POLITECNICO DI TORINO 1859 1906

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

USERS' FINGERPRINTING TECHNIQUES FROM TCP TRAFFIC

LUCA VASSIO

DANILO GIORDANO - MARTINO TREVISAN - MARCO MELLIA - ANA COUTO DA SILVA

Luca Vassio - About me

Italian **Mathematical engineer** with a multidisciplinary profile
PhD in **Telecommunication Engineering** – now postdoc

POLITECNICO DI TORINO
Massachusetts Institute of Technology
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
UNIVERSIDADE FEDERAL DE MINAS GERAIS

GE Aviation
energy way
creditsafe
NOKIA Bell Labs

Research interests

- Data science
  - Big data analytics
  - Data mining and machine learning
- Human behaviour analysis and modelling
  - Recommendation systems
  - Social networks behaviour
  - Web browsing habits
  - Smart cities design
- Optimization
  - Metaheuristics

Are our interests private?

theguardian

How NSA can see 'nearly everything you do online'

Exclusive Secret tool searches email, chat and social media use

THE DATA BROKERS: SELLING YOUR PERSONAL INFORMATION

...at collects,
...Americans with

Not at all!

BIG DATA: HOW TARGET KNOWS YOU ARE PREGNANT

Huge number of trackers that record user web activities with different techniques

Can we be anonymous?

From a **network point of view**

- Does our traffic characterize us?

Users can change

- Application
- Device
- Network

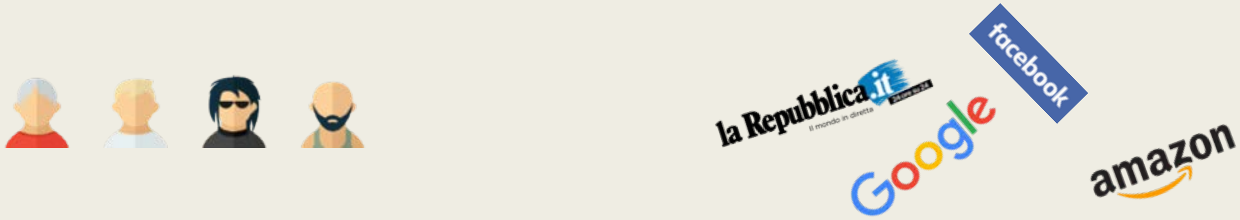
HTTPS limits access to third parties

Can we still use network visible information for user fingerprinting?

Fingerprinting: user or device?

What characterize the behaviour of the **user** when online

- The web-services the user access



What characterize the behaviour of the used **device**

- The web-services the user access
- Services that support these web-services (e.g., CDNs)
- The installed applications (e.g., software updates)

Contributions

Goal

Study capabilities of tracking using only visited domains (server FQDN)
Domains obtained from passive traces (TCP logs)

What is novel?

1. Only consider the name of the contacted server

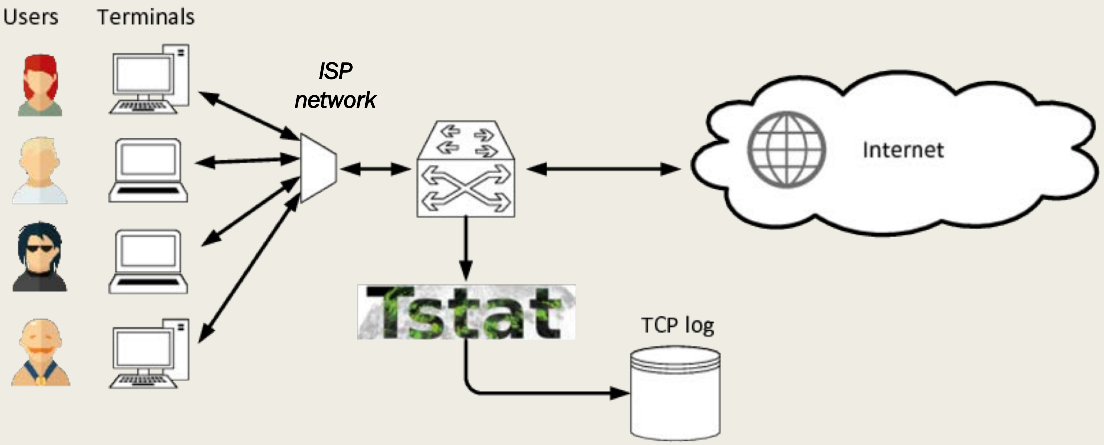
2. Compare different metrics for tracking

