The 21st Privacy Enhancing Technologies Symposium
July 12–16, 2021

Domain Name Encryption Is Not Enough: Privacy Leakage via IP-based Website Fingerprinting

Nguyen Phong Hoang, Arian Akhavan Niaki,
Phillipa Gill, Michalis Polychronakis

Stony Brook University
The State University of New York

UMASS AMHERST

Online surveillance is prevalent

FINANCIAL TIMES

China's tech workers pushed to their limits by surveillance software

Vicious cycle of monitoring and overwork is fuelling productivity — and a backlash

Tech underclass: rapid technological development, paired with poor labour regulations, has created the potential for labour abuse © Michael Tsang

Nikki Sun, Nikkei staff writer JUNE 15 2021

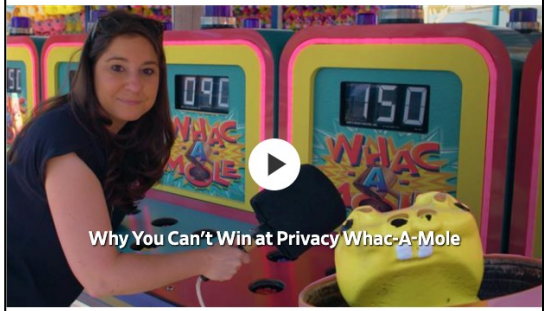
Andy Wang, an IT engineer at a Shanghai-based gaming company, occasionally felt a pang of guilt about his job.

Most of his hours were spent on a piece of surveillance software called DiSanZhiYan, or "Third Eye". The system was installed on the laptop of every colleague at his company to track their screens in real time, recording their chats, their browsing activity and every document edit they made.

THE WALL STREET JOURNAL.

Internet Providers Look to Cash In on Your Web Habits

Broadband operators mine customers' internet-use data, which is valuable to advertisers



Despite new initiatives from Google and Facebook, messing with privacy controls is like playing a carnival game. Knock out one way for advertisers to track you, and they quickly find another way to do it. WSJ's Joanna Stern heads to Coney Island to explain. Photo: Kenny Wassus

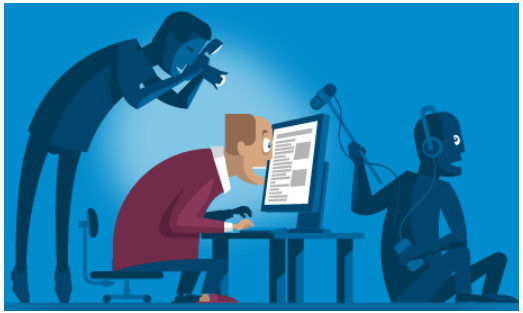
By Sarah Krouse and Patience Haggin
June 27, 2019 5:30 am ET

Internet providers know a lot about what their customers do on the web, including the news sites they read, health ailments they research and entertainment services they use. They often know where those customers shop and manage their finances, too.

Now, they are deciding whether to use that information to sell ads. Some industry titans are being more aggressive than others, even as regulators are pressuring Silicon Valley companies and broadband providers to explain how they use customer data.

Violation of human rights

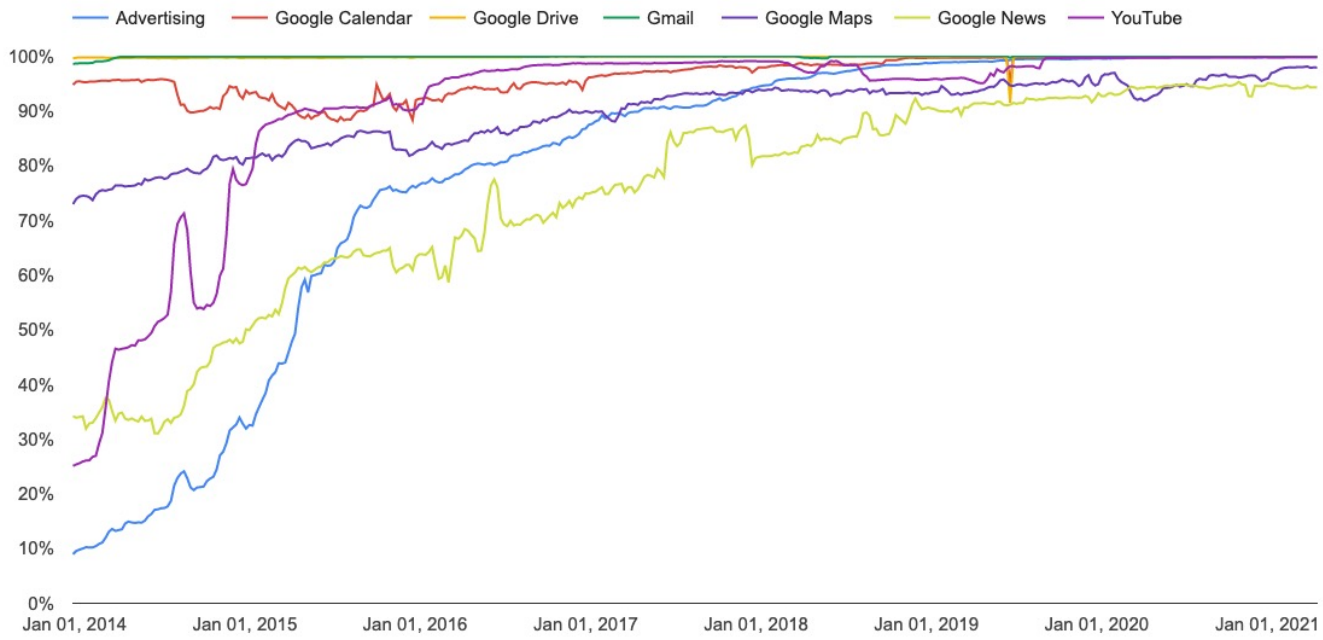
Online surveillance has led to severe violations of many human rights, including the right to privacy



Internet traffic encryption is on the rise

Encryption by product at Google

This chart provides a snapshot of encrypted traffic for several products. Numbers are based on the majority of Google traffic for a given product. We continue to work through the technical barriers that make it difficult to support encryption on some of our products. This chart will change over time to reflect product developments.



