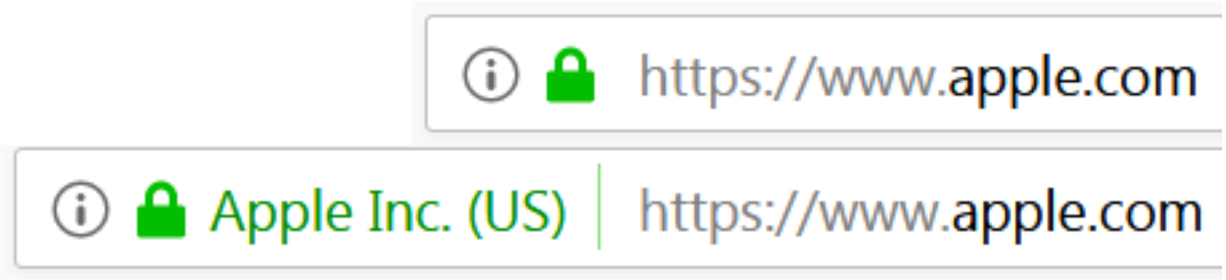
A Reexamination of Internationalized Domain Names: the Good, the Bad and the Ugly

Baojun Liu[1], Chaoyi Lu[1], Zhou Li[2], Ying Liu[1],
Haixin Duan[1], Shuang Hao[3] and Zaifeng Zhang[4]

[1] Tsinghua University, [2] IEEE Member,
[3] University of Texas at Dallas, [4] Netlab of 360

Spot The Difference!