3. Propose a methodology to identify domains intentionally requested

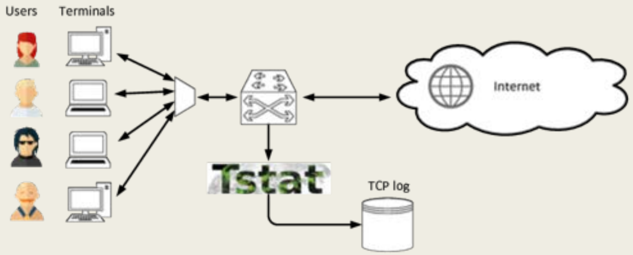
Passive network measurements

- Process of measuring the traffic exchanged between devices interconnected by a network
- Traffic generated by others and observed on client/server/network

Example: through Tstat



TSTAT: TCP Statistics and Analysis Tool



Tstat

- Captures traffic and processes it in real-time
- Logs more than 100 statistics per TCP flow
- Logs information from every HTTP request and response
- Original server domain name retrieved by HTTP/TLS deep packet inspection
- Tracks DNS conversations to retrieve original server domain name (DN-Hunter)

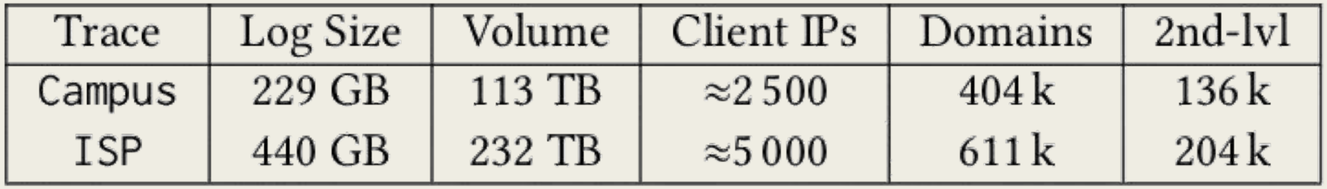
More information at tstat.polito.it

Client IP	Client port	Client bytes	Server IP	Server port	Server bytes	Domain	...
12.132.54.94	1197	18938	87.250.137.92	443	992221	Acme.com	...
12.132.54.94	3441	16541498	123.220.231.13	8080	78661	Example.com	...

Datasets

1. University campus
2. ISP

Users with fixed IP addresses - used as client ID
4 weeks in 2017



Trace	Log Size	Volume	Client IPs	Domains	2nd-lvl
Campus	229 GB	113 TB	≈2 500	404 k	136 k
ISP	440 GB	232 TB	≈5 000	611 k	204 k

Software: Apache Spark in a 20-machine Hadoop cluster

Computation: reading and processing Campus dataset in ~20 minutes

Ethics & Privacy

University ethical board: data collection reviewed and approved

Protect leakages of private information

1. Anonymize IP addresses - irreversible hash functions
2. Save only data strictly needed

   i) Anonymized IP addresses
   ii) Name of domain
   iii) Timestamp of the TCP connection

ISP dataset: reviewed and approved by ISP security board

i) No information about ISP customers
ii) Domain names never saved

Similarity computation

Hypothesis: users are repetitive over time!

