Web Privacy in the Age of Big Data

Martino Trevisan

SmartData@PoliTO Workshop

30 Jan 2020

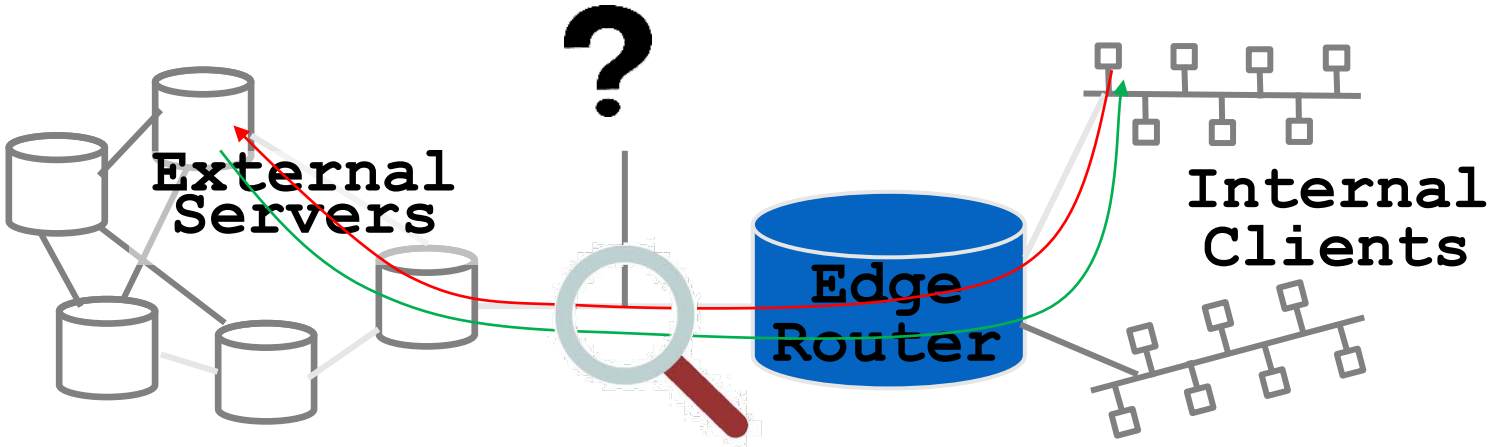
Outline

- What is still visible to the network?
- Can we hide our identity?
- Can we hide the websites we visit?

What is still visible to the network?

Network Monitoring

- Observe (and understand) traffic that flows in the network
  - And eventually take actions: route / block / account
- Performed by:
  - Routers > Traffic management, accounting…
  - Firewalls > Security
  - Network Probes > Knowledge Extraction, troubleshooting