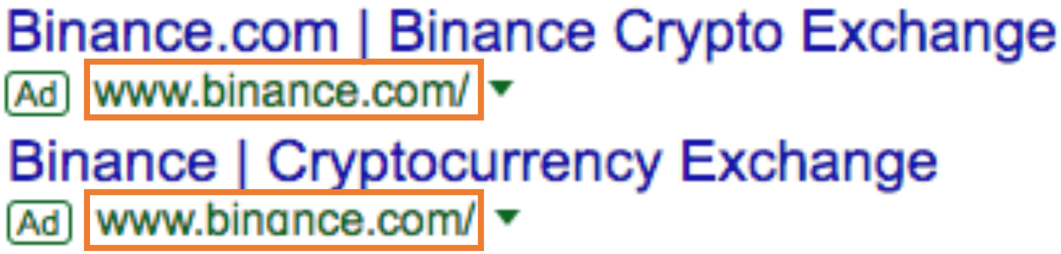


ⓘ 🔒 https://www.apple.com

ⓘ 🔒 Apple Inc. (US) | https://www.apple.com

xn--80ak6aa92e.com

Real Apple

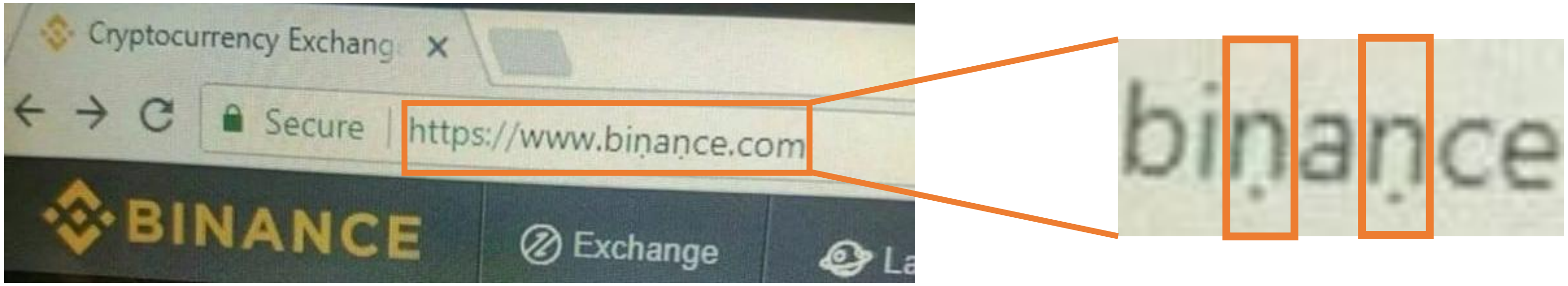


Binance.com | Binance Crypto Exchange

Ad www.binance.com/ ▾

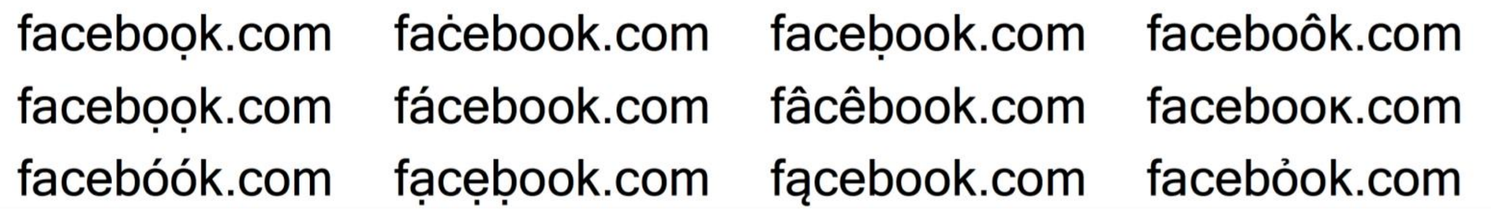
Binance | Cryptocurrency Exchange

Ad www.binance.com/ ▾



The Party Going on…

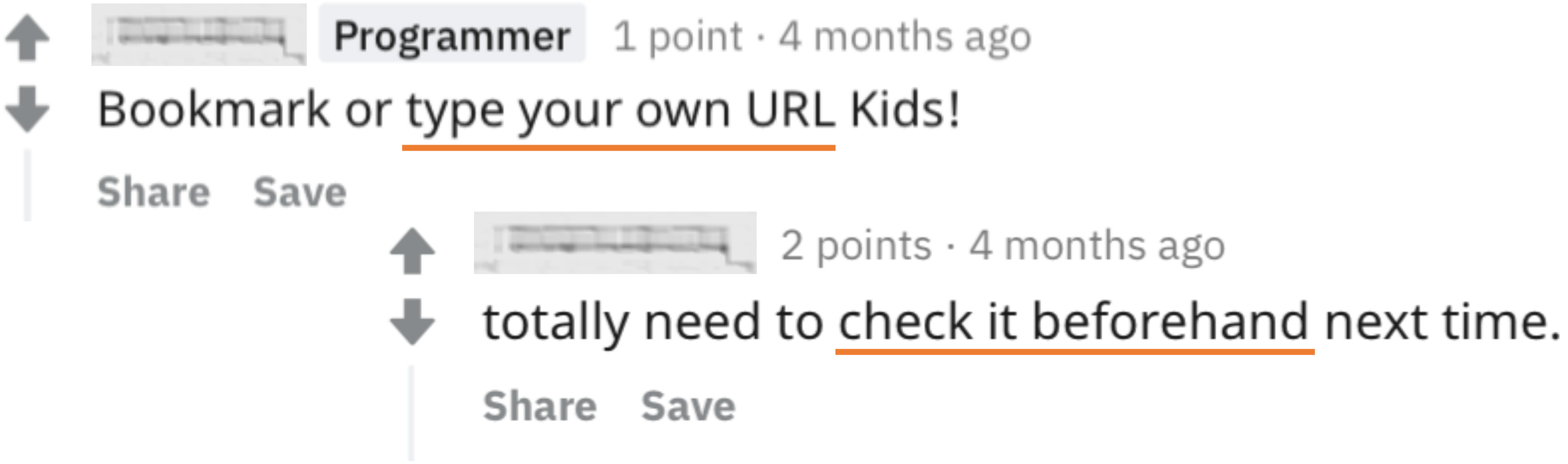
- Can we believe what we see?



faceboọk.com faċebook.com faceḅook.com faceboôk.com

facebọọk.com fácebook.com fâcêbook.com facebooĸ.com

facebóók.com fạcẹḅook.com fącebook.com facebỏok.com



[redacted] Programmer 1 point · 4 months ago

Bookmark or type your own URL Kids!

Share Save

[redacted] 2 points · 4 months ago

totally need to check it beforehand next time.

Share Save

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 June 2018 (e.g., 中国, xn--fiqs8s)
  - Offered under TLDs (e.g., テスト.com)

إختبار.مثال

例子.测试

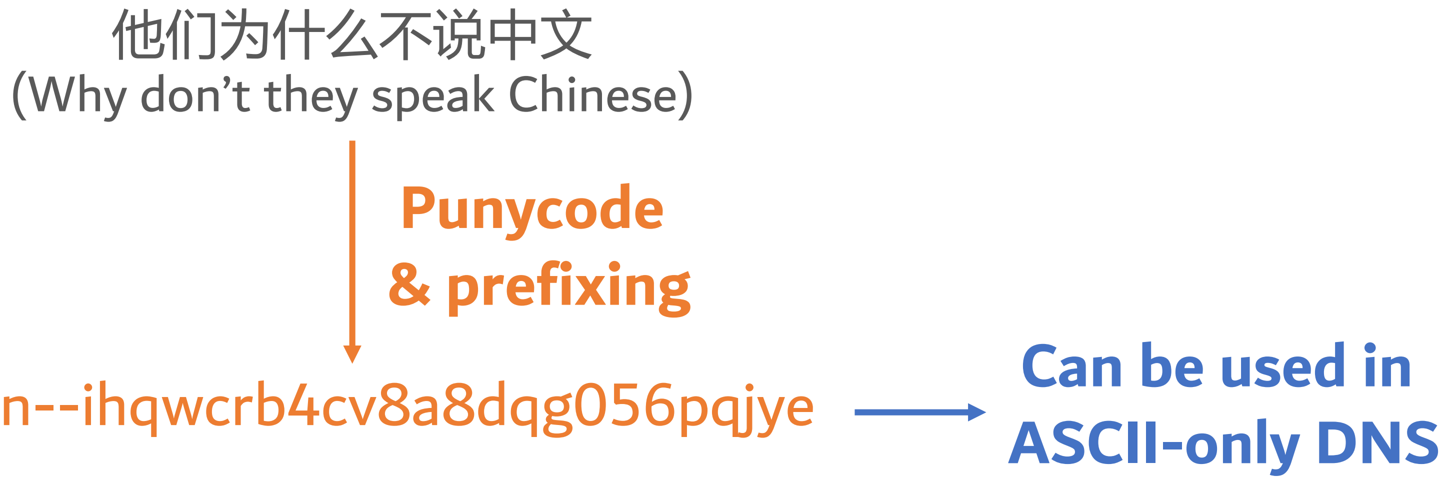
예제들.테스트

例. テスト

(example.test in different languages)