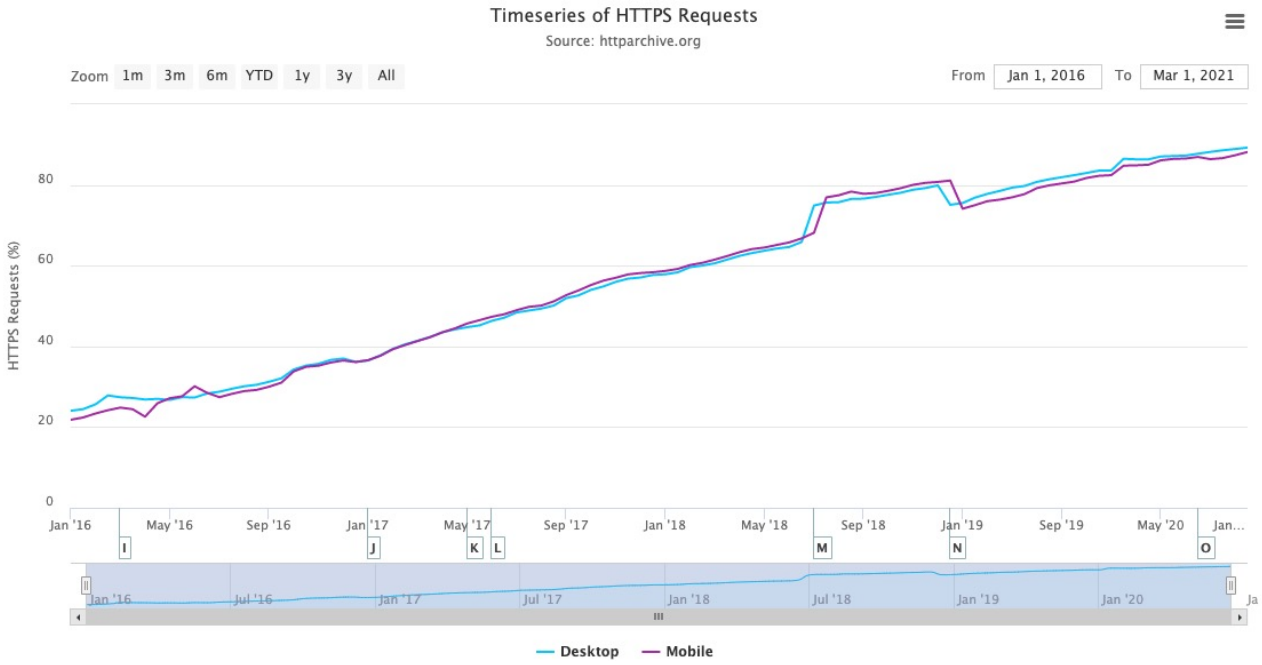
https://transparencyreport.google.com/https/overview

Internet traffic encryption is on the rise

HTTPS Requests

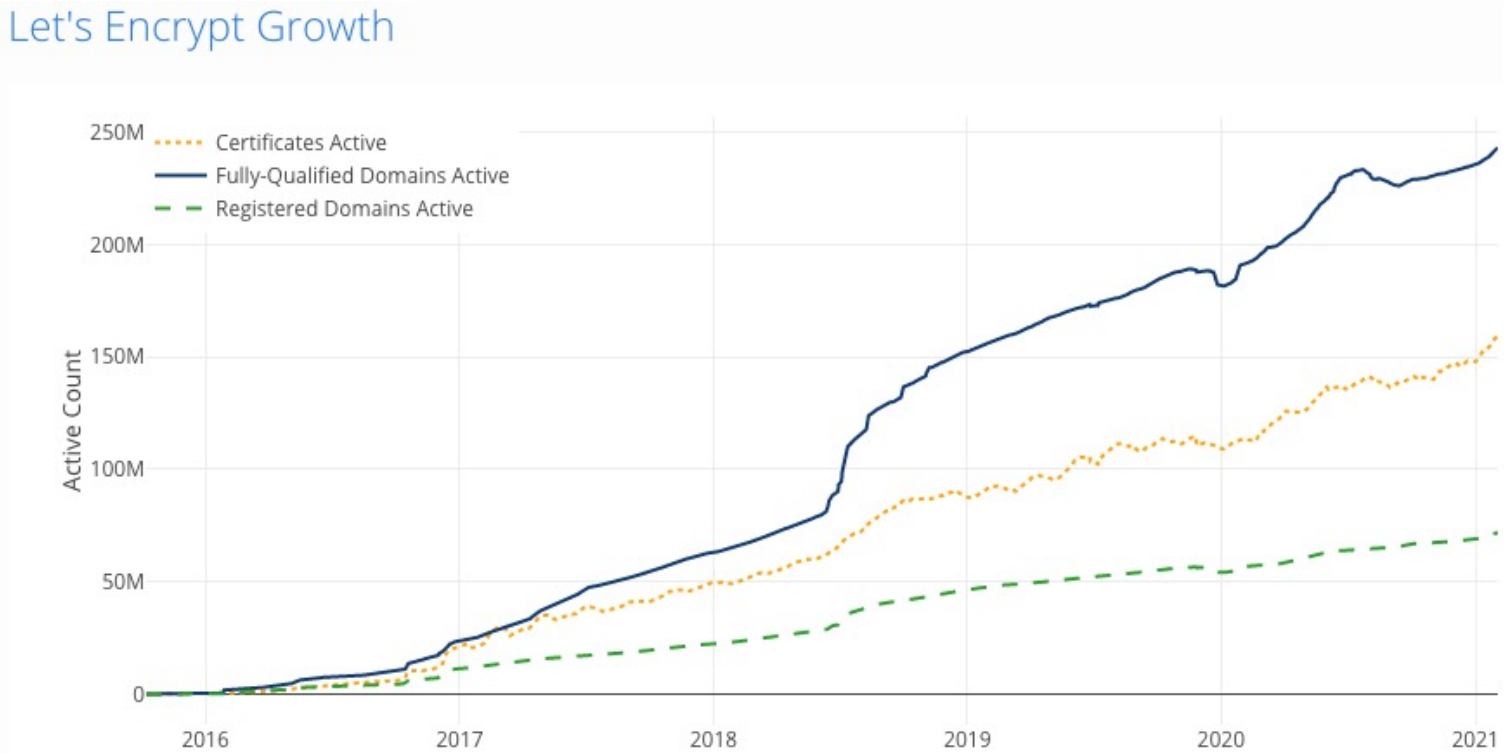
The percent of all requests in the crawl whose URLs are prefixed with `https`.

DESKTOP	MOBILE
89.3%	88.2%
▲272.1%	▲304.6%



https://httparchive.org/reports/state-of-the-web#pctHttps

Internet traffic encryption is on the rise
Thanks to free TLS certificate authorities



https://letsencrypt.org/stats/

Plaintext domain names are the last piece of unencrypted information

DNS query/response packets

Source	Destination	Protocol	Info
192.168.50.194	1.1.1.3	DNS	Standard query 0x5ea5 A example.com
1.1.1.3	192.168.50.194	DNS	Standard query response 0x5ea5 A example.com A 93.184.216.34
192.168.50.194	93.184.216.34	TCP	64895 → 443 [SYN] Seq=3552478921 Win=65535 Len=0 MSS=1460 WS=
93.184.216.34	192.168.50.194	TCP	443 → 64895 [SYN, ACK] Seq=2027449269 Ack=3552478922 Win=6553
192.168.50.194	93.184.216.34	TCP	64895 → 443 [ACK] Seq=3552478922 Ack=2027449270 Win=131712 Le
192.168.50.194	93.184.216.34	TLS…	Client Hello
93.184.216.34	192.168.50.194	TCP	443 → 64895 [ACK] Seq=2027449270 Ack=3552479439 Win=67072 Len

```
▸ Compression Methods (1 method)
  Extensions Length: 403
▸ Extension: Reserved (GREASE) (len=0)
▾ Extension: server_name (len=16)
    Type: server_name (0)
    Length: 16
  ▾ Server Name Indication extension
      Server Name list length: 14
      Server Name Type: host_name (0)
      Server Name length: 11
      Server Name: example.com
```