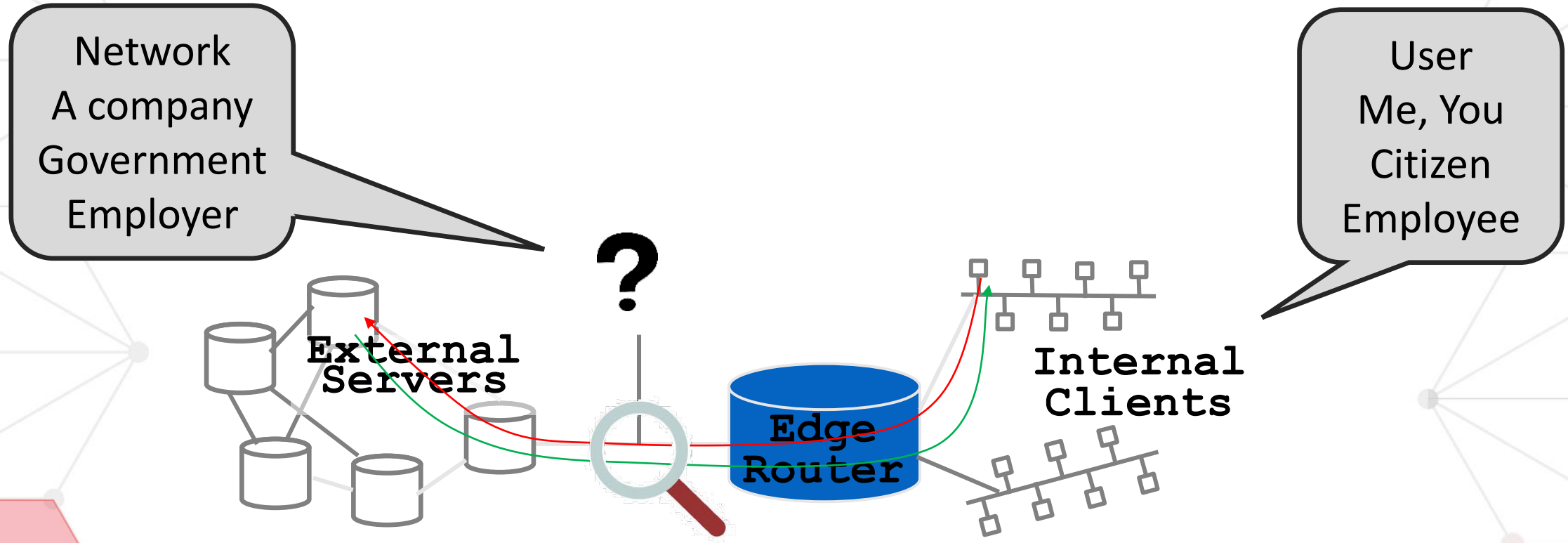


Privacy is a must!

Personal Information travels in the network

Users want privacy

Traffic is going encrypted to prevent the network from eavesdropping users' traffic



The history of encryption

The trend is from less encryption to more encryption

Three chapters of the history:

- Until ≈ 2010: **No encryption** -> Everything was visible
  - The URLs you visit
  - Your emails and social messages
  - Your Credit Card Number

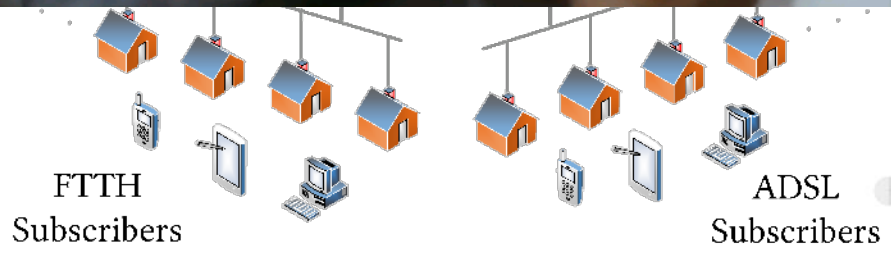
- ≈ 2010 - 2019: Deployment of **HTTPS**: Payload is encrypted
  - Only the name of the website is visible
    - Through DNS and HTTPS non-encrypted headers

- From 2020: **Signalling (e.g., DNS) is encrypted**
  - No information at all
    - Except for the server address (cannot encrypt!)

Can big data break your privacy?

With Big Data, an attacker can:

- Collect and process large datasets of network traffic
- Train ML models on big data
- Use these models to break users' privacy
  - Identify users changing their identifiers
  - Unveil the visited websites even under encryption