Encoding of IDN

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



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

Punycode & prefixing

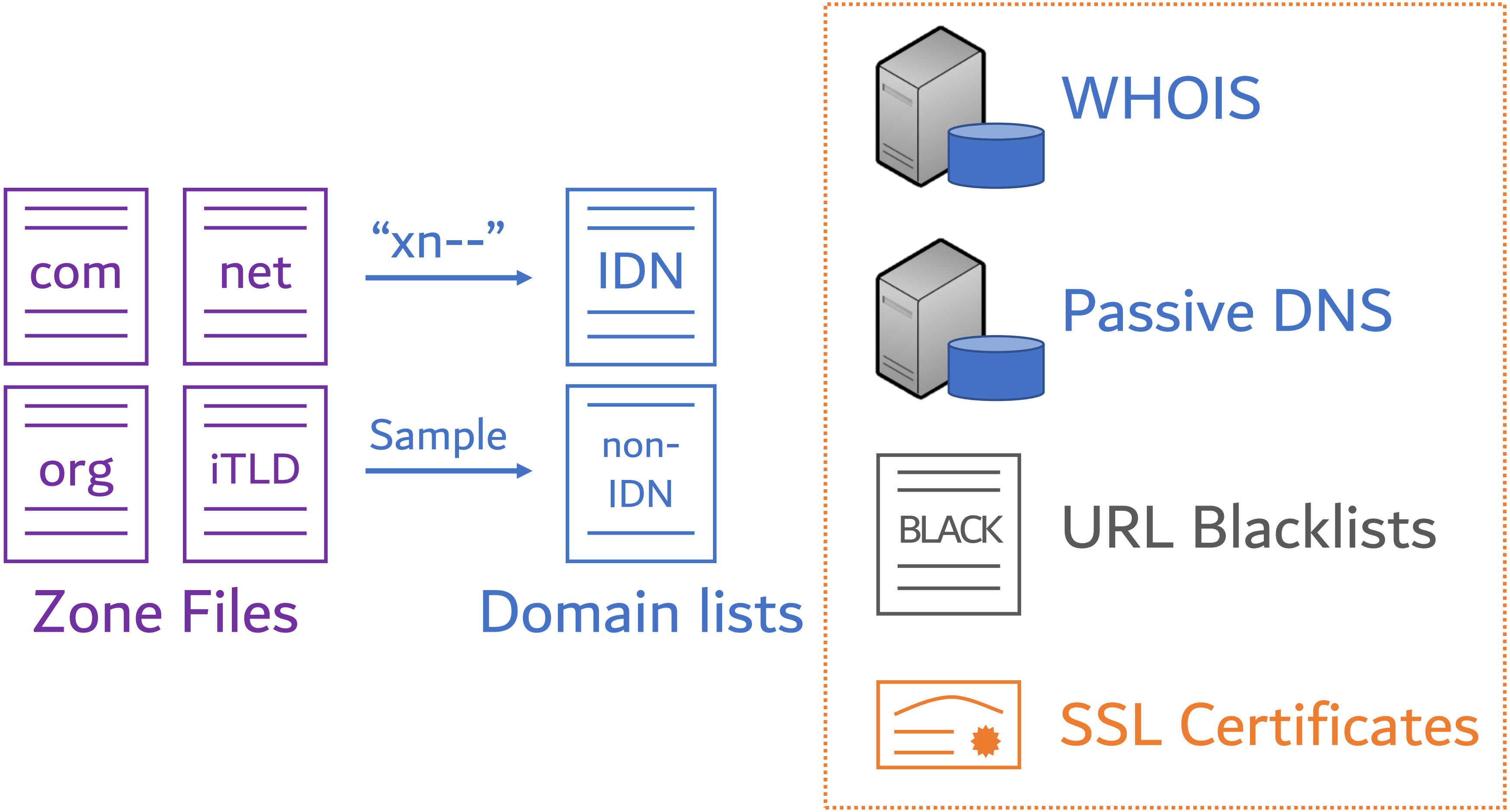
xn--ihqwcrb4cv8a8dqg056pqjye

Can be used in ASCII-only DNS

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



com

net

org

iTLD

Zone Files

"xn--"

Sample

IDN

non-IDN

Domain lists

WHOIS

Passive DNS

BLACK

URL Blacklists

SSL Certificates

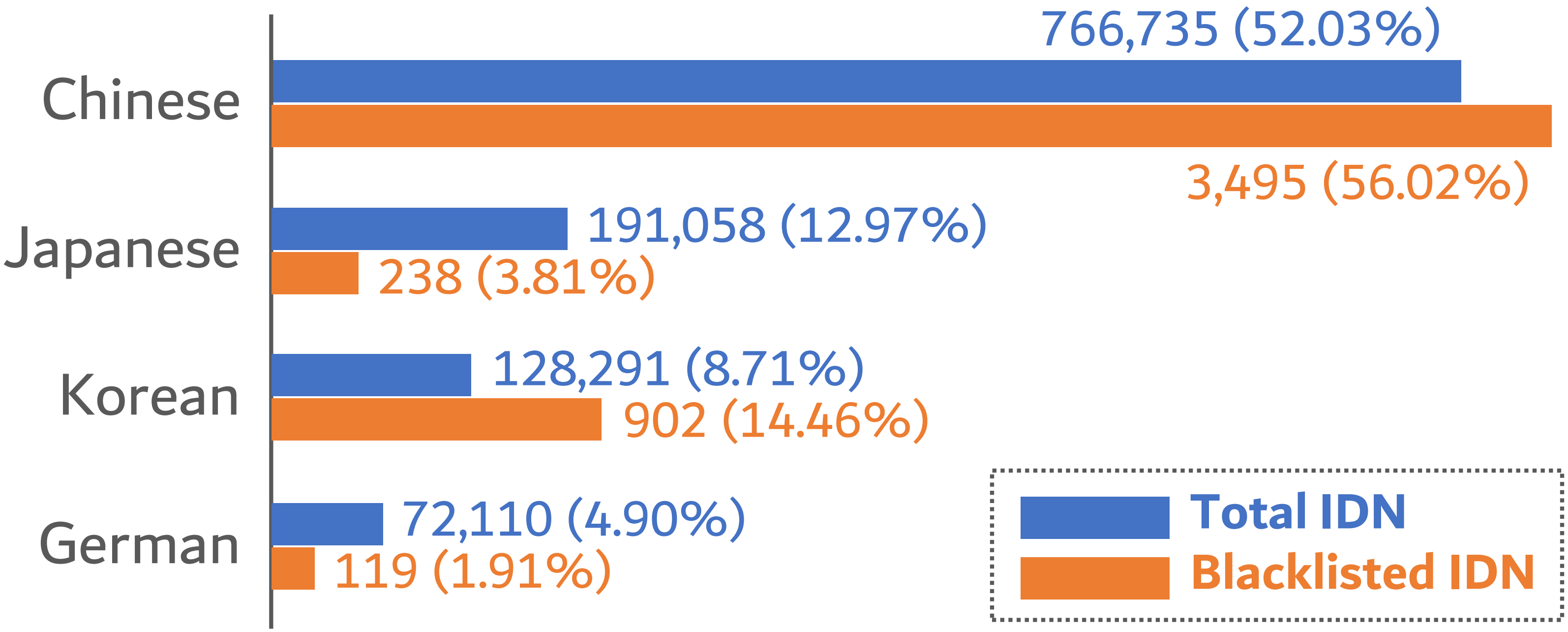
Dataset Collection

- Collected dataset

TLD	Snapshot on	# IDN (SLD)	WHOIS	Blacklisted
com	Sept 21, 2017	1,007,148	590,542	5,284
net	Sept 21, 2017	231,896	131,573	746
org	Oct 5, 2017	25,629	19,271	59
iTLD (53)	Oct 5, 2017	208,163	2,226	152
Total	**-**	**1,472,836**	**739,160**	**6,241**

