On the challenges of network traffic classification with NetFlow/IPFIX

Pere Barlet-Ros

Associate Professor at UPC BarcelonaTech

(pbarlet@ac.upc.edu)

Joint work with: Valentín Carela-Español, Tomasz Bujlow and Josep Solé-Pareta

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 726763.

Background

- What do we refer to as *traffic classification*?
  – Identifying the ***application*** that generated each ***flow***

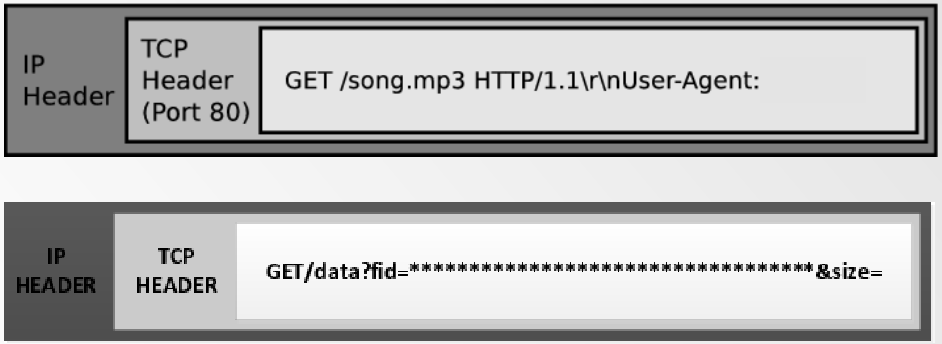
- What is traffic classification used for?
  – Network planning and dimensioning
  – Per-application performance evaluation
  – Traffic steering / QoS / SLA validation
  – Charging and billing

Background: *Ports*

- Port-based
  - Computationally lightweight
  - Payloads not needed
  - Easy to understand and program
  - Low accuracy / completeness (but most NetFlow products still use it!)

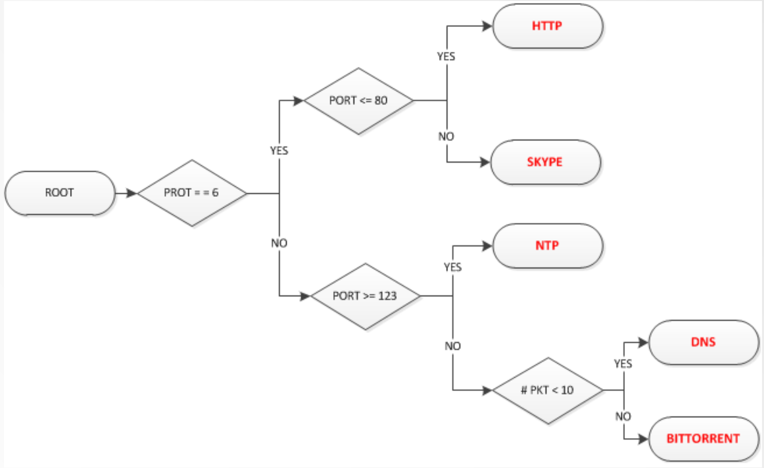
Background: *DPI*

- Deep packet inspection (DPI)
  - High accuracy and completeness
  - Computationally expensive
  - Needs payload access
  - Privacy concerns
  - Cannot work with encrypted traffic



Background: *ML*

- Machine Learning
  - High accuracy and completeness
  - Computationally viable
  - Payloads not needed
  - Can work with encrypted traffic
  - Needs frequent retraining



Main limitations of ML-TC

- Introduction in real products and operational environments is *limited* and *slow*
  - Current proposals suffer from practical problems
  - Actual products rely on simpler methods or DPI

- 3 main real-world challenges:
  1) The **deployment** problem
  2) The **maintenance** problem
  3) The **validation** problem