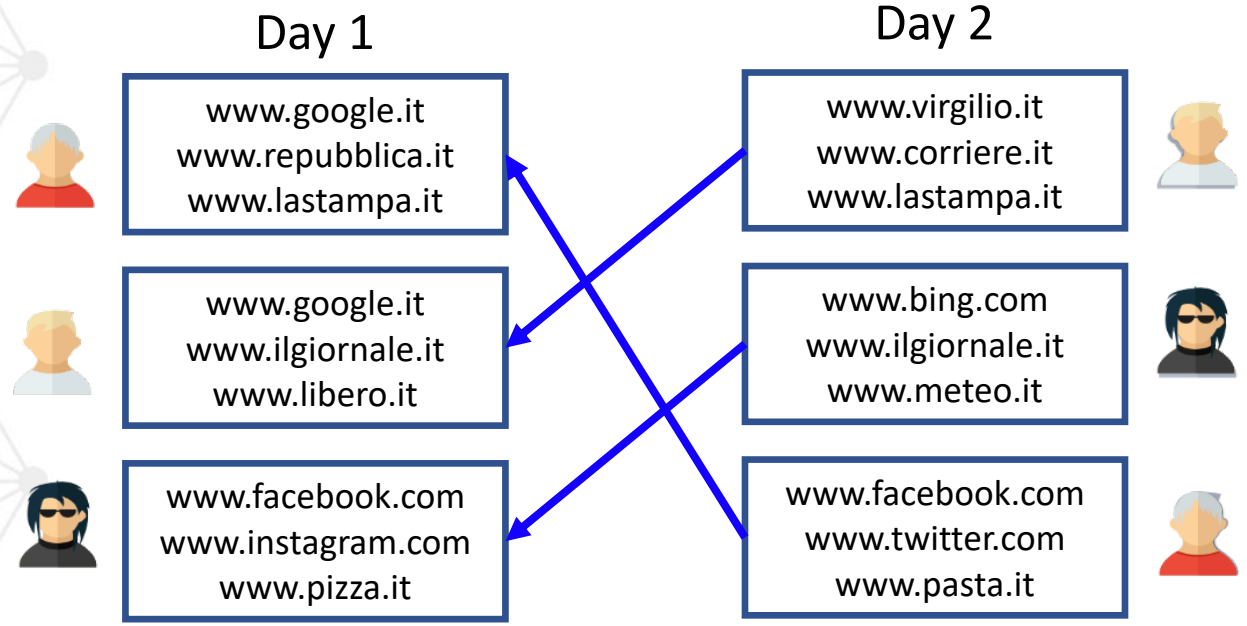


Can we hide our identity?

Scenario

Question: can the network re-identify us based on the websites we visit?

Scenario: the network can collect the list of websites we visit (second scenario)



Fingerprint similarity computation

Create **profiles** for users:

- A profile is the **set** of contacted **websites**

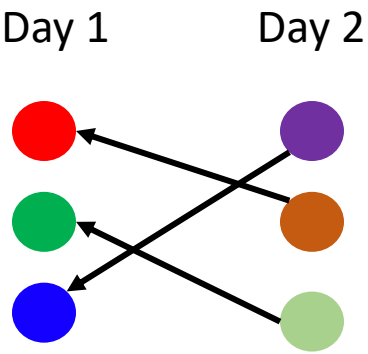
Hypothesis: users stay similar (correlation between different time windows)!

Goal: correctly identify a user among the profiles built in the **past**

Challenge: compute a suitable similarity metric

Three methodologies for similarity among **sets**

1. JACCARD INDEX
2. MAXIMUM LIKELIHOOD ESTIMATION
3. COSINE SIMILARITY BASED ON TF-IDF



Core / support domains

Websites (domain names) can be naturally divided in two types

Core domains
www.nytimes.com
www.repubblica.com
twitter.com
www.lastampa.it
www.youtube.com

Support domains
static01.nyt.com
abs.twimg.com
upload.wikimedia.org
cdns.gigya.com
gstatic.com

We use a simple tree-based model to automaticallty identify them

We create profiles **separately** for core and support domains

Goal: what works better for re-identification?

- What we access **intentionally**?
- The "**background noise**" generated by our devices?