IDN Characteristics

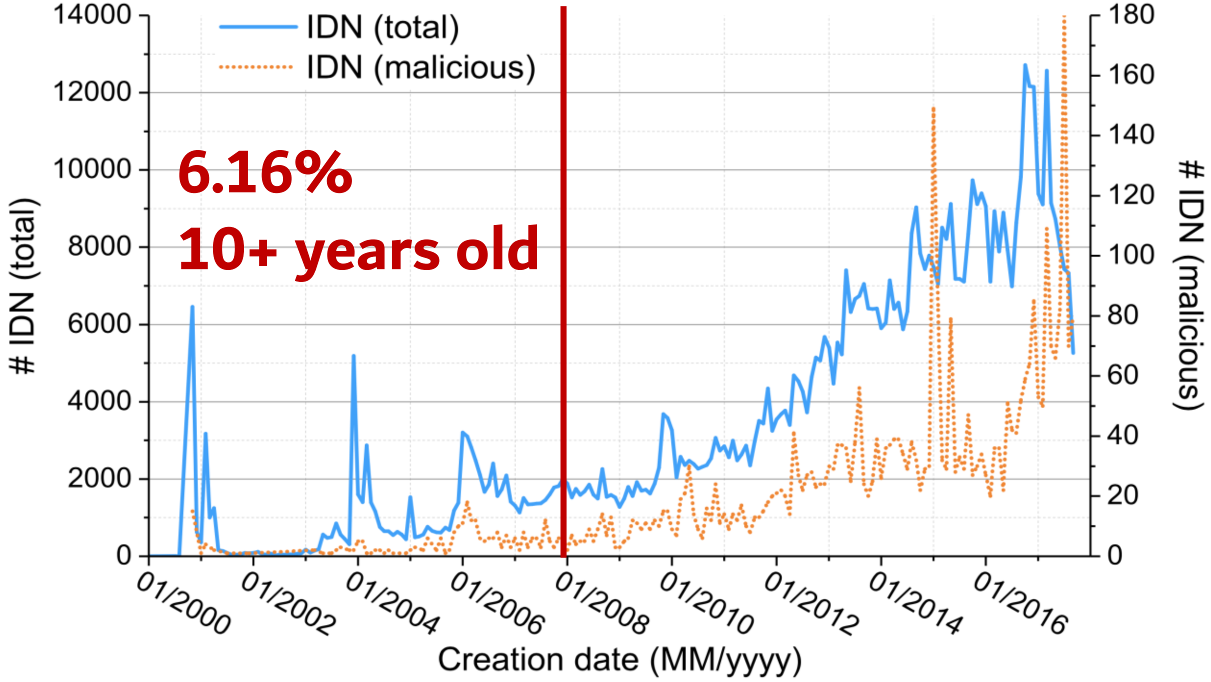
- A. Language
  - Using LangID[*] for language identification
  - **75%+** IDN are in languages of east Asian countries



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

IDN Characteristics

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



6.16%

10+ years old

IDN Characteristics

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

Email	# IDN	Remarks
776053229@qq.com	2,609	All are southwest city names in China.
daidesheng88@gmail.com	1,562	All are about online gambling.
tetetw@gmail.com	1,453	All are short words in Chinese.

Large-scale opportunistic registrations, of specific pattern / topic

IDN Characteristics

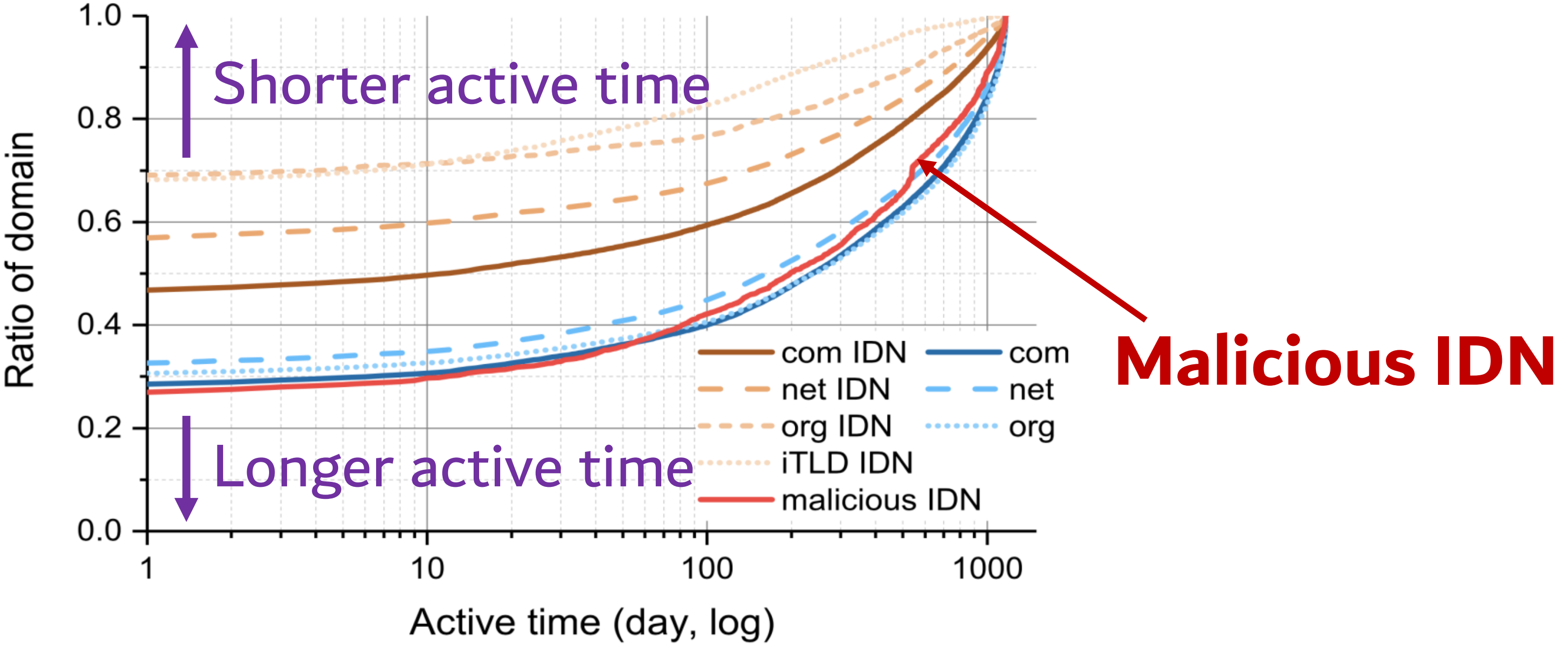
- B. Registration
  - Correlating with WHOIS data
  - Creation date
  - Registrant
  - Registrar (% registered IDN)

GMO	22.99% (JP)	gabia.	4.02% (KR)
万网 www.net.cn 阿里云旗下品牌	10.86% (CN)	GoDaddy™	1.88%

East Asian markets are more active.