TLS handshake's Client Hello
Server Name Indication (SNI)

→ Security and privacy issues

Domain names reveal semantic info

- amazon.com, walmart.com, ebay.com

  → online shopping preferences

- HIV.gov , Cancer.gov

  → health condition

- Islamicity.org, Quran.com, Bible.com

  → religion

- LGBT.foundation, Grindr.com

  → gender identity

- Xvideos.com, Pornhub.com

  → sexual habits

MOTHERBOARD
TECH BY VICE

Comcast Is Lobbying Against Encryption That Could Prevent it From Learning Your Browsing History

Motherboard has obtained a leaked presentation internet service providers are using to try and lobby lawmakers against a form of encrypted browsing data.

By Joseph Cox

23.10.19

Internet giant Comcast is lobbying U.S. lawmakers against plans to encrypt web traffic that would make it harder for internet service providers (ISPs) to determine your browsing history, according to a lobbying presentation obtained by Motherboard.

The plan, which Google intends to implement soon, would enforce the encryption of DNS data made using Chrome, meaning the sites you visit. Privacy activists have praised Google's move. But ISPs are pushing back as part of a wider lobbying effort against encrypted DNS, according to the presentation. Technologists and activists say this encryption would make it harder for ISPs to leverage data for things such as targeted advertising, as well as block some forms of censorship by authoritarian regimes.

Domain encryption: DoT/DoH & ESNI

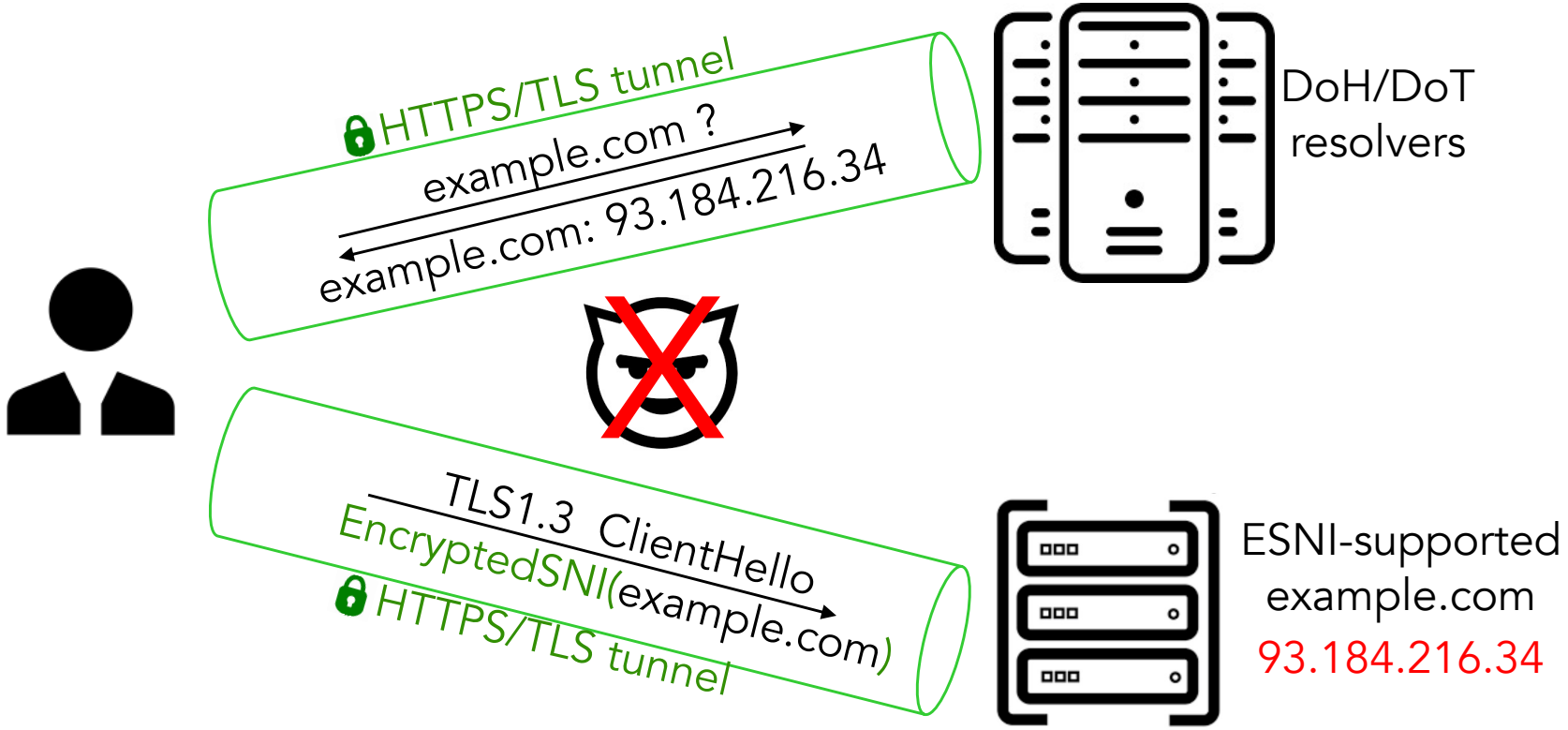
- **DoT:** DNS queries and responses are tunneled over TLS (RFC7858)

- **DoH:** DNS resolution is performed over HTTPS, inheriting all security benefits of the HTTPS protocol (RFC8484)

- **Encrypted SNI**: Starting from TLS1.3, the Server Name Indication extension in the Client Hello message during the TLS handshake can be *optionally* encrypted (RFC8744)

  → being reworked to Encrypted Client Hello (Internet draft)

Domain encryption: DoH/DoT and ESNI



Motivation

Domain name encryption → user privacy?