Experimental setup

Dataset from a University campus
Users with fixed IP addresses -> we get a ground truth

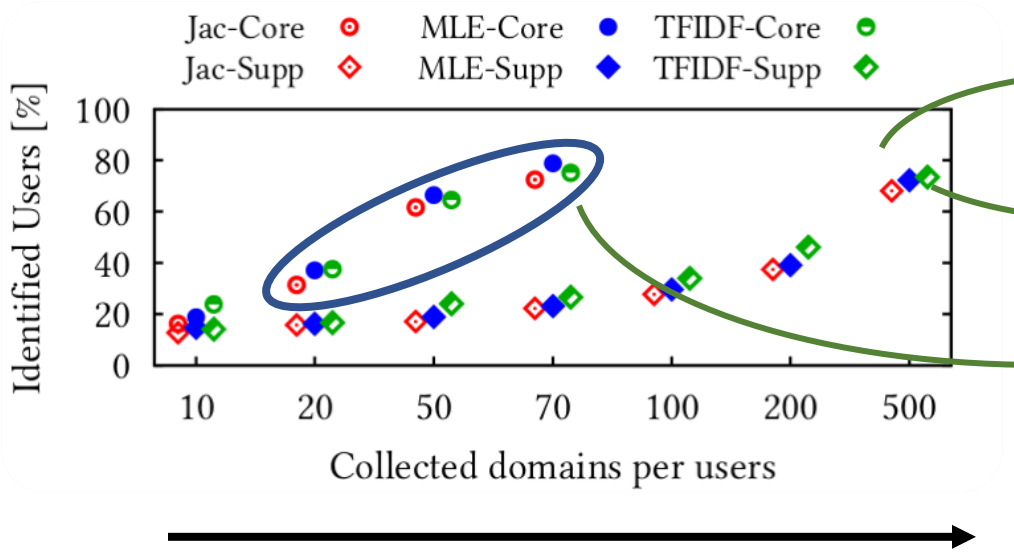
Log Size	Volume	Client IPs	Domains	2nd-lvl
229 GB	113 TB	≈2 500	404 k	136 k

- Load and process the logs using **Apache Spark** in a 20-machine Hadoop cluster
- Reading and processing the Campus dataset in about **20 minutes**.
- 1 hour for **classifying** 404 k domains as Core or Support domains.

Identification accuracy

Results separate for Core and Support domains

Core domains (websites) are more important than Support domains (CDN domains, background apps, etc.)



- ✓ Jaccard performs worst in all the cases
- ✓ TFIDF has the best results in most of the experiments, but
- ✓ MLE performs a slightly better with Core domains.

The larger is the data, the better is the identification Accuracy Up to **85% (on 2 k users)**

We are repetitive. An attacker with a big dataset can us this to re-identify us!

Can we hide the websites we visit?

Scenario

Question: can we use ML to understand the website of an encrypted connection

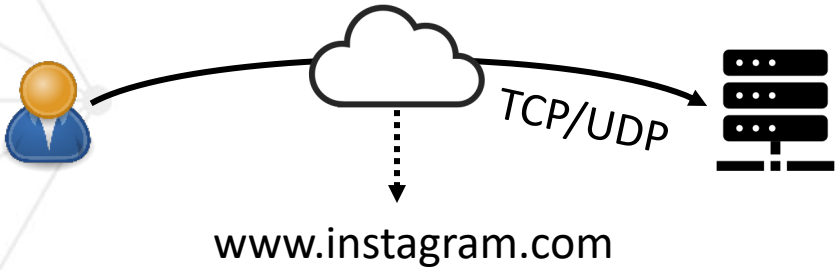
Scenario: Signalling protocols are encypted (third scenario)

- DNS is encrypted over HTTPS
- HTTPS uses the Encrypted server name indication (eSNI)

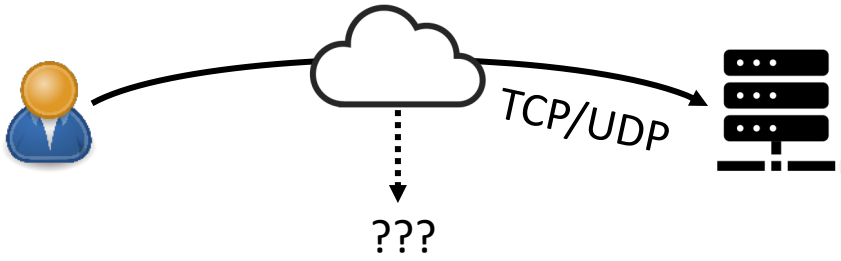
➢ The network cannot associate a website to a flow