Goal: identify user among profiles built in the past by checking visited domains

1. Profile the users creating fingerprints
2. Identify user in a later trace

Performance metric: percentage of users correctly identified

Select a fixed number of users in all tests → meaningful comparisons

Similarity computation

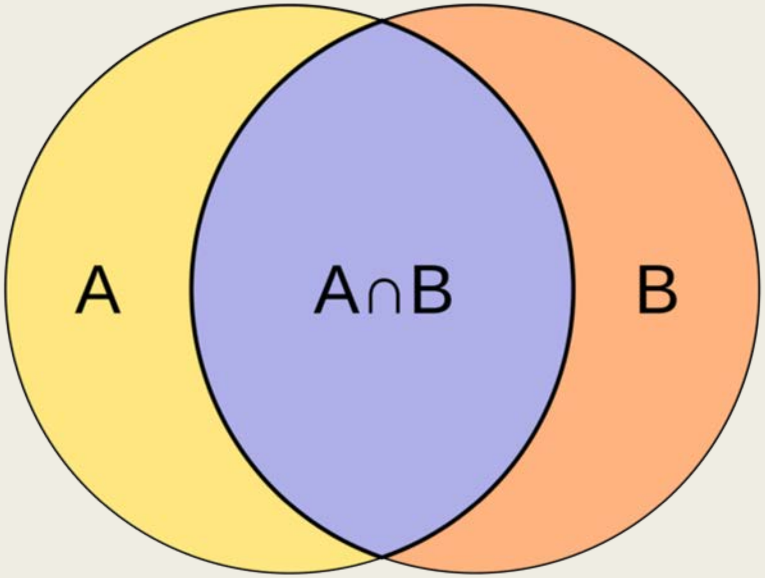
Three methodologies for similarity among fingerprinting sets

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

1. Jaccard index

Size of intersection divided by size of the union of two sample sets

- Just depends on two domains sets
- Fast computation time



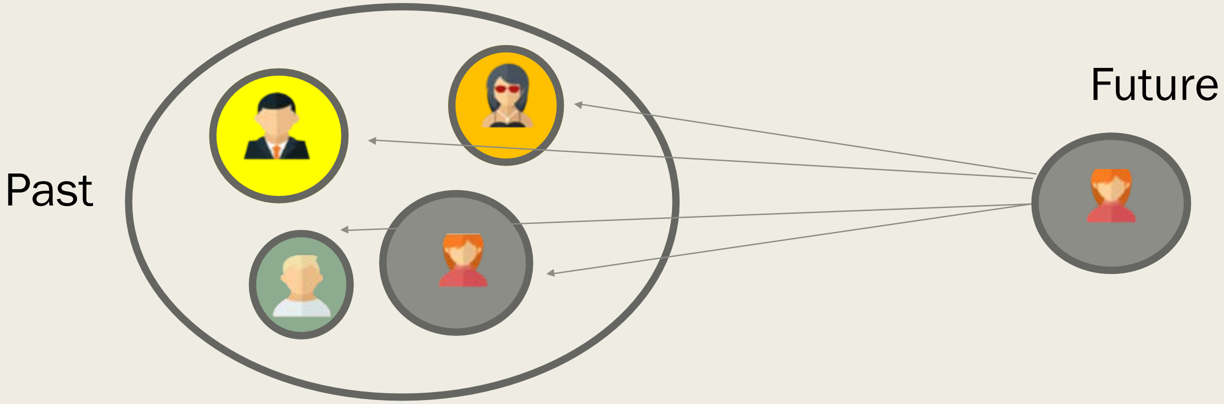
2. Maximum likelihood estimation

Behavioral model: user’s likelihood of visiting a domain governed by

(i) Domain overall popularity
(ii) Whether domain already appeared in her previous domains set

Each user has personalized factor of attraction towards past domains

Identification: for each user, likelihood of generating the future set with the model



Model proposed in:

J. Su, A. Shukla, S. Goel, and A. Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. Proocedings of the WWW 2017. 1261-1269.