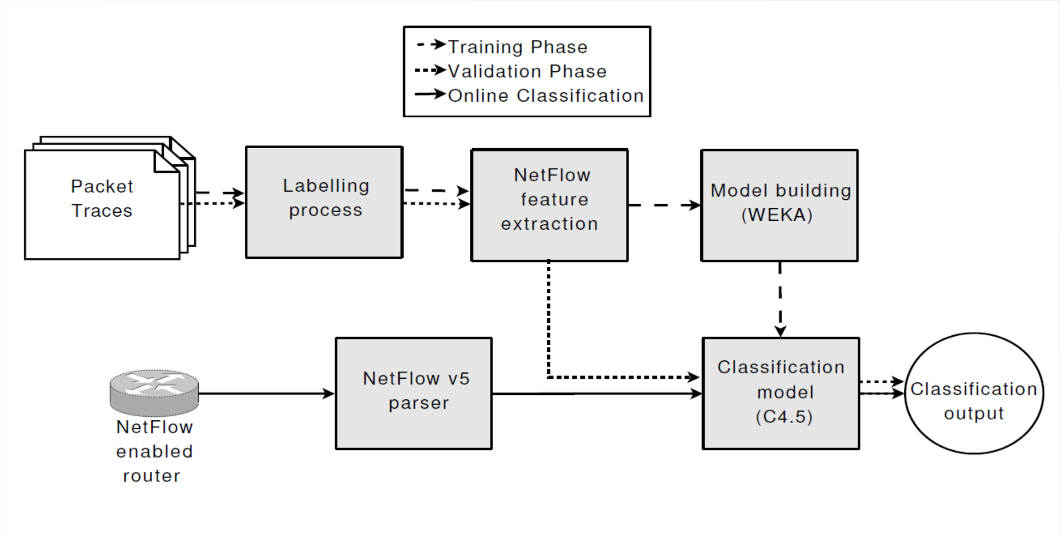
1) Deployment problem

- Current solutions are **difficult to deploy**
  - Need dedicated hardware appliances / probes
  - Need packet-level access (e.g. compute features, …)

- How to address this problem?
  - Work with flow level data (e.g. Netflow / IPFIX)
  - Support packet sampling (e.g. Sampled Netflow)

NetFlow w/o sampling

- Challenge: NetFlow v5 features are very limited
  - IPs, ports, protocol, TCP flags, duration, #pkts, …
- State-of-the-art ML technique: C4.5 decision tree



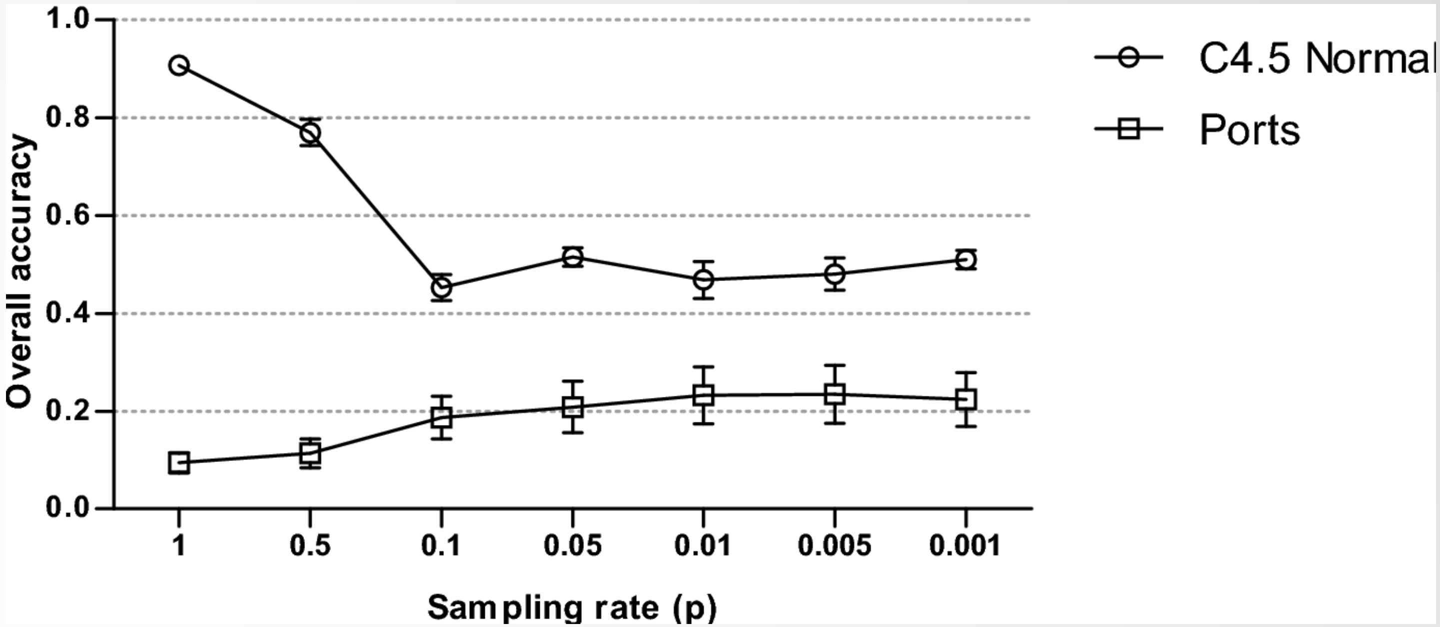
Results (NetFlow w/o sampling)