Less than 2% of clients already updated

Before: Non-encrypted signaling



Now (close future): Encrypted signaling



Use ML -> www.instagram.com

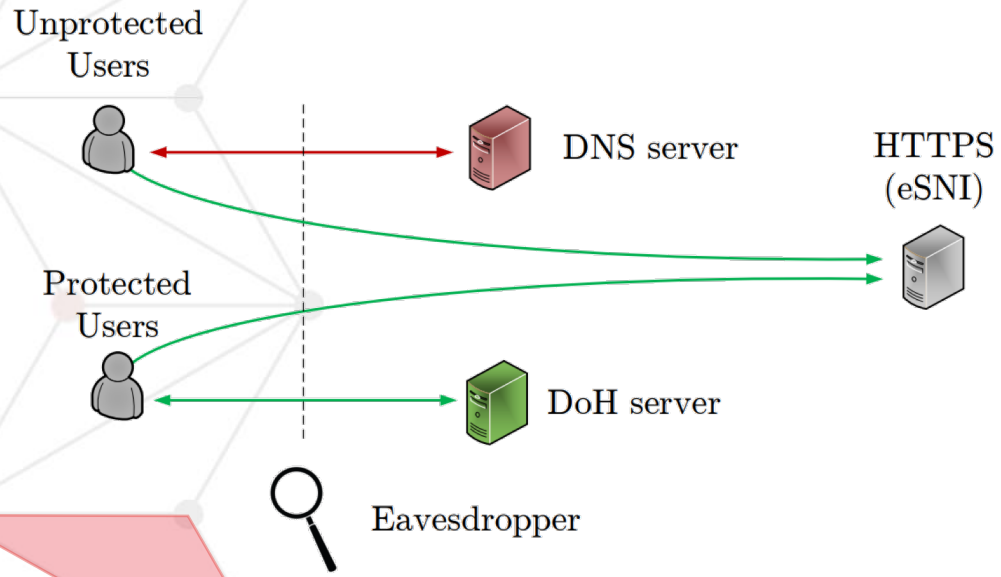
Experimental setup

Use a dataset from a University Campus

- Flow records for 1 month
- 3,900 users
- 900 M contacted websites

Encrypted signalling used by 2% of users

➢ We have the ground truth for all the dataset (= we have the website for each TCP/UDP flow)



We assume that the attacker:

Has the ground truth for 50% of clients → **Training**

- Because he controls a DNS server, or creates a testbed

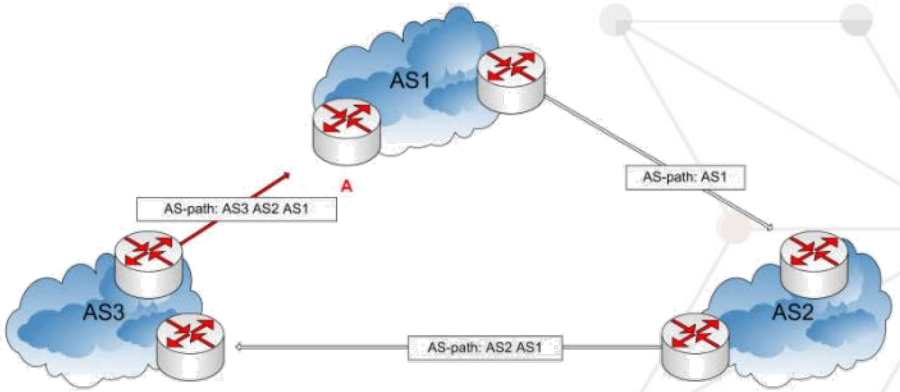
Wants to classify the remaining traffic → **Testing**