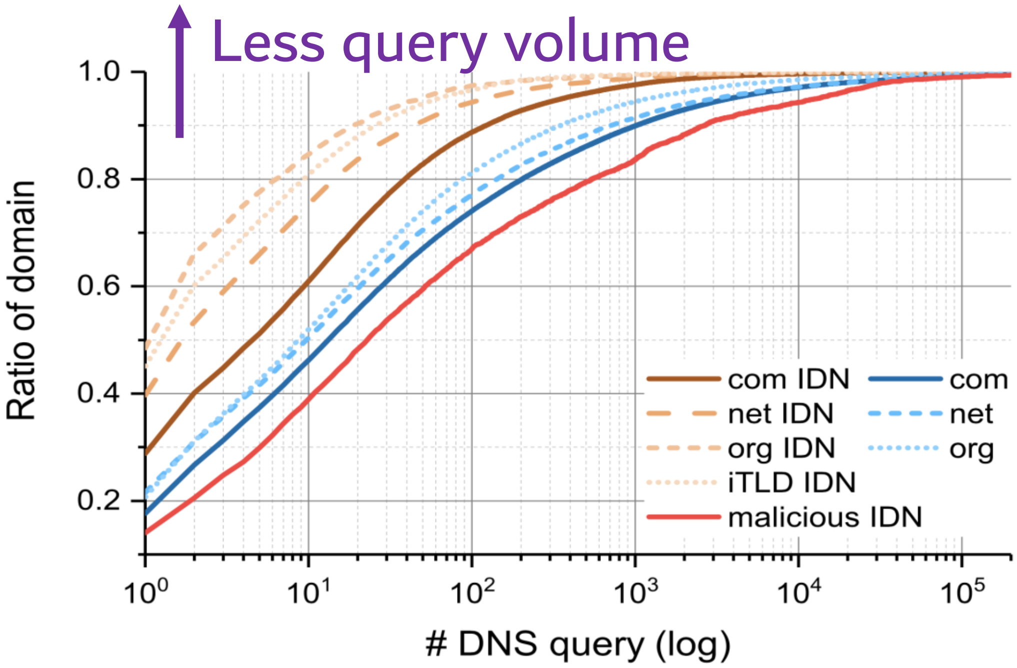
IDN Characteristics

- 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

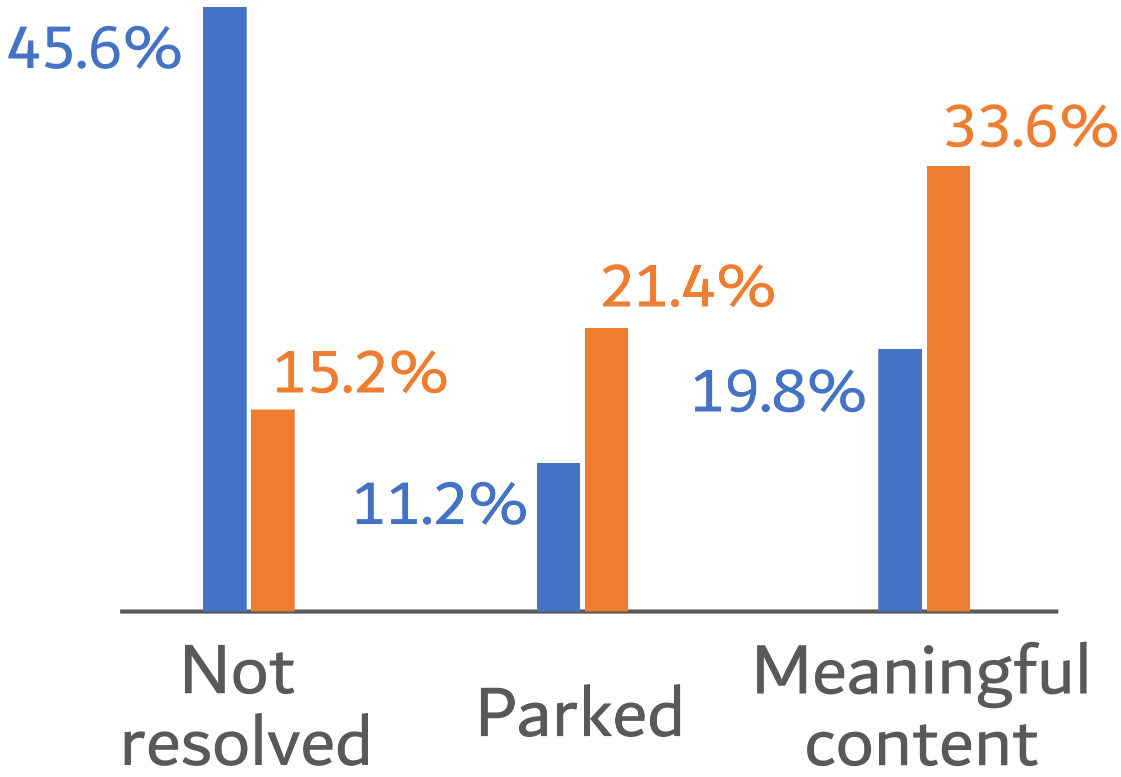


Less query volume

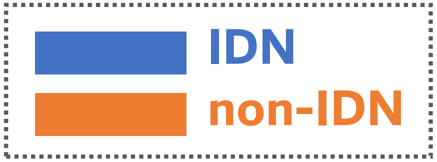
Malicious IDNs are effective at trapping users.

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.