- UPC dataset: Real traffic from university access link
  - 7 x 15 min traces (collected at different days / hours)
  - Labelled with L7-filter (strict version with less FPR)
  - Public data set available at: https://cba.upc.edu/monitoring/traffic-classification

Name	Overall accuracy			
	C4.5			Port-based[8]
	Flows	Packets	Bytes	Flows
UPC-I	89.17%	66.37%	56.53%	11.05%
UPC-II	93.67%	82.04%	77.97%	11.68%
UPC-III	90.77%	67.78%	61.80%	9.18%
UPC-IV	91.12%	72.58%	63.69%	9.84%
UPC-V	89.72%	70.21%	61.21%	6.49%
UPC-VI	88.89%	68.48%	60.08%	16.98%
UPC-VII	90.75%	61.37%	40.93%	3.55%

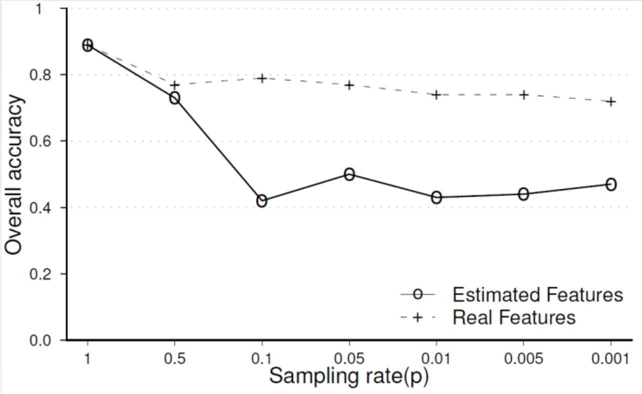
Results (Sampled NetFlow)