- Associate a TCP/UDP flow to the corresponding website

Machine Learning Methodology

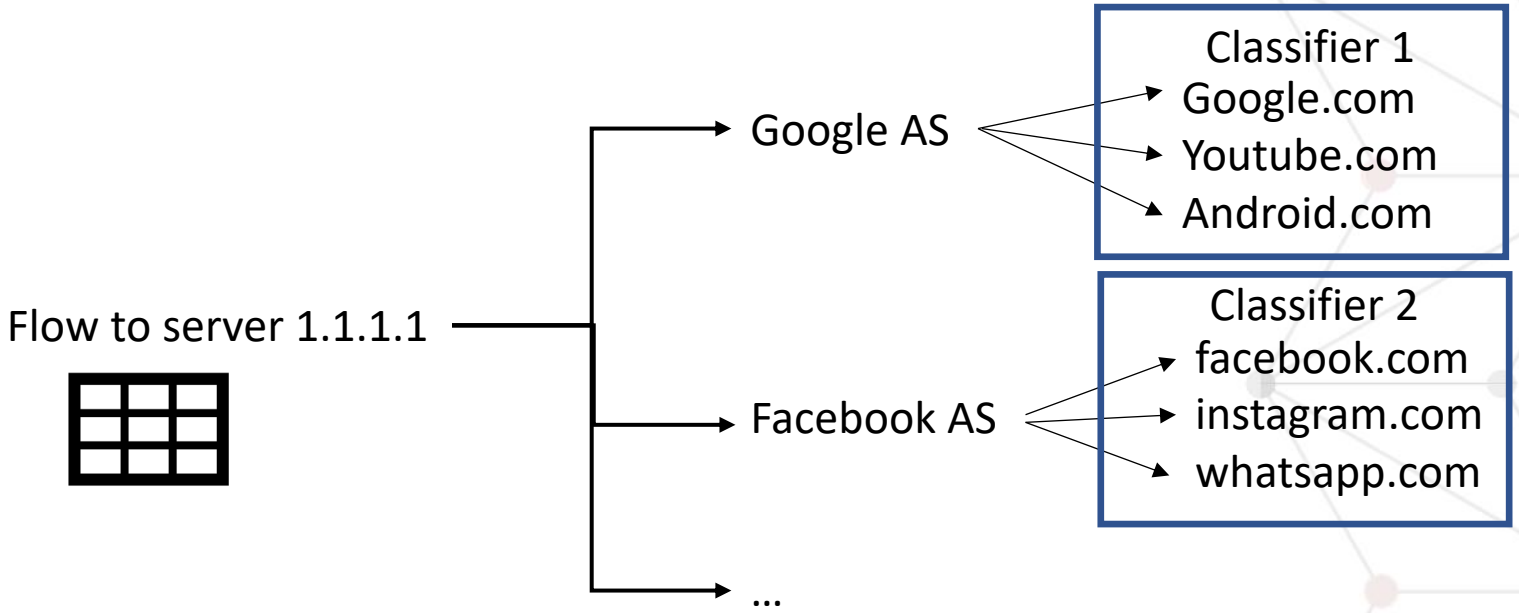
On the Internet, the set of networks owned by the same body are called **«Autonomous Systems»**
Google, Facebook, Microsoft, Amazon have their AS
The IP addresses associated to an AS are public

- ➢ We split our classification problem in many subproblems



Features extracted from flow characteristics

- Packet size
- Timing
- TCP level flags
- …. More than 100 ….