Investigate whether network-level browsing tracking *at scale* is still possible, given that destination IPs are visible to on-path observers

The extent to which domain inference can be made depends on:

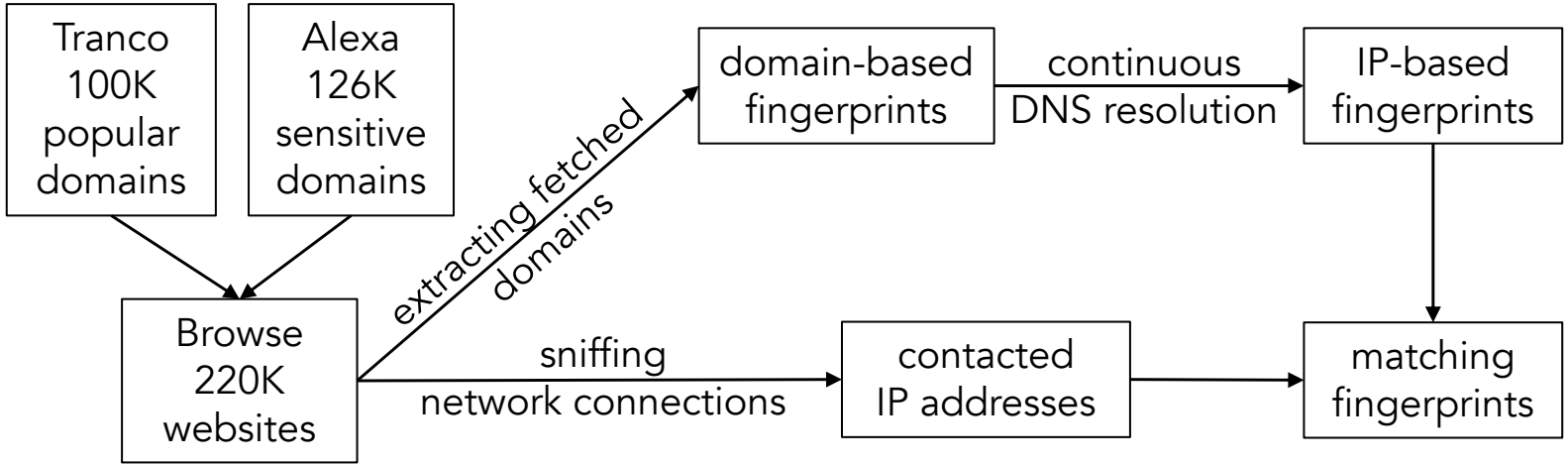
- Whether one or many domains are hosted on a given IP
- The stability of the mapping of a domain and its IP address(es)

Website Fingerprinting

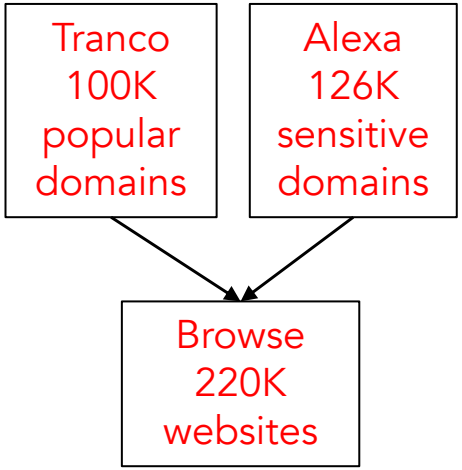
- Website fingerprinting (WF): a type of traffic analysis attack, based on unique traffic patterns (fingerprints)
- Fingerprints: constructed from network packets' visible metadata

→ We introduce a lightweight website fingerprinting (WF) technique that allows a network-level observer to identify with high accuracy the websites a user visits based on IP address information

Methodology



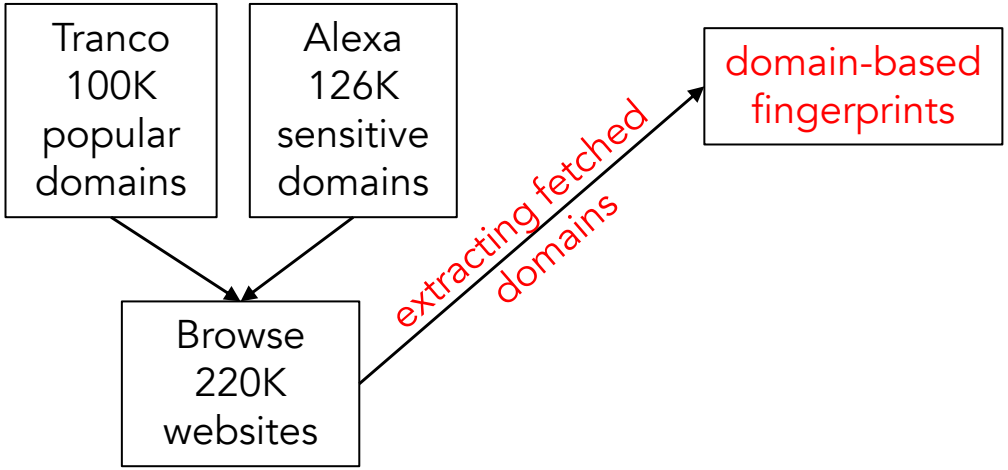
Selection of test domains



Website type	Total
All websites crawled	208,191
Popular websites	93,661
Sensitive websites	120,293
Sensitive and popular	5,763

The number of websites crawled is lower than the number of domains selected due to some unresponsive websites

Domain-based fingerprints

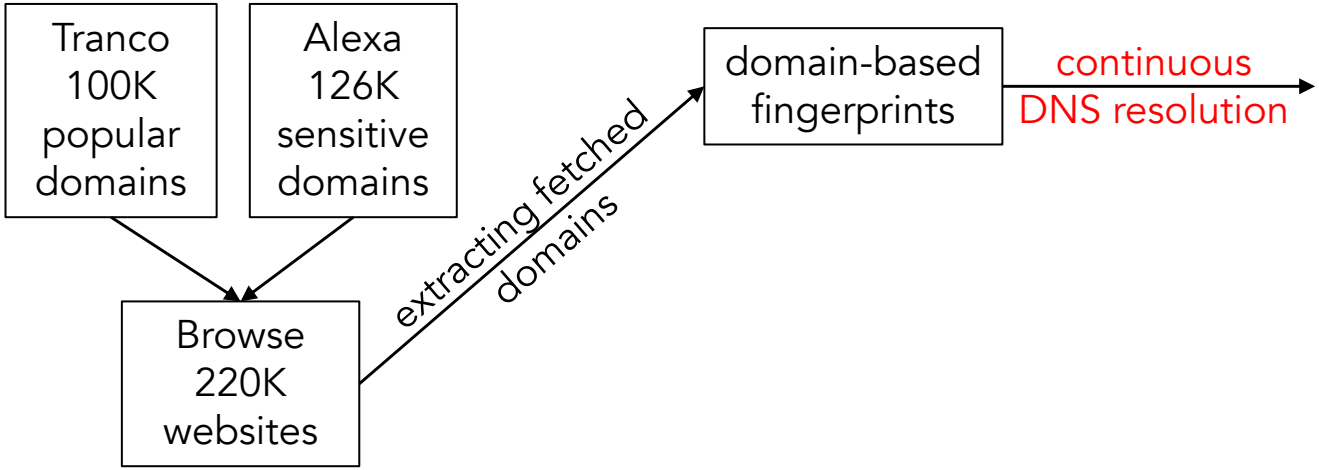



```
{'twitter.com':
    {'api.twitter.com', 'abs.twimg.com',
    'pbs.twimg.com', 'www.google-analytics.com'}
}
```