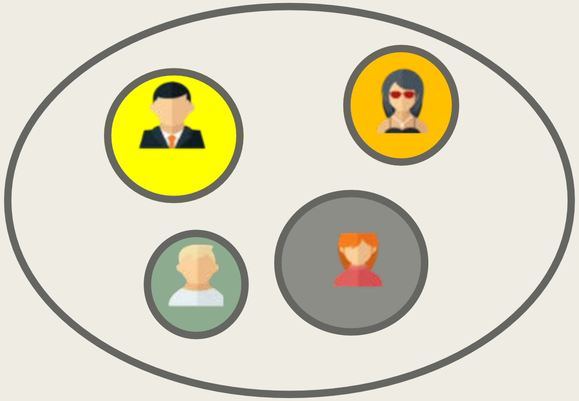
3. Cosine similarity with TF-IDF

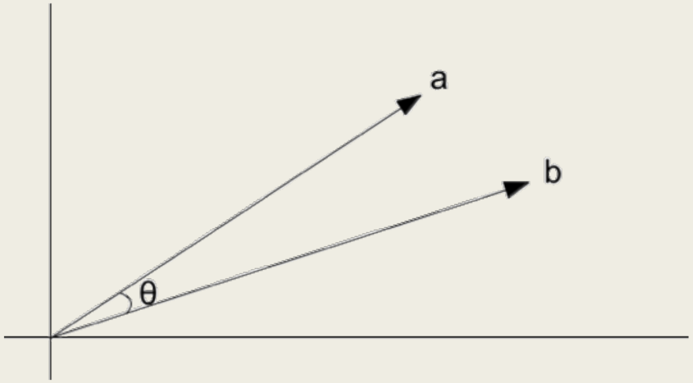
TF-IDF

- Statistic per domain and per user
- Reflects
  - How important a domain is for a user (TF – Term Frequency)
  - With respect to all users (IDF – Inverse Document Frequency)

Identification:

Cosine distance of vectors of TF-IDFs of domains





Computational complexity

M users each with ≈ N domains
P is the total number of domains seen by all the M users, with
$N \leq P \leq M \cdot N$

Jaccard index computation costs at most O (M · N^2)
TFIDF and MLE methodologies cost at most O (M · N · P), using larger set of all domains

Hashing methodology could be used on top of our computation to speed-up the process. Computation cost decrease to O (M · N) for Jaccard and O (M ·P) for MLE and TFIDF

In case of very large population, it is therefore much faster to use the simpler Jaccard similarity

Core and support domains

Are all domains equal?

Core domain: are intentionally requested by a user to download the page HTML document

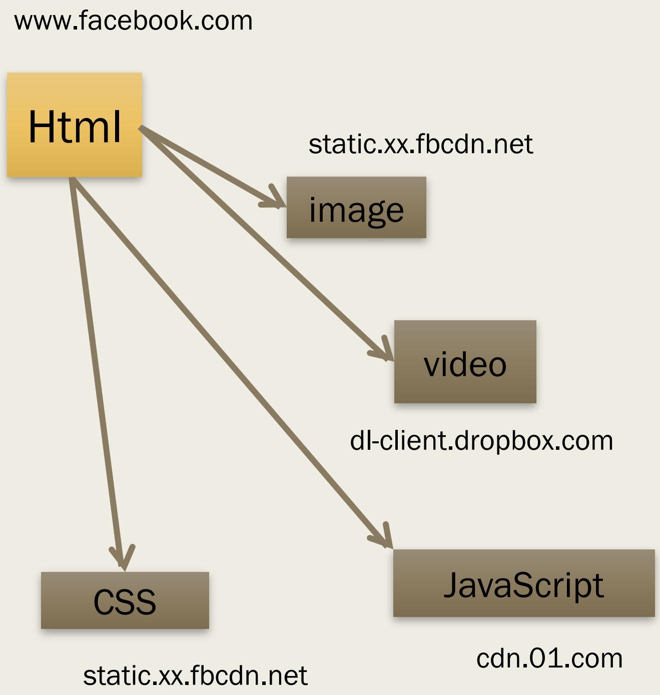
e.g.: www.facebook.com and en.wikipedia.org

