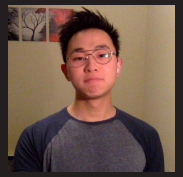
Assessing Browser-level Defense against IDN-based Phishing




Hang Hu[1*] Steve T.K. Jan[1,2*] Yang Wang[1] Gang Wang[1]

[1] University of Illinois at Urbana and Champaign
[2] Virginia Tech

* Equal contribution

Apple

apple.com

Mac iPad iPhone Watch TV Music Support

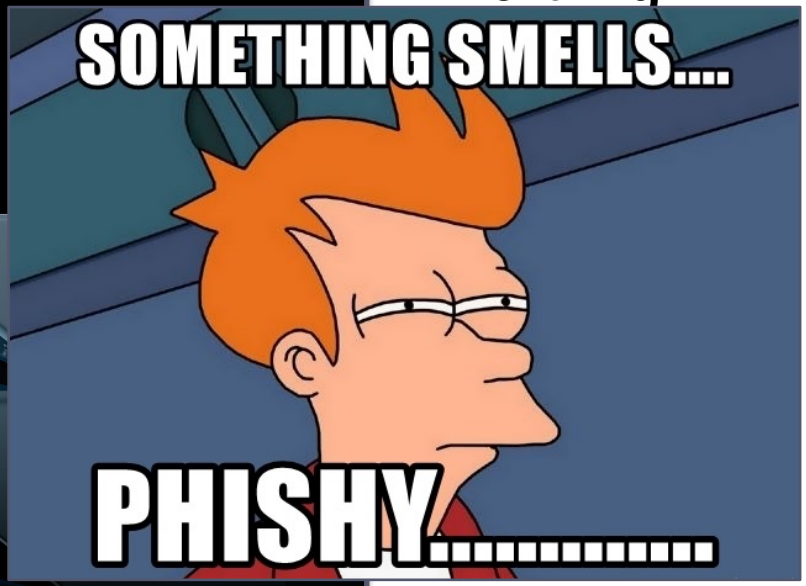
iPhone 12 Pro

It's a leap year.

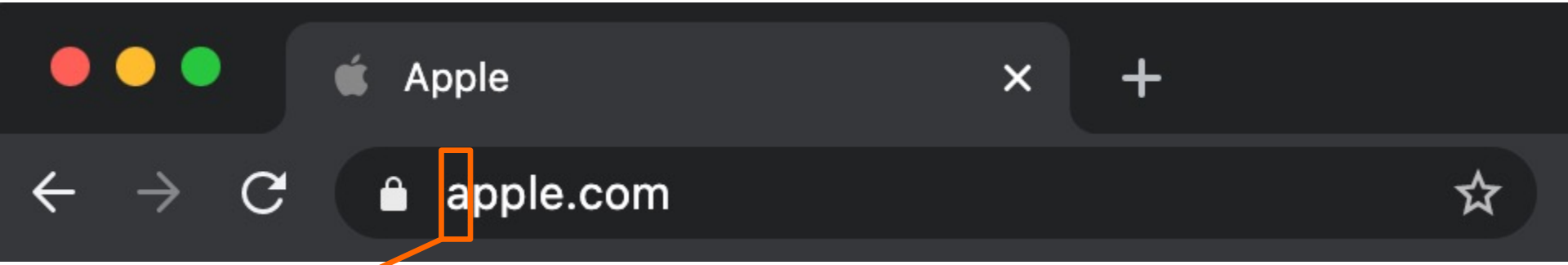
From $41.62/mo. for 24 mo. or $999 before trade-in.[3]
Buy directly from Apple with special carrier offers.