Category	# IDN (% certificates)	# non-IDN (% certificates)
Expired	8,411 (12.5%)	8,730 (24.9%)
Invalid Authority	12,169 (18.1%)	5,801 (16.7%)
Invalid Common Name	**45,133 (67.3%)**	**19,527 (45.5%)**

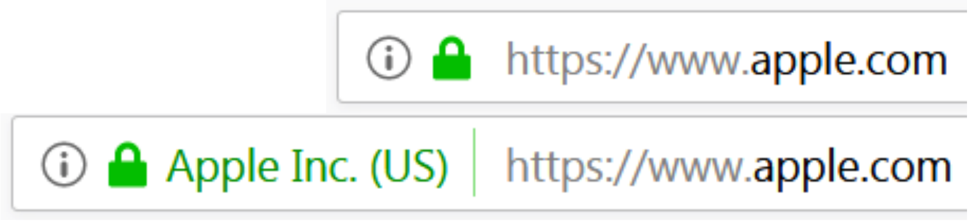
[*] Analysis of the HTTPS certificate ecosystem. IMC 2013

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



ⓘ 🔒 https://www.apple.com

ⓘ 🔒 Apple Inc. (US) | https://www.apple.com

- Semantic attack
  - Type-1: brand name + keyword

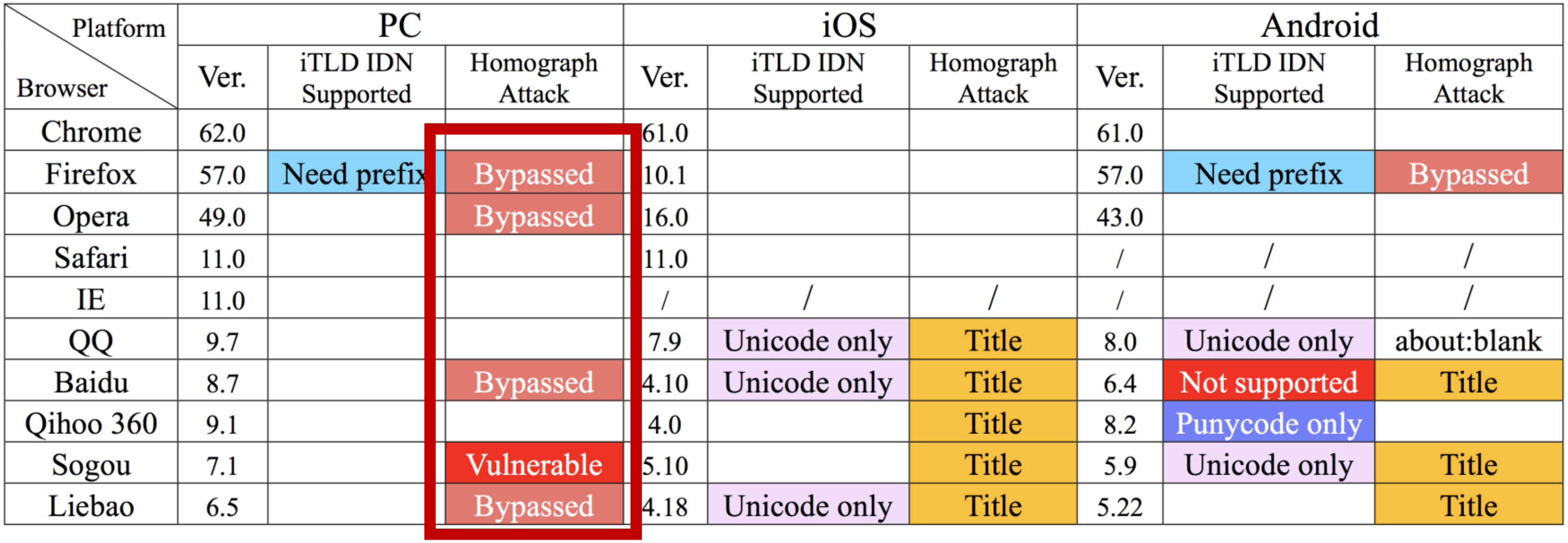
    icloud登录.com apple邮箱.com

  - 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**



Platform / Browser	PC			iOS			Android		
	Ver.	iTLD IDN Supported	Homograph Attack	Ver.	iTLD IDN Supported	Homograph Attack	Ver.	iTLD IDN Supported	Homograph Attack
Chrome	62.0			61.0			61.0		
Firefox	57.0	Need prefix	Bypassed	10.1			57.0	Need prefix	Bypassed
Opera	49.0		Bypassed	16.0			43.0		
Safari	11.0			11.0			/	/	/
IE	11.0			/	/	/	/	/	/
QQ	9.7			7.9	Unicode only	Title	8.0	Unicode only	about:blank
Baidu	8.7		Bypassed	4.10	Unicode only	Title	6.4	Not supported	Title
Qihoo 360	9.1			4.0		Title	8.2	Punycode only	
Sogou	7.1		Vulnerable	5.10		Title	5.9	Unicode only	Title
Liebao	6.5		Bypassed	4.18	Unicode only	Title	5.22		Title

Homograph Attack

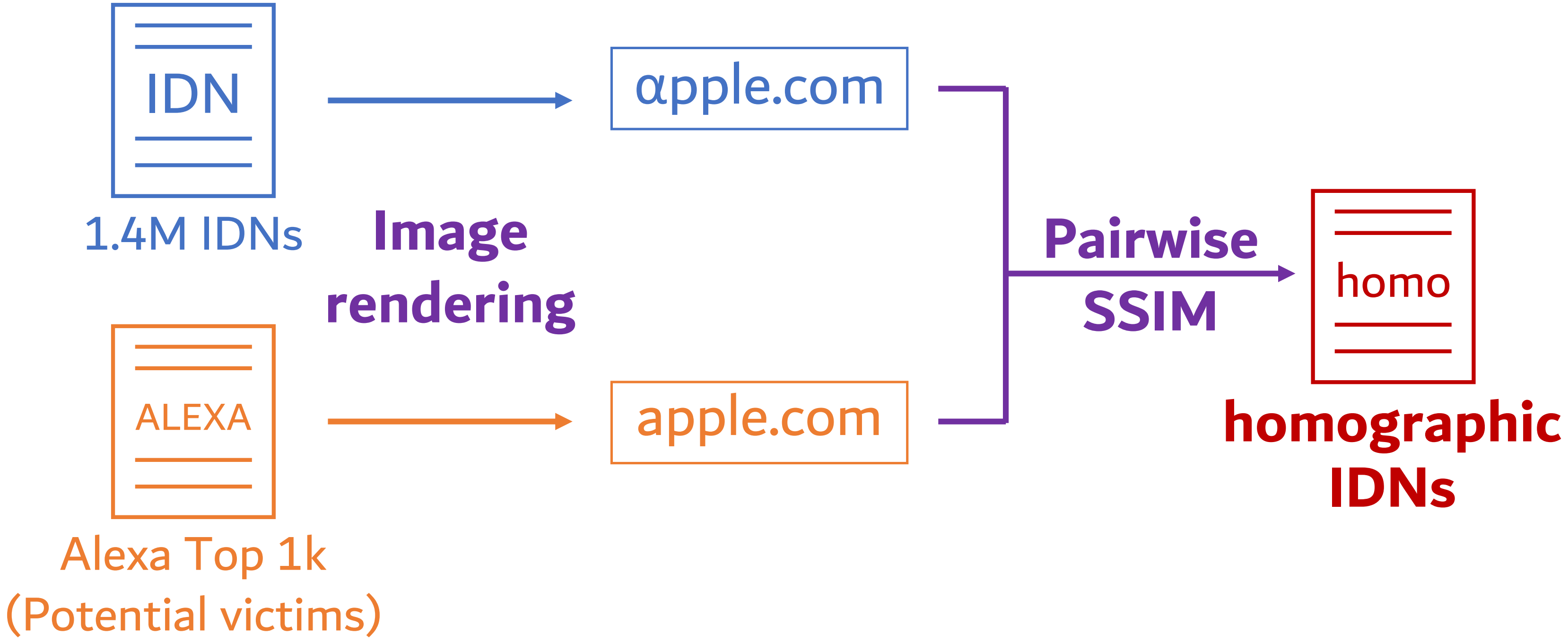
- A. Browser policies
  - RFC3490 (IDNA): avoid exposing raw ACE encoding
  - Firefox & Chrome: display based on **character sets**
  - Manual survey