Support domains: remaining ones triggered by website visits, or by background applications

e.g.: static.xx.fbcdn.net and client.dropbox.com

Core domains might be more important for fingerprinting

- Accuracy
- Interpretability
- Independence from device

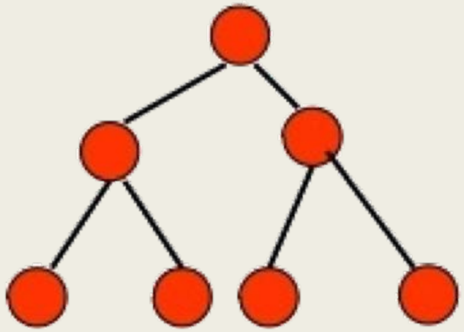


Identify core domains

Classification task

Machine learning methodology: decision tree classifier

1. Build a labeled dataset for training and testing
2. Get features: active crawling visiting home page of each domain (Selenium)
3. Features selection: HTML document size and redirection to external domain
4. Classifier training: C4.5



Results: accuracy 96%

Execution time: 1 hour for classifying 404 k domains

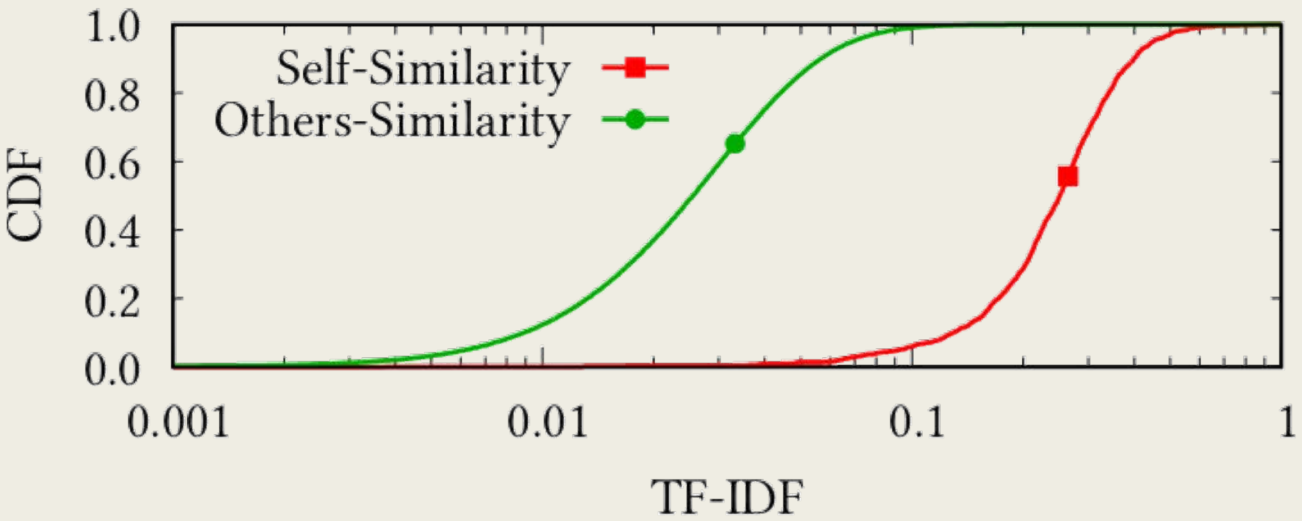
Bottleneck: Internet access speed (1Gbps)

RESULTS

Is user browsing repetitive?