Learn more › Buy ›

Imagine
Visiting



Internationalized Domain Names (IDN)



а

U+0430

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

a

U+0061

- 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
books

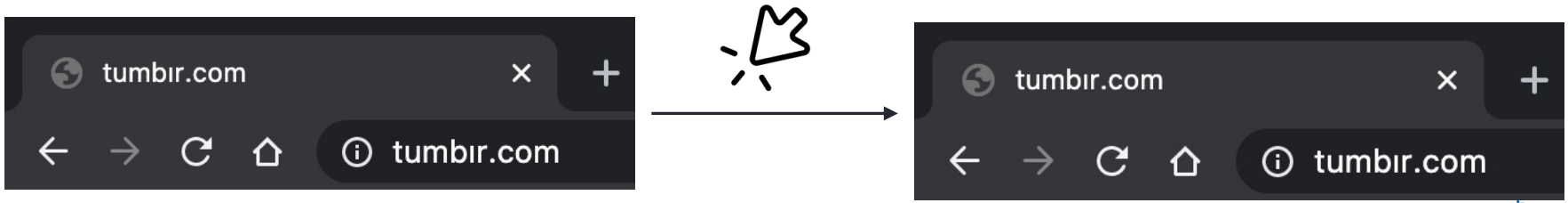
Unicode: “bücher.de”
Punycode: “xn--bcher-kva.de”

- IDN homograph enable highly deceptive phishing
  - Exploits the fact that different Unicode characters look alike

Browser Defense

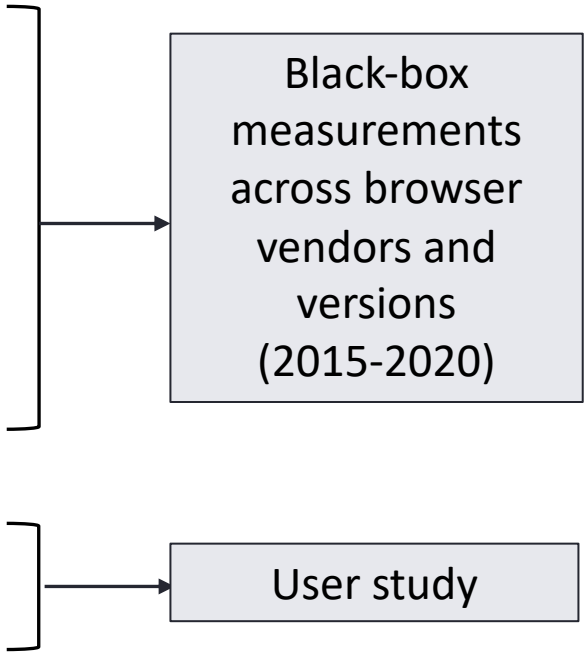
- Displaying Punycode as a defense

- But we observe inconsistent reactions sometimes
  - Punycode not shown when a phishing site mimics a popular domain name



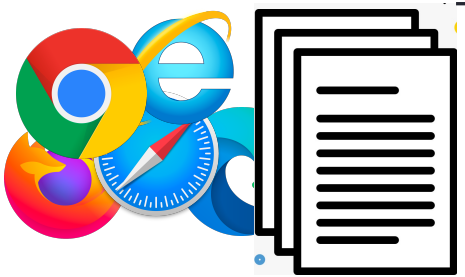
This Paper: Research Questions

- 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?



Blackbox Testing (1): Claimed Policies

- Claimed policies vary across browsers



Publicly available Documentations/code

Unicode script mixing (blocked)

Unicode script mixing (allowed)

Skeleton rule (top domains)

Whole-script confusable + TLD

Confusable characters (blocked)

Unicode scripts (allowed)

