RFC3490 (IDNA): avoid exposing raw ACE encoding
  • Firefox & Chrome: display based on character sets
  • Manual survey

<table>
<thead>
<tr>
<th>Input</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple.com (xn--80ak6aa92e.com)</td>
<td>Punycode</td>
</tr>
<tr>
<td>Only the ‘l’ is Cyrillic.</td>
<td></td>
</tr>
<tr>
<td>soso.com (xn--n1aa1eb.com)</td>
<td>Unicode</td>
</tr>
<tr>
<td>ALL characters in the SLD are Cyrillic.</td>
<td></td>
</tr>
</tbody>
</table>

Some up-to-date policies still need to be revised.
B. Detecting homographic IDNs

SSIM index*: a metric of visual resemblance

1.4M IDNs

Image rendering

Pairwise SSIM

homographic IDNs

### Homograph Attack

**C. Registered homographic IDNs**

- **1,516** homographic IDNs detected (100 blacklisted)
- Brands: few defensive registration

<table>
<thead>
<tr>
<th>Brand Domain</th>
<th># Homographic IDN (% of 1,516)</th>
<th># Defensive Registration</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>121 (8.0%)</td>
<td>19</td>
</tr>
<tr>
<td>facebook.com</td>
<td>98 (6.5%)</td>
<td>0</td>
</tr>
<tr>
<td>amazon.com</td>
<td>55 (3.6%)</td>
<td>14</td>
</tr>
<tr>
<td>icloud.com</td>
<td>42 (2.8%)</td>
<td>0</td>
</tr>
<tr>
<td>youtube.com</td>
<td>41 (2.7%)</td>
<td>0</td>
</tr>
</tbody>
</table>
Homograph Attack

• **C. Registered homographic IDNs**
  - 1,516 homographic IDNs detected (100 blacklisted)
  - Brands: few defensive registration
  - Long active time & considerable visits

![Graphs showing active time and DNS queries](image-url)
Homograph Attack

• **C. Registered homographic IDNs**
  - **1,516** homographic IDNs detected (100 blacklisted)
  - Brands: few defensive registration
  - Long active time & considerable visits
  - Few (15%-) are in active use, from manual sampling
**D. Available homographic IDNs**

- Generate 128,432 IDNs from brand domains, using homoglyphs* to replace the original characters
- **42,671** are homographic (only 237 are registered)

[*] The methodology and an application to fight against unicode attacks. SOUPS 2006
To sum up

- **Browsers** have responded to the homograph threat; some up-to-date policies still need to be revised
- **Defensive registrations** are in the minority
- Most homographic IDNs are not yet delivering useful content
- **Choices of homographic IDNs** are substantial
A. Detection

- Remove the non-ASCII characters from each IDN
- Compute the pairwise SSIM with brand domains
- Only if SSIM says identical
- Which means: the IDN contains an intact brand
Semantic Attack

- **B. Registered abusive IDNs**
  - **1,497** abusive IDNs detected
  - Long active time & considerable visits
  - **85%+** are inactive

![Graph 1: Ratio of abusive IDN vs Active time (day)](image1)
- 50%+
- 600+ days active

![Graph 2: Ratio of abusive IDN vs # DNS query (log)](image2)
- 40%
- 200+ queries
Discussion

• **Mitigating IDN abuse**
  
  • **Registry**: check for abusive registration
  
  • **Registrar**: avoid parking for abusive IDNs
  
  • **Browser**: enforce a proper IDN policy
  
  • **Users**: education; check when visiting websites
Summary

• **IDN development**
  • Volume of IDN is **steadily growing, 1.4M+** registered
  • East Asian countries are active at registration
  • IDNs’ visits and content are **still under expectation**

• **IDN Abuse**
  • Homograph attack & semantic attack
  • Efforts should be spread by various entities
Thanks for your attention!

Questions?