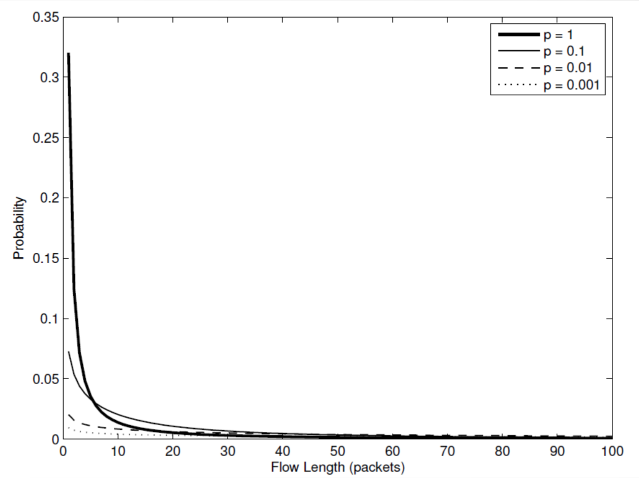
- Impact of packet sampling

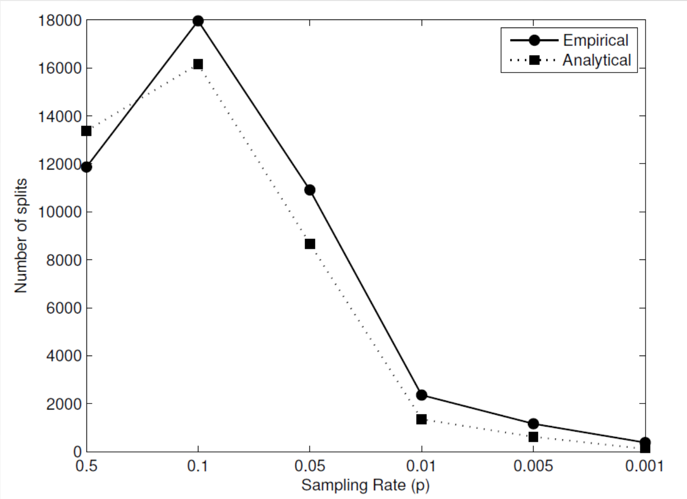


Sources of inaccuracy

1) Error in the estimation of the traffic features



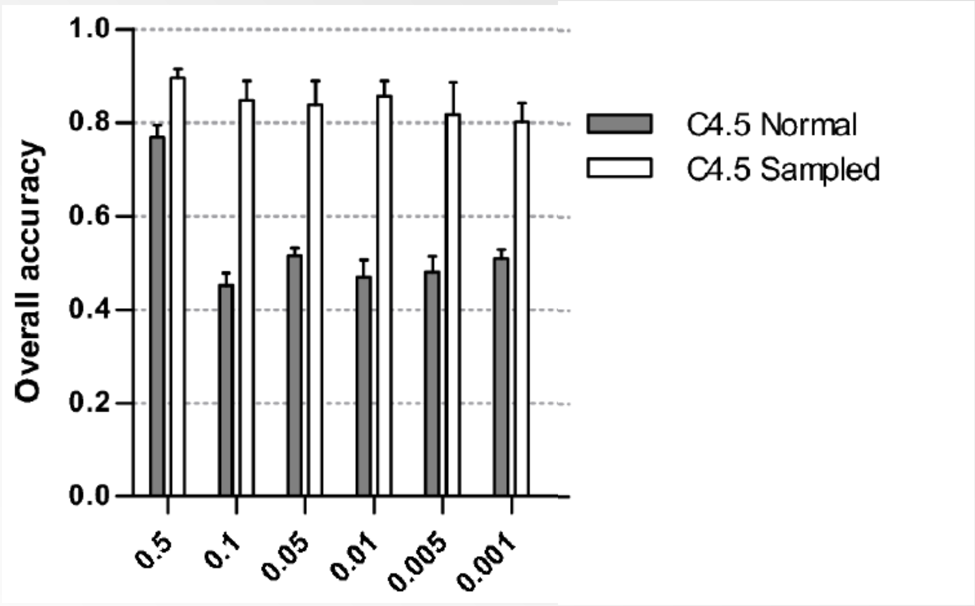




2) Changes in flow size distribution

3) Changes in flow splitting probability

Solution (Sampled NetFlow)



V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, J. Solé-Pareta. **Analysis of the impact of sampling on NetFlow traffic classification**. *Computer Networks*, 55(5), 2011.