Input	Display
apple.com (xn--80ak6aa92e.com) Only the ‘l’ is Cyrillic.	**Punycode**
soso.com (xn--n1aa1eb.com) ALL characters in the SLD are Cyrillic.	**Unicode**

Some up-to-date policies still need to be revised.

Homograph Attack

- B. Detecting homographic IDNs
  - SSIM index*: a metric of visual resemblance



[*] Image quality assessment: From error visibility to structural similarity. IEEE TIP 2004.

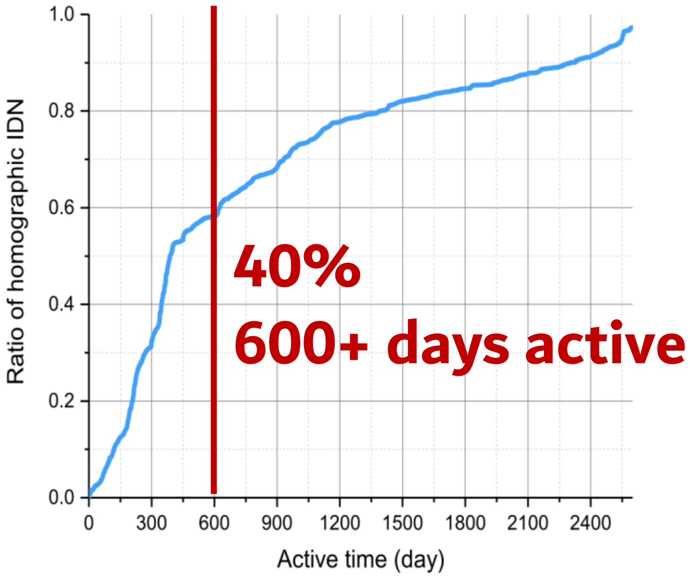
Homograph Attack

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

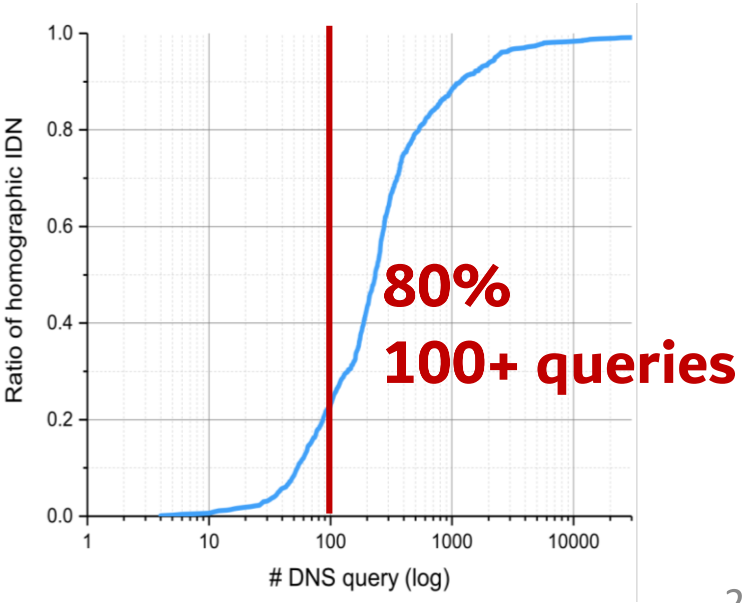
Brand Domain	# Homographic IDN (% of 1,516)	# Defensive Registration
google.com	121 (8.0%)	19
facebook.com	98 (6.5%)	0
amazon.com	55 (3.6%)	14
icloud.com	42 (2.8%)	0
youtube.com	41 (2.7%)	0