Campus dataset

One week for profiling and identification

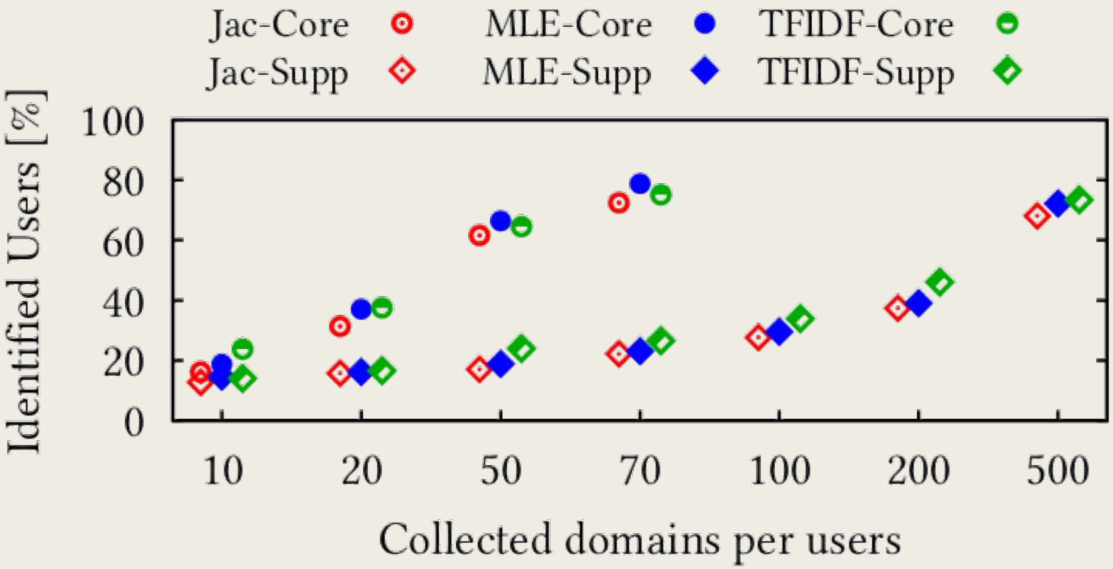


- Users behave similar over time
- Users are different among themselves
- Discriminative power of the profiles

Number of domains

Campus dataset

Core/support domains for profiling and identification tasks



- The bigger the set, the better the identification
- Core domains better characterizing - 500 support ~ 70 core

- Jaccard index worst method
- TFIDF best results
- MLE slightly better with core domains

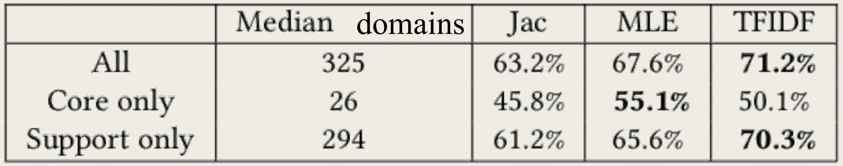
Observation time

Campus dataset

Users have different rates for discovering domains

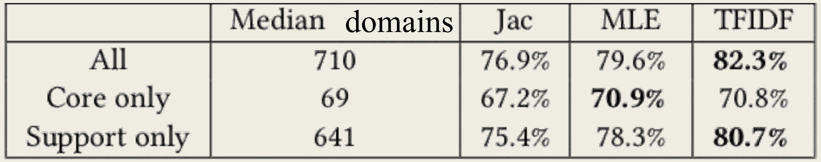
Keep constant observation time (profiling/identification)

Two consecutive days



	Median domains	Jac	MLE	TFIDF
All	325	63.2%	67.6%	**71.2%**
Core only	26	45.8%	**55.1%**	50.1%
Support only	294	61.2%	65.6%	**70.3%**

Two consecutive weeks



	Median domains	Jac	MLE	TFIDF
All	710	76.9%	79.6%	**82.3%**
Core only	69	67.2%	**70.9%**	70.8%
Support only	641	75.4%	78.3%	**80.7%**

- Much more support domains than core ones
- Quantity of support domains helps identification (personalized ads/tracking)

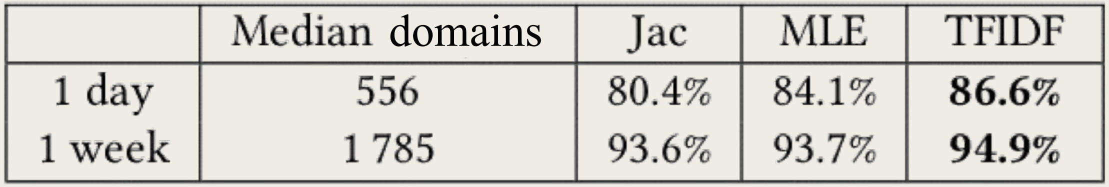
- Jaccard worst performance
- TFIDF overall best performance
- With few core domains MLE better