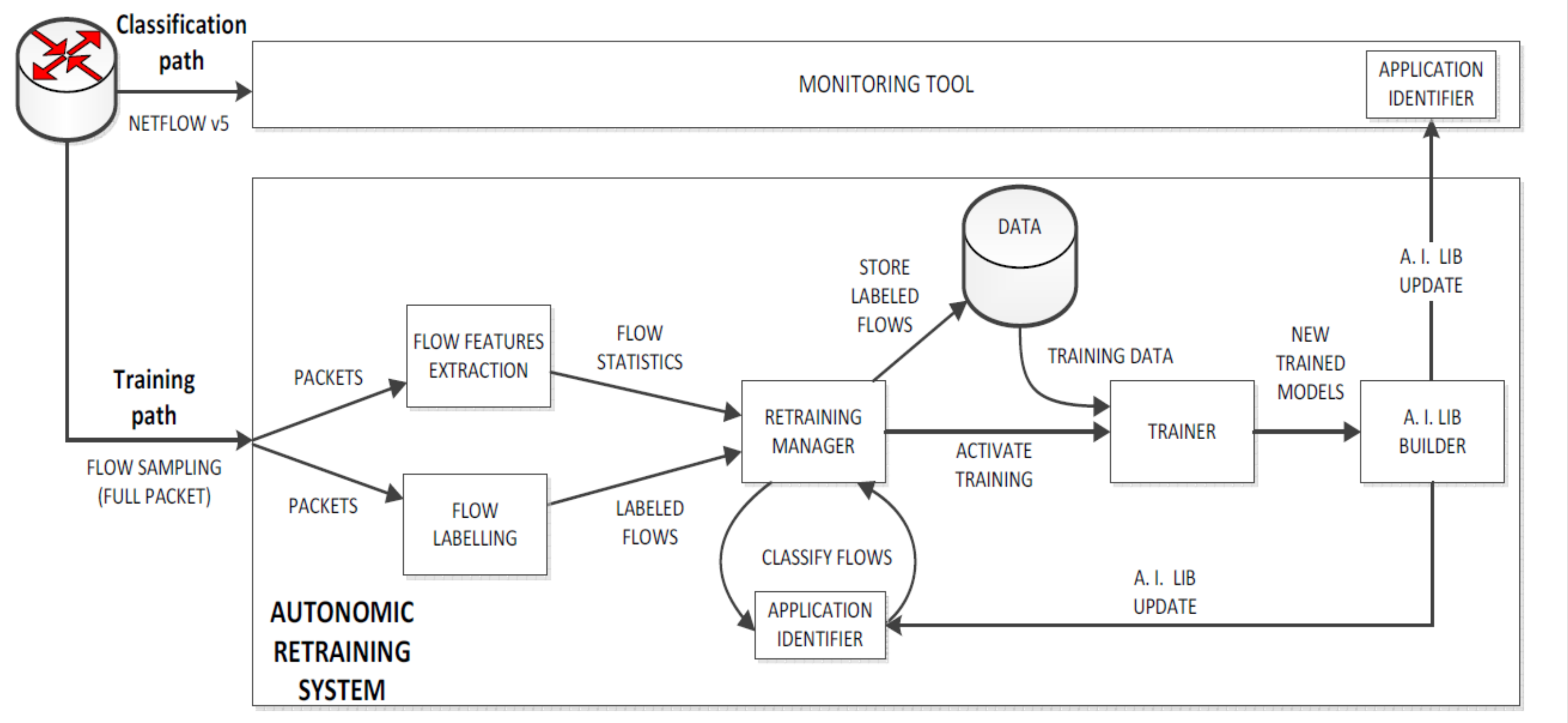
2) Maintenance problem

- Difficult to keep classification model updated
  - Traffic changes, application updates, new applications
  - Involve significant human intervention
  - ML models need to be frequently retrained

- Possible solution to the problem
  - Make retraining automatic
  - Computationally viable
  - Without human intervention