primary domain

secondary domains

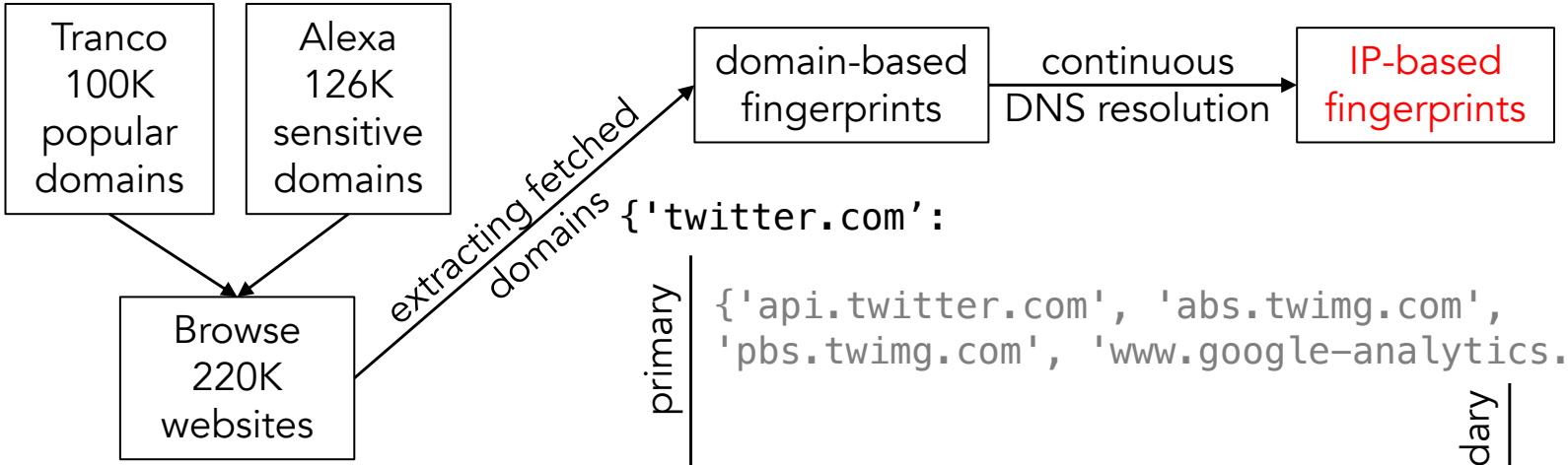
Domain-IP mapping




```
twitter.com;{1760832065, 1760832129, 1760832193, 1760832001} 4 IPs

api.twitter.com;{1760832194, 1760832002, 1760832066, 1760832130} 4 IPs
abs.twimg.com;{2540008607, 3353879711, 1209359174, 2540032159, …} 7 IPs
pbs.twimg.com;{1209359174, 2540042399, 2540008607, 3353886879, …} 10 IPs
www.google-analytics.com;{2899905678, 2899904206, 2899905038, …} 16 IPs
```

IP-based fingerpints




```
{'twitter.com':
{'api.twitter.com', 'abs.twimg.com',
'pbs.twimg.com', 'www.google-analytics.com'}}
```

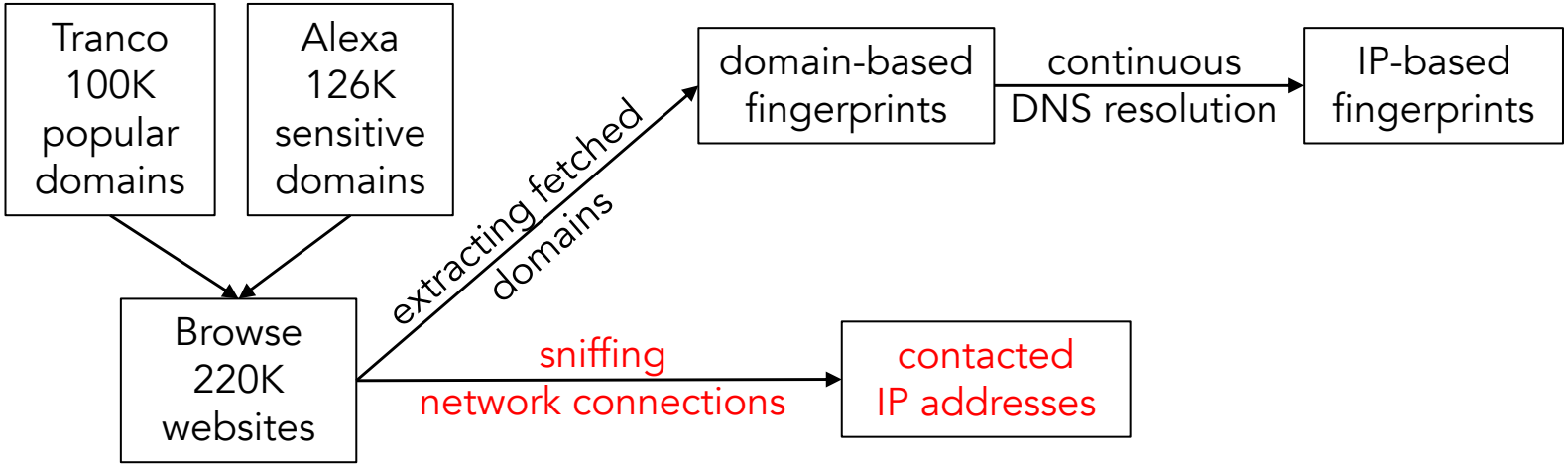
primary

secondary

```
{[1760832001, 1760832065, 1760832129, 1760832193];

[385967085, 385968877, 1209359174, 1760832002, 1760832066, 1760832130,
1760832194, 2540008607, 2540030111, 2540032159, 2540042399, 2899903342,
2899904206, 2899904238, 2899904270, 2899905006, 2899905038, 2899905070,
2899905102, 2899905134, 2899905166, 2899905262, 2899905294, 2899905326,
2899905678, 2899905710, 2899905742, 3089042157, 3236277520, 3353879711,
3353886879]}
```

