Towards Web Services Classification

**Problem**

- Traffic classification difficult in modern web, traffic more and more encrypted, CDNs and Cloud Providers complicate the scenario
- A classifier can still rely on flow level metrics (flow records)
- Server IP address doesn’t give much information

**Goal**

- Account traffic to web service generating it
- Evaluate which features of traffic are useful for classification
- Leverage machine learning techniques
- Self-learning approach when possible

**How many names have IP addresses?**

- Hostnames per IP address
- Traffic per IP address class

The majority of addresses associated to 1 hostname.
Those addresses carry a little share of the traffic (20%).

**Bags of IP addresses**

- Enumerate all the IP addresses of some popular services.
- Consider all traffic going to those addresses as belonging to the respective service.

**How stable are IP addresses in the time?**

- Facebook addresses over 2 weeks
- Persistence of addresses over 2015

A classifier relying only on server IP address doesn’t achieve high performance (about 20-30% of coverage).

**Temporal correlation**

Characterize interesting domains neighborhood

**Future work**

- A classifier relying only on server IP address doesn’t achieve high performance (about 20-30% of coverage)
- Other traffic characteristic can be exploited for classification:
  - DNS requests and replies
  - Packet level features (e.g., packets size, arrival time...)
  - Temporal and spatial correlation among flows
- Combine such approaches in a unique system

---

Martino Trevisan  
Politecnico di Torino  
martino.trevisan@polito.it  
Supported by Cisco