日本語.jp

googlé.com

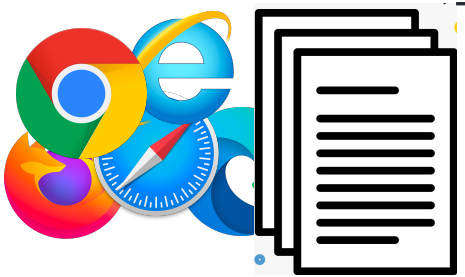
Domain “skeleton” matches with top domain names (5000 p

apple.com

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



Publicly available Documentations/code

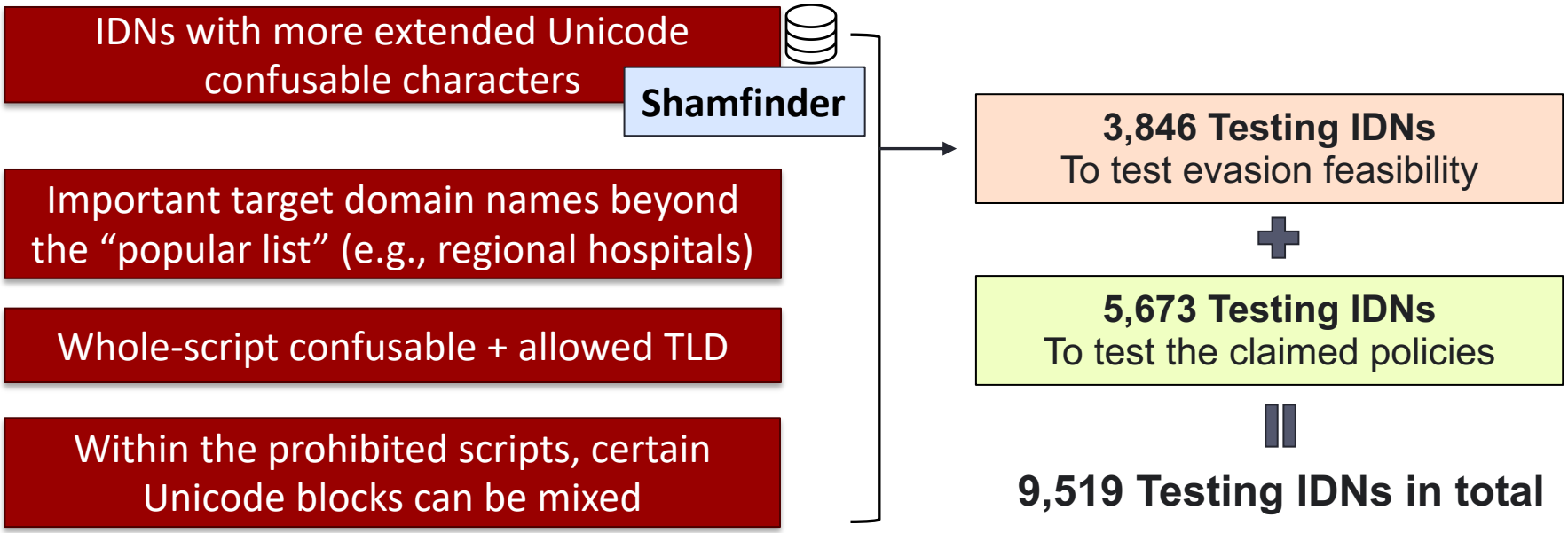
- Unicode script mixing (blocked)
- Unicode script mixing (allowed)
- Skeleton rule (top domains)
- Whole-script confusable + TLD
- Confusable characters (blocked)
- Unicode scripts (allowed)

5,673 Testing IDNs

To test the claimed policies

Blackbox Testing (2): Evasion

- Construct potentially evasive testing cases



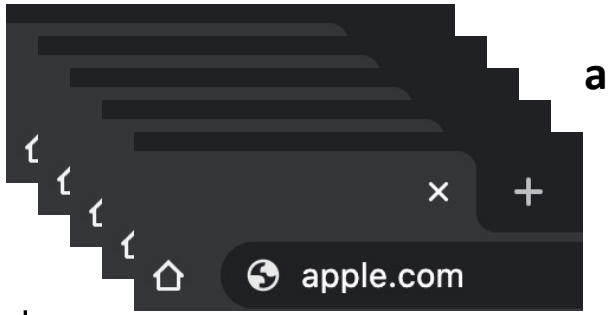
Shamfinder: An automated framework for detecting IDN homographs. In Proc. of IMC, 2019