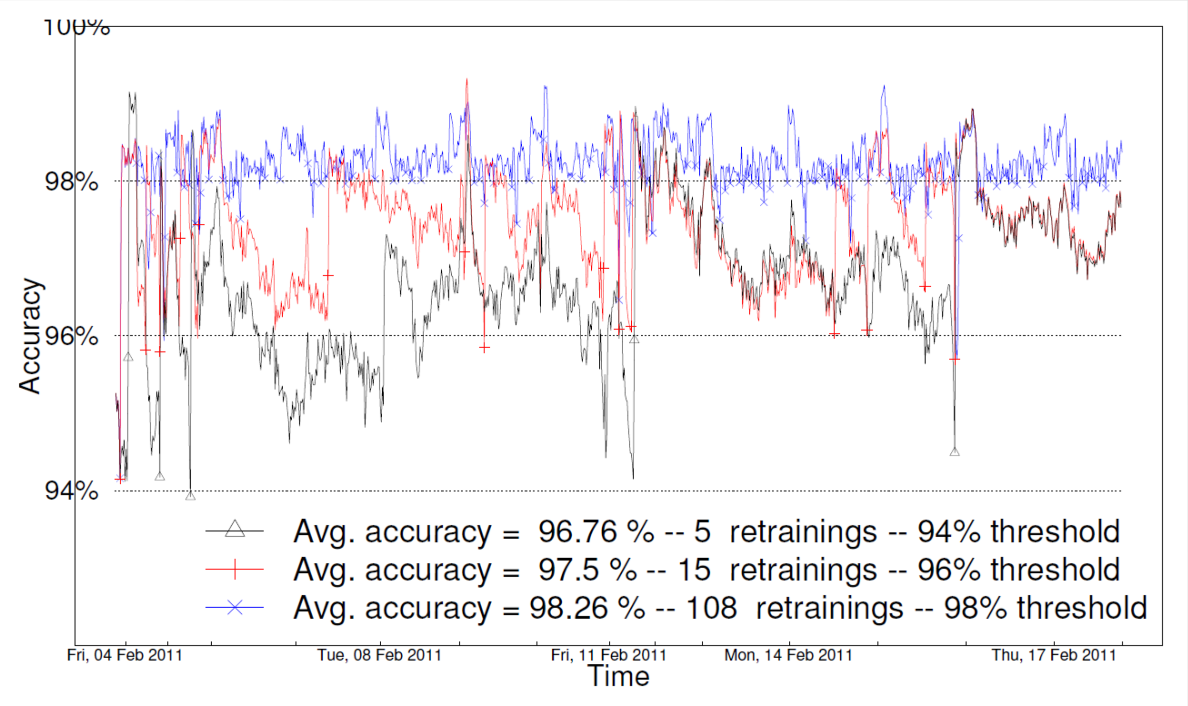
Autonomic Traffic Classification

- *Lightweight* DPI for retraining
  - Small traffic sample (e.g. 1/10000 flow sampling)



Results

- 14-days trace collected at the *Anella Científica* (Catalan RREN) managed by CSUC (www.csuc.cat)



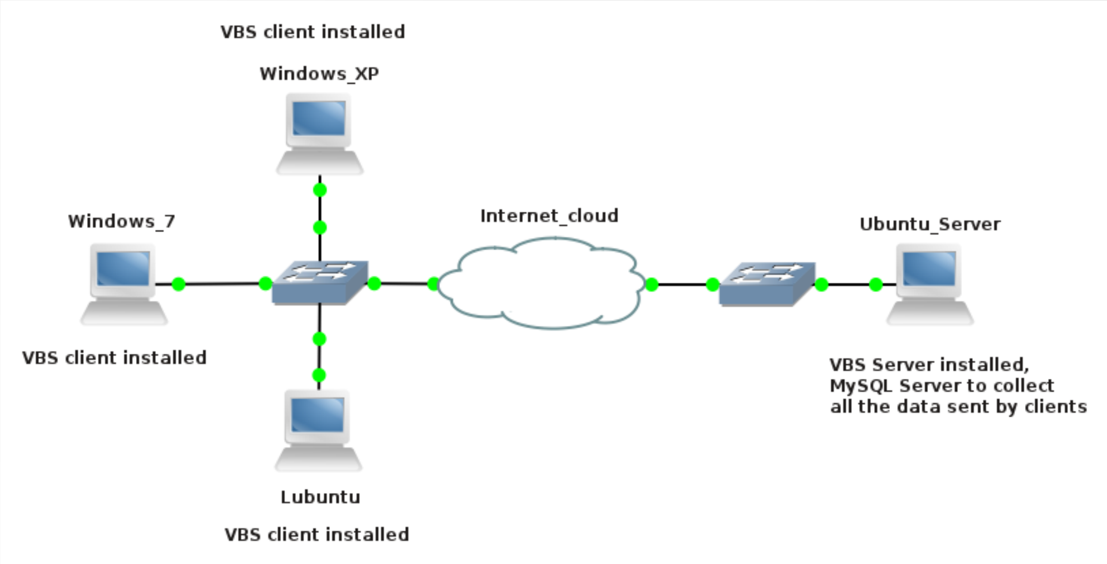
V. Carela-Español, P. Barlet-Ros, O. Mula-Valls, J. Solé-Pareta. **An autonomic traffic classification system for network operation and management**. *Journal of Network and Systems Management*, 23(3):401-419, 2015.