Sniffing network connections

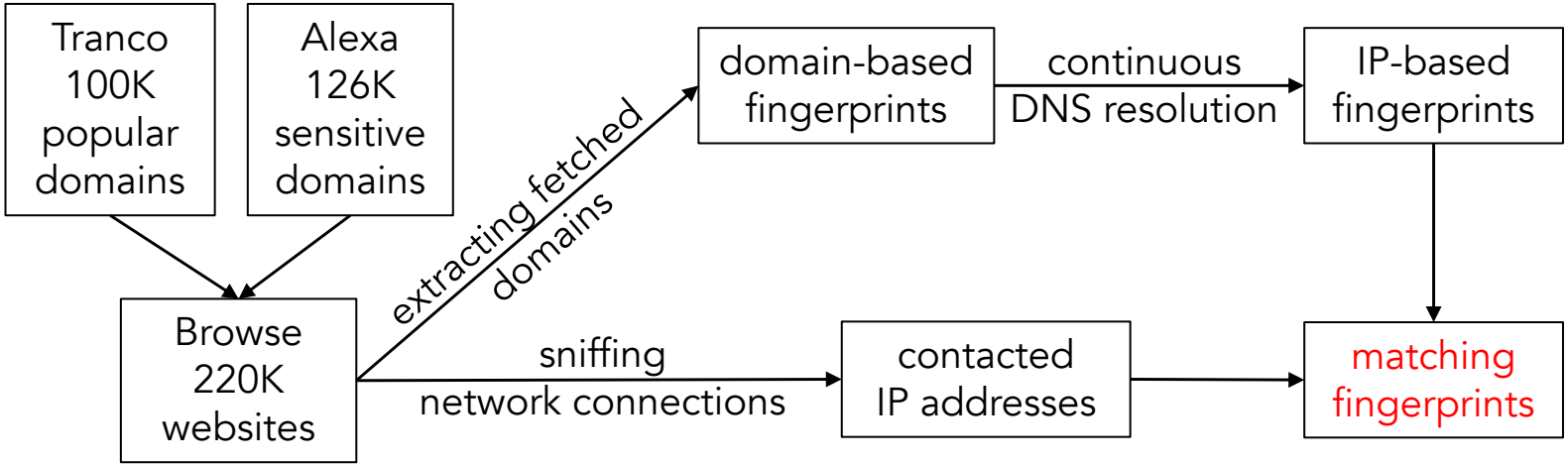



```
Sequence of unique IP addresses connected:
{1760832065, 1760832002, 1209359174, 2899904270}
```

primary IP

secondary IPs

Matching fingerprints




```
Sequence of IPs: {1760832065, 1760832002, 1209359174, 2899904270}
```

```
twitter.com;{[1760832001, 1760832065, 1760832129, 1760832193];[385967085,
385968877, 1209359174, 1760832002, 1760832066, 1760832130, 1760832194,
2540008607, 2540030111, 2540032159, 2540042399, 2899903342, 2899904206,
2899904238, 2899904270, 2899905006, 2899905038, 2899905070, 2899905102,
2899905134, 2899905166, 2899905262, 2899905294, 2899905326, 2899905678,
2899905710, 2899905742, 3089042157, 3236277520, 3353879711, 3353886879]}
```

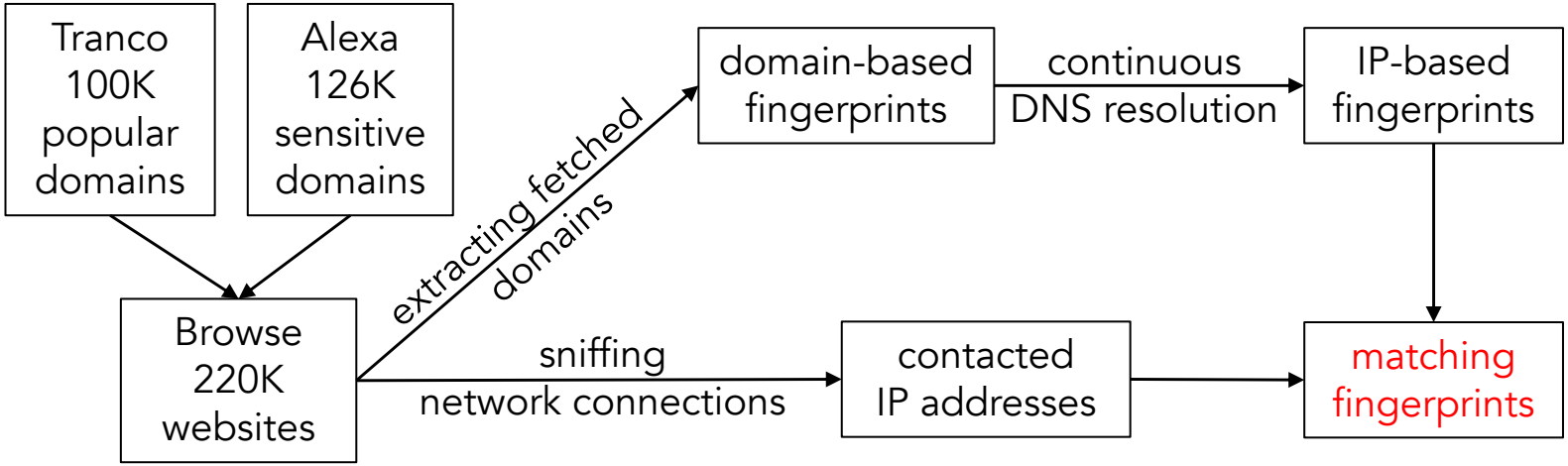
Single-hosted primary domains

- twitter.com;{1760832065, 1760832129, 1760832193, 1760832001}
- hrw.org;{1224469683}
- grindr.com;{583195696, 65110341, 885721358}
- xvideos.com;{3109598466, 3109598467, 3109598468, 3109598469, 3109598470, 3109598471, 3109598472, 3109598473, 3109598474, 3109598475}

When a primary domain is single-hosted on one IP or multiple IPs without sharing its hosting server(s) with any other domains, it is straightforward to infer the website being visited

Assessing the Privacy Benefits of Domain Name Encryption, *ACM AsiaCCS 2020.*

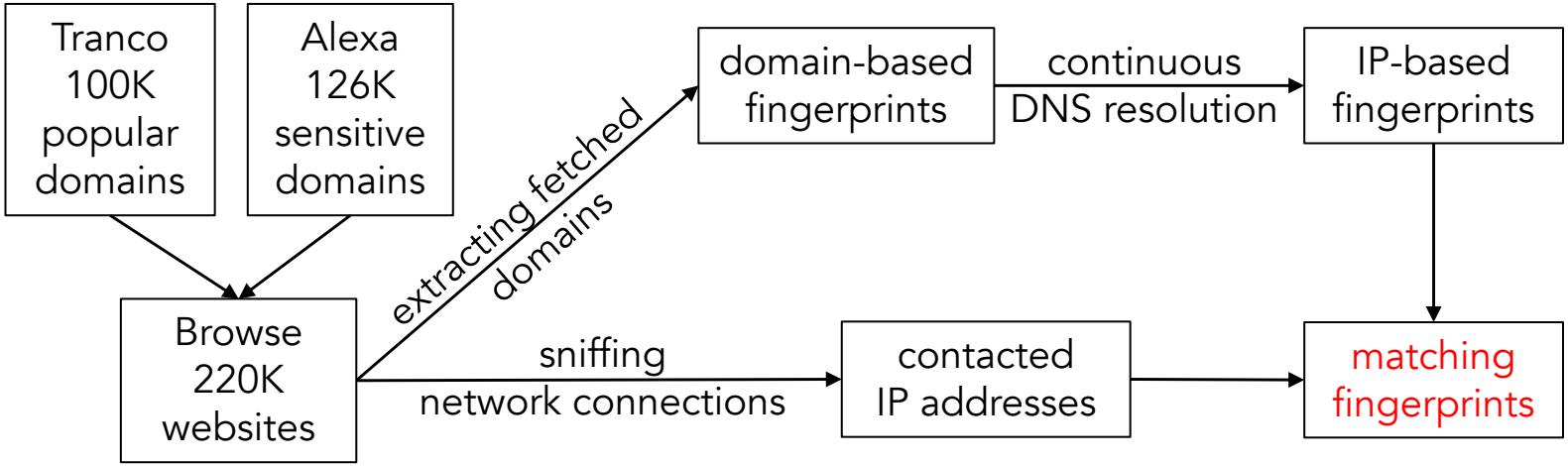
Matching fingerprints




```
Sequence of IPs: {1760832065, 1760832002, 1209359174, 2899904270}
```

```
twitter.com;{[1760832001, 1760832065, 1760832129, 1760832193];[385967085,
385968877, 1209359174, 1760832002, 1760832066, 1760832130, 1760832194,
2540008607, 2540030111, 2540032159, 2540042399, 2899903342, 2899904206,
2899904238, 2899904270, 2899905006, 2899905038, 2899905070, 2899905102,
2899905134, 2899905166, 2899905262, 2899905294, 2899905326, 2899905678,
2899905710, 2899905742, 3089042157, 3236277520, 3353879711, 3353886879]}
```