The ISP case

Let’s apply what we have learned to the ISP case

Residential access

- More heterogeneous users
- Devices multiplexed on same IP address of the access gateway
- All domains – no distinction between core/support



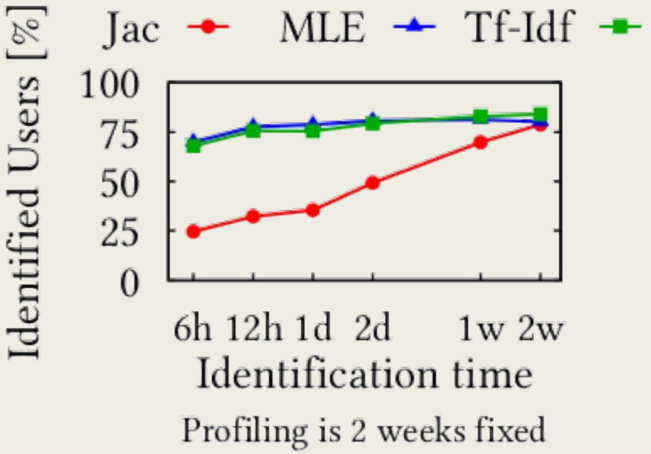
	Median domains	Jac	MLE	TFIDF
1 day	556	80.4%	84.1%	**86.6%**
1 week	1 785	93.6%	93.7%	**94.9%**

- Bigger amount of data
- More distinct and repetitive behavior: easier to fingerprint and identify

Longer profiling or identification?

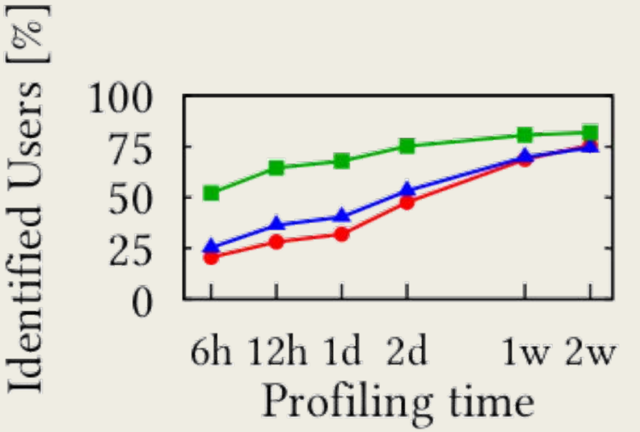
Campus dataset

1. Fixed 2 weeks profiling, variable identification time



Profiling is 2 weeks fixed

2. Fixed 2 weeks identification, variable profiling time



Identification is 2 weeks fixed

- Jaccard is symmetric
- TFIDF and MLE account for whole population when profiling
- TFIDF and MLE good even with few hours of identification
- Large profiling sets more important than identification

Conclusions

Still no privacy and anonymity online, even from the network point of view

Good similarity metrics and machine learning application can improve forensics applications:

- A simple TFIDF approach can solve identification problem
- Web-services intentionally requested (core) better characterize users
- But support domains are also important due to their quantity

Future work: analyze point of view of tracker applications

Foster new studies and permit results reproducibility!
Data and models available at bigdata.polito.it/content/domains-web-users

Luca Vassio, Danilo Giordano, Martino Trevisan, Marco Mellia, Ana Paula Couto da Silva, Users' Fingerprinting Techniques from TCP Traffic, ACM SIGCOMM Workshop on Big Data Analytics and Machine Learning for Data Communication Networks, 2017

Want to talk with me? Propose something interesting for my future?

Luca Vassio

luca.vassio@polito.it
lucavassio.wordpress.com