Homograph Attack

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



40%
600+ days active



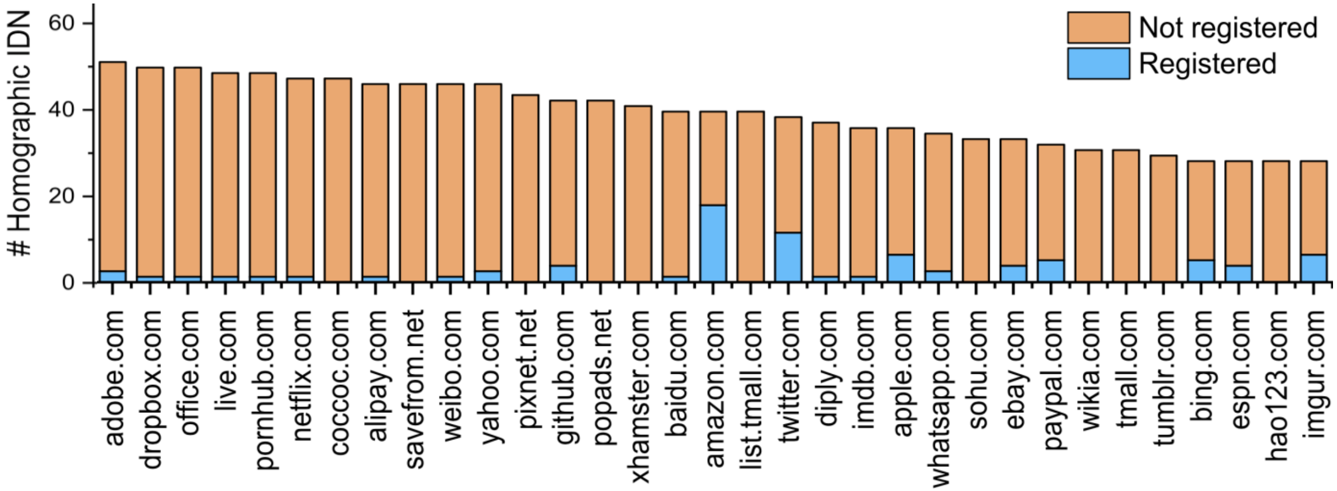
80%
100+ queries

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

Homograph Attack

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



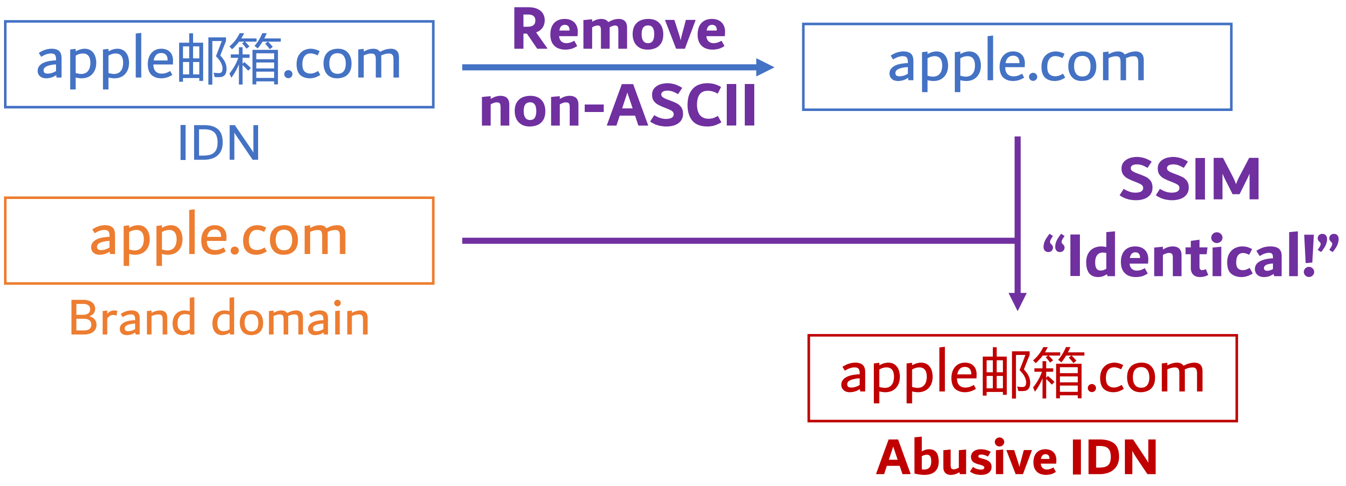
[*] The methodology and an application to fight against unicode attacks. SOUPS 2006

Homograph Attack

- 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

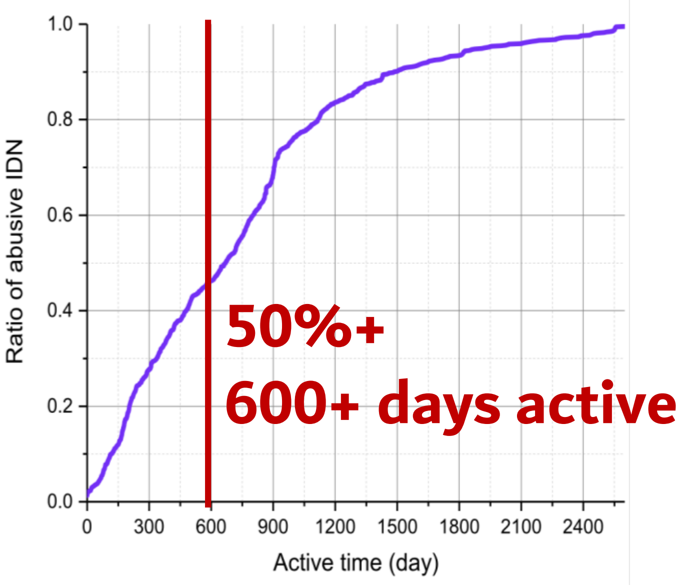
Semantic Attack

- 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

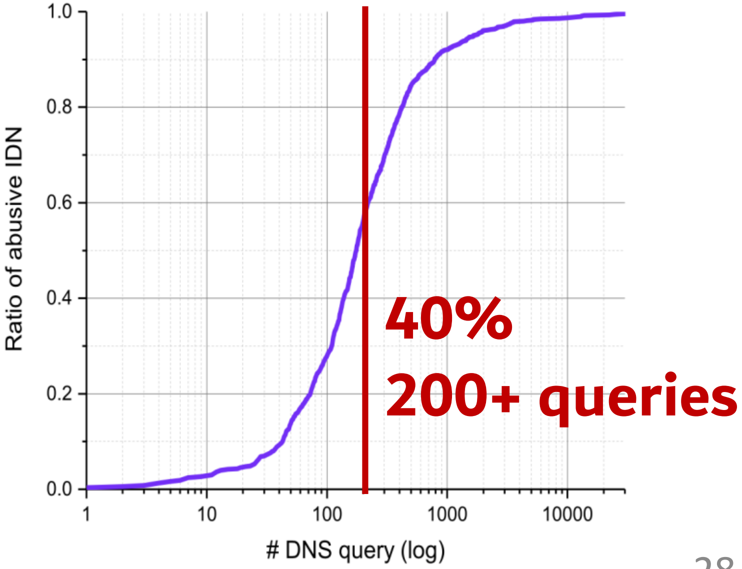


Semantic Attack

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



50%+
600+ days active



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?