Matching fingerprints




```
Sequences of IPs: {1760832065,                                    }
```

```
twitter.com;{[1760832001, 1760832065, 1760832129, 1760832193];
                                                                  }
```

Analysis result
Single-hosted primary domains

Website type	Total	Primary Domain	IP-based Fingerprinting	Connection Bucketing
All websites crawled	208,191	107,455 (52%)	174,662 (84%)	**189,527 (91%)**
Popular websites	93,661	58,989 (63%)	86,147 (92%)	90,231 (96%)
Sensitive websites	120,293	51,538 (43%)	93,988 (78%)	104,983 (87%)
Sensitive and popular	5,763	3,072 (53%)	5,473 (95%)	5,687 (99%)

52% of the websites studied have their primary domain hosted on their own IP(s) → an adversary could already infer 52% of the targeted websites based solely on the IP address of the first connection to the primary domain, without having to consider secondary connections.

Analysis result
Basic IP-based fingerprinting

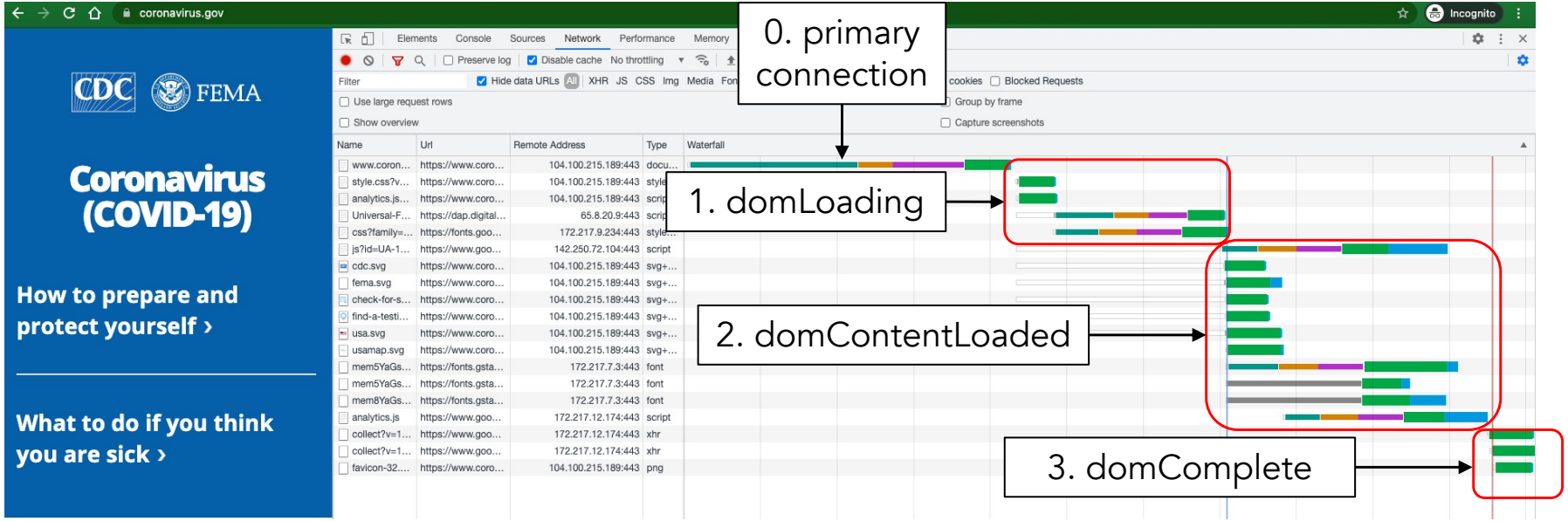
Website type	Total	Primary Domain	IP-based Fingerprinting	Connection Bucketing
All websites crawled	208,191	107,455 (52%)	174,662 (84%)	189,527 (91%)
Popular websites	93,661	58,989 (63%)	86,147 (92%)	90,231 (96%)
Sensitive websites	120,293	51,538 (43%)	93,988 (78%)	104,983 (87%)
Sensitive and popular	5,763	3,072 (53%)	5,473 (95%)	5,687 (99%)

Considering secondary connections → an increased accuracy of 84%

Of the fingerprinted websites, we could match 92% of the popular and 78% of the sensitive websites. More worrisome is the fact that 95% of sensitive and popular websites can be fingerprinted.

Enhanced IP-based fingerprinting

Viewing all requests as a whole → a high-level ordering relationship



Clustering connections → increased fingerprints' discriminatory

Enhanced IP-based fingerpints

Enhanced domain-based fingerprint

```
0: {'twitter.com'},
1: {'abs.twimg.com'},
2: {'api.twitter.com', 'abs.twimg.com', 'pbs.twimg.com'},
3: {'twitter.com', 'api.twitter.com', 'abs.twimg.com', 'www.google-analytics.com'}
```

Enhanced IP-based fingerprint

```
0: {1760832001, 1760832065, 1760832129, 1760832193};

1: {1209359174, 2540008607, 2540030111, 2540032159, 2540042399, 3236277520, 3353879711};

2: {385967085, 385968877, 1209359174, 1760832002, 1760832066, 1760832130, 1760832194,
2540008607, 2540030111, 2540032159, 2540042399, 3089042157, 3236277520, 3353879711,
3353886879};

3: {1209359174, 1760832001, 1760832002, 1760832065, 1760832066, 1760832129, 1760832130,
1760832193, 1760832194, 2540008607, 2540030111, 2540032159, 2540042399, 2899903342,
2899904206, 2899904238, 2899904270, 2899905006, 2899905038, 2899905070, 2899905102,
2899905134, 2899905166, 2899905262, 2899905294, 2899905326, 2899905678, 2899905710,
2899905742, 3236277520, 3353879711}
```