3) Validation problem

- Current proposals are difficult to **validate**, **compare** and **reproduce**
  – Private datasets
  – Different ground-truth generators

- Our contribution
  – Publication of labeled datasets (with payloads)
  – Common benchmark to validate/compare/reproduce
  – Validation of common ground-truth generators

Methodology



- Manually generate representative traffic
  - Create fake accounts (e.g. Gmail, Facebook, Twitter)
  - Interact with the service simulating human behavior (e.g. posting, gaming, watching videos, skype calls …)

Data set

- Public **labeled** data set with **full payloads**
  - Accurate: VBS (label from the application socket)
  - Avoids privacy issues: Realistic “artificial” traffic
  - Limitations: Traffic mix might not be representative

- Data set is publicly available at:
  - http://www.cba.upc.edu/monitoring/traffic-classification
  - Shared with 200+ researchers over the world
  - Cited in 100+ scientific articles (source: Google Scholar)

Data set

- > 750K flows, ~55 GB of data
- 17 application protocols
  – DNS, HTTP, SMTP, IMAP, POP3, SSH, NTP, RTMP, …
- 25 applications
  – Bittorrent, Dropbox, Skype, Spotify, WoW, …
- 34 web services
  – Youtube, Facebook, Twitter, LinkedIn, Ebay, …

T. Bujlow, V. Carela-Español, P. Barlet-Ros. **Independent comparison of popular DPI tools for traffic classification**. *Computer Networks*, 76:75-89, 2015.
V. Carela-Español, T. Bujlow, P. Barlet-Ros. **Is our ground-truth for traffic classification reliable?** In Proc. of *Passive and Active Measurement Conf.* (PAM), 2014.

DPI tools compared

Table 1: DPI tools included in our comparison

Name	Version	Released	Apps. identified
PACE	1.47.2	November 2013	1000
OpenDPI	1.3.0	June 2011	100
nDPI	rev. 7543	April 2014	170
L7-filter	2009.05.28	May 2009	110
Libprotoident	2.0.7	November 2013	250
NBAR	15.2(4)M2	November 2012	85

Results: Application protocols

- Most tools achieve 70%-100% accuracy
- nDPI and Libprotoident showed highest completeness (15/17)
- Only Libprotoident identified encrypted protocols (e.g., IMAP TLS, POP TLS, SMTP TLS)
- L7-filter suffered from false positives (9/17)

Results: Applications

- 20-30% less accuracy compared to protocols
- PACE (20/22) and nDPI (17/22) obtained highest completeness
- Libprotoident showed reasonable acc. (14/22)
  - Note it only uses 4 bytes of the payload
- NBAR showed very low performance (4/22)
  - Unable to classify most applications

Results: Web services

- PACE: 16/34 (6 over 80%)
- nDPI: 10/34 (6 over 80%)
- OpenDPI: 2/34
- Libprotoident: 0/34
- L7-filter: 0/44 (high FPR)
- NBAR: 0/34

Implications for operators

- Current DPI products are **expensive** and **difficult to deploy**

- Accurate traffic classification with *Sampled NetFlow* is **possible** and **easy to deploy**