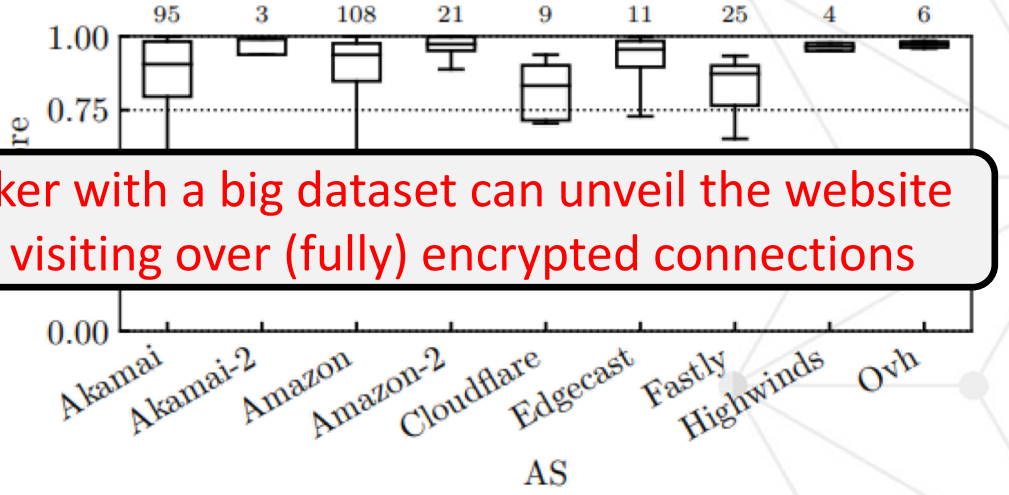


Does it work?

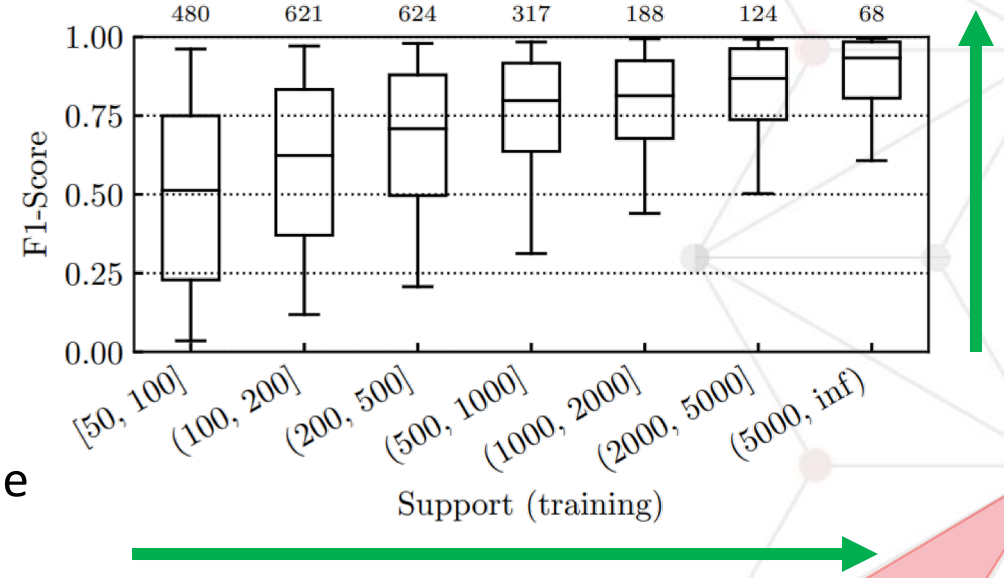
We consider 1 month of traffic
3900 users
- 50% training
- 50% testing

Try different off-the-shielf classification algorithms
Use Spark more most of processing
Focus on 9 ASes of top-Internet players
- Consider only cloud providers (e.g., Amazon)
- Google, Facebook would be too easy ☺

Goal:
associate the website to TCP/UDP flows



An attacker with a big dataset can unveil the website we are visiting over (fully) encrypted connections



Results:

80% of domains can be classified with F1-Score > 0.8
- **On 280 most popular websites**

Random Forest the best classification algo
Most impacting factor: **dataset size**
- The more you observe a website during training, the better you classify it at testing time

Conclusion

The Privacy trade-off

- Network monitoring is useful for cybersecurity, traffic engineering
- Users want privacy

Currently, users' privacy is triumphing – driven by content providers

➢ Everything is going encrypted

Encryption is not a miracle cure

- Also attackers can play with Big Data and ML
- Large datasets allow to:
  - Re-identify users based on their website visits
  - Identify websites behind encrypted connections


Perguntas
Fragen
Domande
Galdera
Otázky
Questions
Spørgsmål
Pertanyaan
kysymykset
Frågor
Spørsmål
Cwestiynau
вопросы
Preguntes
Sorular
Въпроси
Vragen
Pytania