Implementing the Test Framework

- Testing browsers across planforms and versions

Testing IDNs

Browsers	Versions
Chrome (21)	51.0-81.0
Firefox (15)	61.0-75.0
Microsoft Edge (6)	15.0-18.0 79.0-81.0
Safari (4)	10.0-13.0
IE (4)	8.0-11.0
Android Chrome (7)	5.0-9.0
iOS Safari (13)	10.2-13.2



Browser automation + Screen recording

Video frame analysis

OCR (image → text)

Classify Punycode

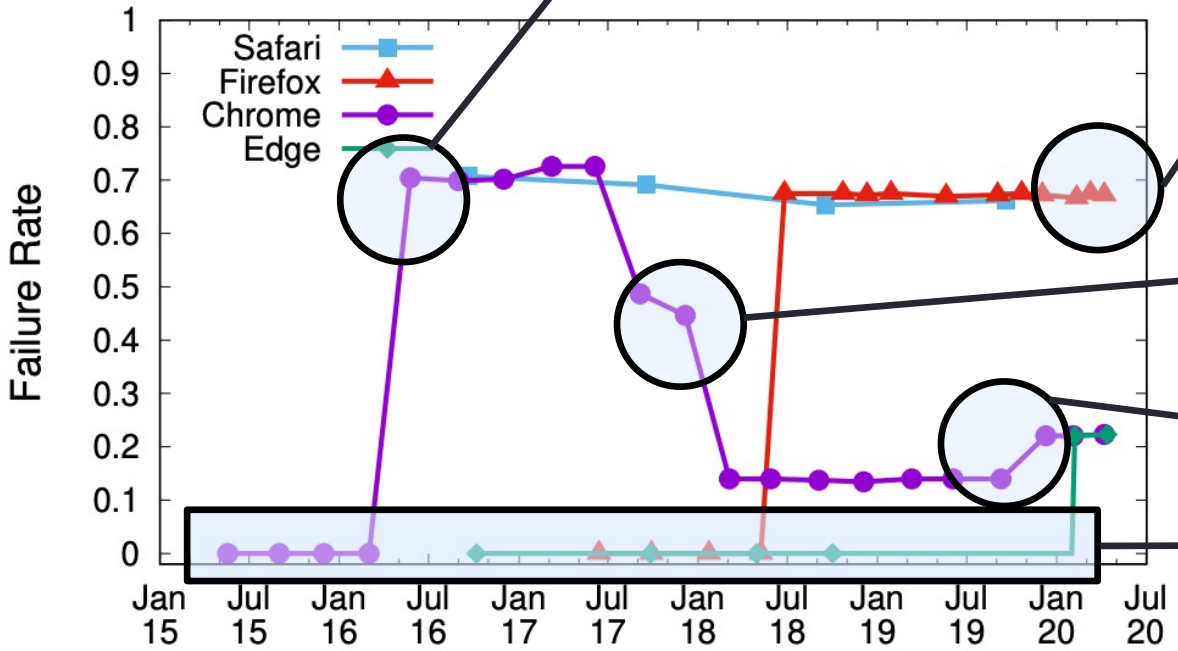
Result Analysis (on 9K Testing IDNs)

Defense Failed

Browsers	Chrome	Firefox	Safari	Edge
Unicode	1,963	4,233	4,085	1,963
Failure Rates	20.62%	44.46%	42.91%	20.62%

- 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

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?

.com Zone file	IDNs	Homograph IDNs impersonating top 10K sites
400 million	916, 805	1,855

google.com, mıcrosoft.com, ąsos.com,
spótify.com,wellsfärgo.com,
amȷazon.com, coinbasė.com,
gȯȯgle.com, bitçoin.com, bitcọin.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[*2], Steve T.K. Jan[*1,2], Yang Wang[1], Gang Wang[1]
[1]University of Illinois at Urbana-Champaign [2]Virginia Tech
{hanghu, tekang}@vt.edu, {yvw, 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

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