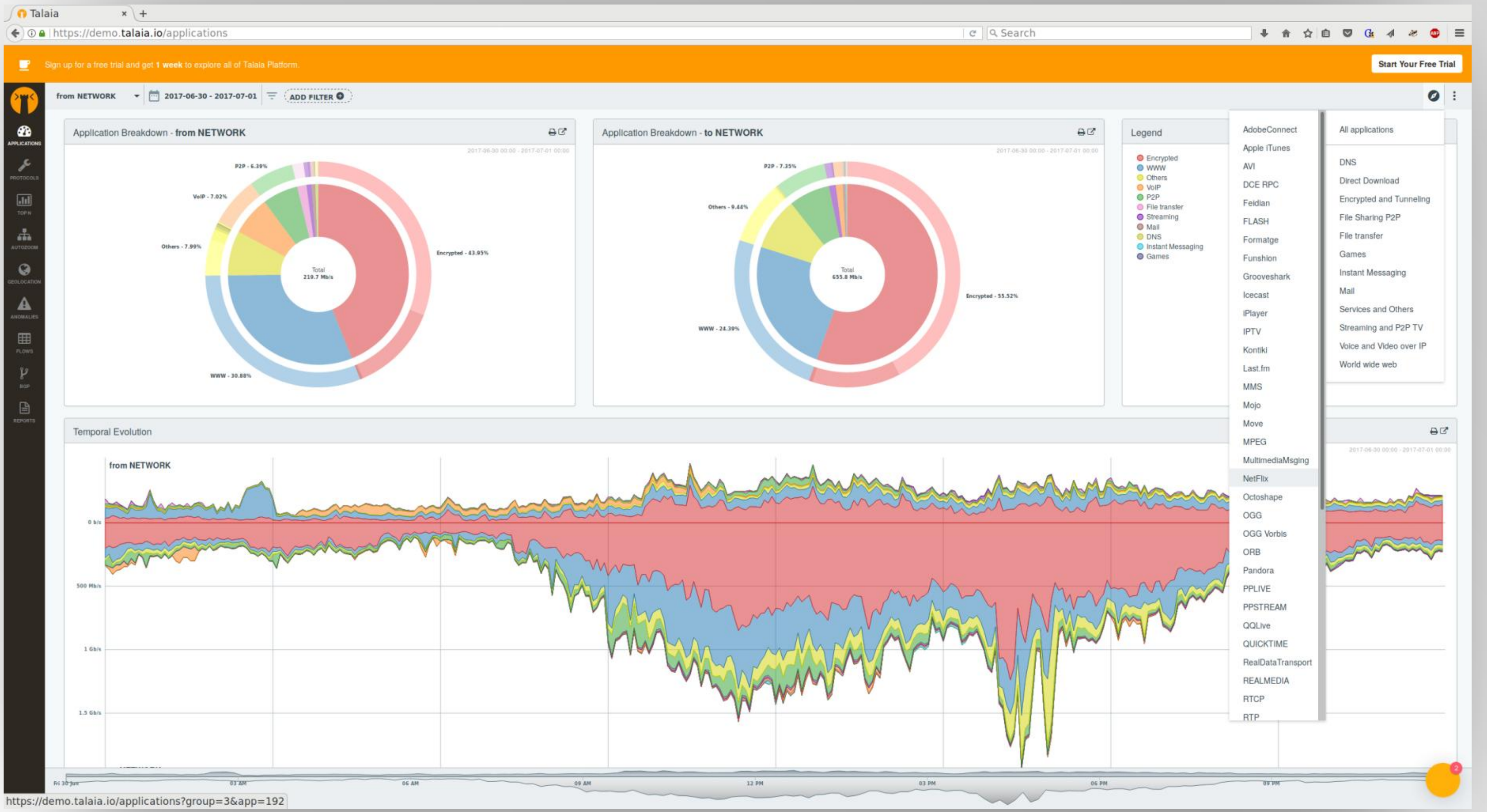
- *Sampled NetFlow* traffic volumes are low
  - Flows can be easily sent (encrypted) to the cloud
  - Monitoring can be offered **as a service** (SaaS)

Real implementation

- Received funding from EU H2020 to convert technology into a commercial product
  - SME Instrument Phase 2 project
  - Grant agreement No. 726763

- Talaia Networks, S.L. (www.talaia.io)
  - Spin-off of UPC Barcelona-Tech
  - Monitoring and security service (SaaS and on-prem)
  - Customers worldwide (operators, ISPs, cloud prov., …)

On-Line Demo



https://www.talaia.io