Clustered sequence of IPs from network trace, using K-means

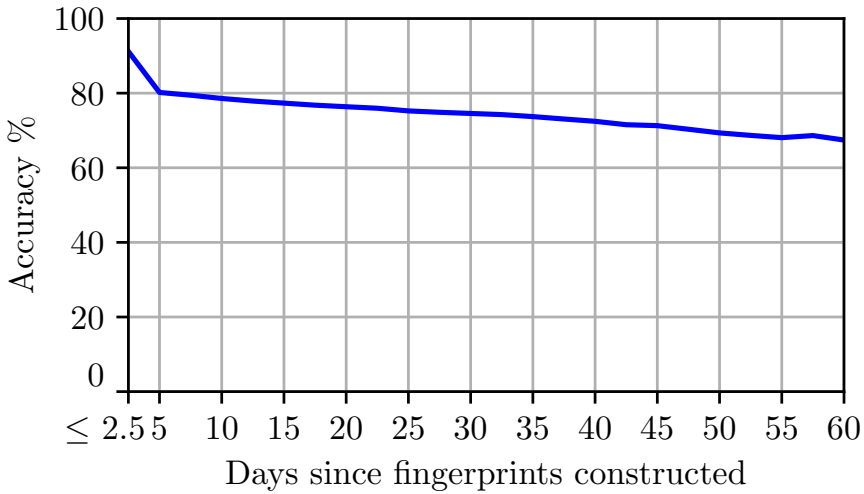
```
0: {1760832065},
1: {1209359174},
2: {1760832002, 1209359174},
3: {1760832065, 1760832002, 1209359174, 2899904270}
```

Analysis result
Enhanced IP-based fingerprinting

Website type	Total	Primary Domain	IP-based Fingerprinting	Connection Bucketing
All websites crawled	208,191	107,455 (52%)	174,662 (84%)	**189,527 (91%)**
Popular websites	93,661	58,989 (63%)	86,147 (92%)	90,231 (96%)
Sensitive websites	120,293	51,538 (43%)	93,988 (78%)	104,983 (87%)
Sensitive and popular	5,763	3,072 (53%)	5,473 (95%)	5,687 (99%)

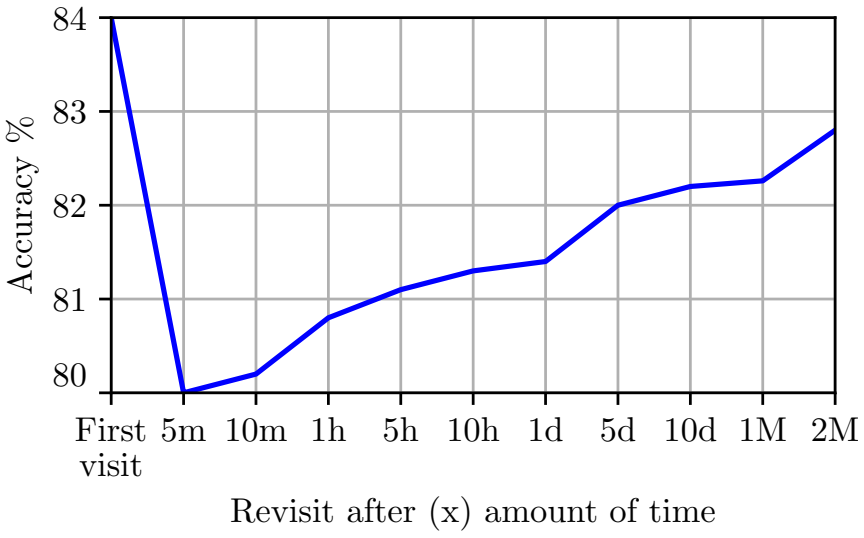
Enhanced fingerprinting improves the accuracy to 91%. For the popular and the sensitive websites, we obtain an accuracy of 96% and 87%, respectively. An alarming result: 99% of sensitive *and* popular websites can be precisely fingerprinted, posing a severe privacy risk.

Fingerprinting stability



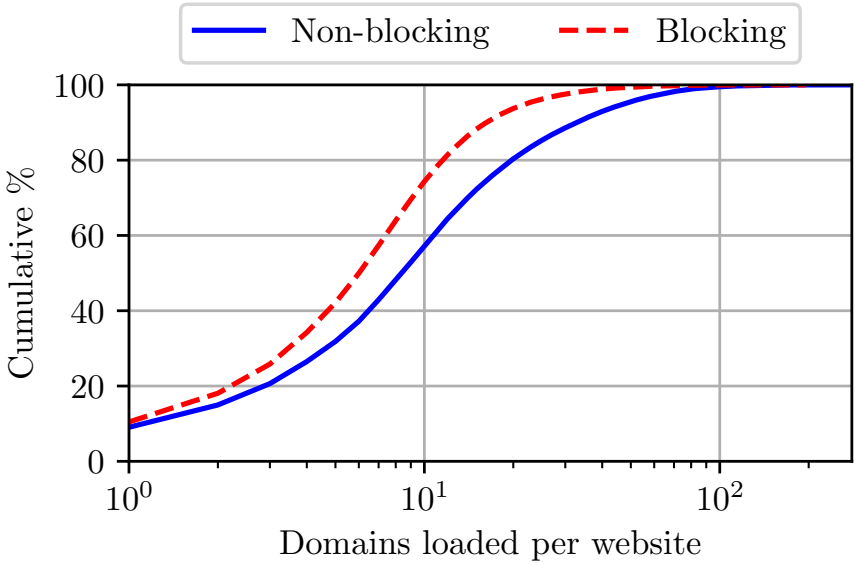
By conducting our measurement in a longitudinal manner for two months, we show that our enhanced IP-based fingerprints are still effective at correctly identifying about 70% of the tested websites.

Fingerprinting robustness
HTTP caching



Regardless of the impact of HTTP caching, an accuracy of 80% can still be obtained—a decrease of just 4% (from 84%) compared to when websites are visited for the first time.

Fingerprinting robustness
Ad blocking



When browsing with Brave, ad and tracking domains are blocked, leading to:

- changes in resource loading order
- fewer IP connections observed

Still obtain an accuracy of 80% when using the basic fingerprints, in which the ordering structure of loaded resources is not considered

Key takeaway

Regardless of the increasing trend of web co-location [*] and an idealistic future in which domain name encryption is universally adopted, network-level adversaries can still rely on destination IP addresses of contacted web servers for IP-based website fingerprinting to track users' browsing history at scale for the vast majority of websites.

Dataset is made available to stimulate future studies in this research domain at https://homepage.np-tokumei.net/publication/publication_2021_popets

[*] The Web is Still Small After More Than a Decade. SIGCOMM Computer Communication Review 2020.

Potential countermeasures

- **Full domain name confidentiality** must be preserved on both DNS and TLS channels; otherwise, neither technology can provide any actual privacy benefit if deployed individually
- **Domain owners** can seek providers that offer an increased co-hosting ratio per IP and/or highly dynamic domain-IP mappings
- **Hosting providers** can help to increase the co-hosting degree by grouping more websites under the same IP and dynamically rotate domain-IP mappings to hinder straightforward IP-based fingerprinting and further improve privacy