Network traffic analysis (for encrypted traffic and security monitoring)

Jérôme François
Inria Nancy Grand Est, France
jerome.francois@inria.fr

NMRG - IETF100
November 14th, 2017

Outline

Why traffic analysis?

Supporting network management operations

- detect / prohibits illegitimate behaviors
- enforce access control
- resource provisioning
- QoS...

Traffic classification

- know your traffic profiles $\rightarrow$ monitor, predict evolution
- no single definition of a profile (level of classification): protocol, application type, user, service use type, service providers, user location...

Challenges

Legacy techniques

Level	Discriminative feature(s)
Protocol / application	ports
User	IP address, hostname
Service provider	IP address, domain name

- + content for all (signatures)

Changes over the last years

- Applications relies on same framework and protocols to ease integration in multiple OS and devices → predominance of web-based applications
- Outsource servers and processes (clouds, CDN...)
- Privacy concerns raise: encryption (HTTPS generalization), VPNs, ToR...

The encryption Dilemma

Security vs. Privacy

- Secure protocols are now widely used (many relies on TLS)
- Despite SSL/TLS good intentions, it may be used for illegitimate purposes.
- By default solution: enforce use of proxies to decrypt communications

The main research question

Can we rely on the monitoring techniques that do not decrypt encrypted traffic (e.g. HTTPS)?

Outline

A Multi-Level Framework to Identify HTTPS Services

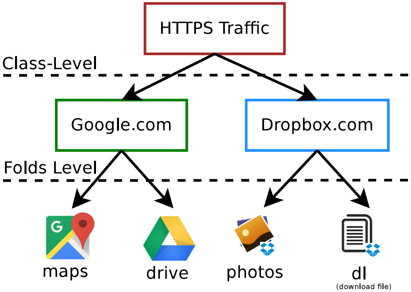


Figure : Multi-level presentation

Multi-level method

- Reform the training dataset into a tree-like fashion.
- The top level is refereed as Class-level (Root domain)
- The lower Level contains individual Folds-level (Sub-domain)

The 18 selected features

Client $\leftrightarrow$ Server
Inter Arrival Time (75th percentile)
Client $\rightarrow$ Server
Packet size (75th percentile, Maximum), Inter Arrival Time (75th percentile), **Encrypted Payload(Mean, 25th, 50th percentile, Variance, maximum)**
Server $\rightarrow$ Client
Packet size (50th percentile, Maximum), Inter Arrival Time (25th, 75th percentile), **Encrypted payload(25th, 50th, 75th percentile, variance, maximum)**

Evaluation Results

Second Level Evaluation

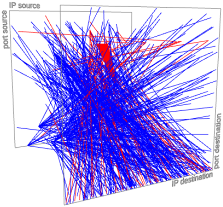
- We can identify the service provider of HTTPS traffic with 93.6% overall accuracy.
- From 68 distinct service providers, 51 service providers have more than 95% of good classification
- For example, we can differentiate between 19 services run under Google.com, with 93% of Perfect identification.

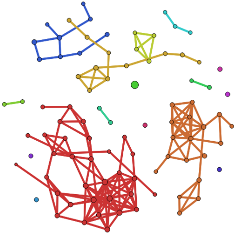
Accuracy Range	Nb of service providers		
-	Classical Features	Full Features	Selected Features
100-95%	50	51	51
95-90%	5	5	5
90-80%	6	6	6
Less than 80%	7	6	6

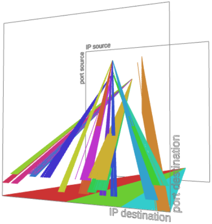
Without encryption, no challenge?

/20 darknet monitoring

- 1 month = 3 millions packets per day
- apply Topological Data Analysis to extract attack patterns (scanning, DDoS) → correlation
  - Analysis of high dimensional and complex data by extracting invariant geometrics features → discover relationships and patterns in data
  - Mapper algorithm: partition-based clustering







Outline

1 Challenges

2 Few examples

3 Going further

Target the right problem...

What is the goal of the traffic monitoring / classification ?

- characterize all the traffic: numerous signatures, many privacy concerns (e.g. identify all users)
- detect some particular traffic (for whitelisting / blacklisting purposes): simple models/signature to maintain, compliant with non massive surveillance
- identify individual patterns vs joint patterns

What is the final use?

- real-time, near real-time, batch (forensics)
- soft vs. strong impact: alerts vs. access control

… to define the proper methodology

General methodology (to be refined)

1. collect packet information
2. flow reassembly (e.g. extracting the TLS application data is useful for encrypted traffic)
3. Collect (application) specific information (= out of network information)
4. (train) and test the model

Feature engineering in the core of the process

1. limited set of features, widely used in literature (no real consensus)
2. packet-, flow-, application data-level statistics, end-points (number and variety), timing information

Need for network-specific ML

Commons errors

- suppose that there is no necessity to customize the model with context-specific information (*e.g.* the structure and semantics of data)
- use blackbox approaches (It is actually very hard to benchmark the best algorithms to use)

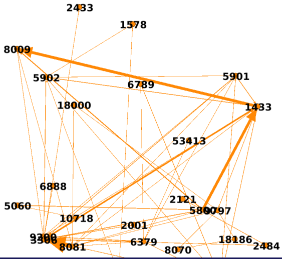
Distances between network flows (Euclidian distance?)

- Not all features are numeric
- Numeric features are not in the same space
- Usual distance may not catch the real semantic (e.g. port numbers)

TCP/UDP Port similarities

Towards a distance/similarity metrics between port numbers

- security → leverage attacker semantics from darknet monitoring
- graph mining (community detection) over scans
  - Database service ports: **mysql**: 3306, **redis**: 6379, **ms-sql-s**: 1443 (Microsoft-SQL-Server), **radg**: 6789 (GSS-API for the Oracle), **ttc-ssl**: 2484 (Oracle TTC SSL)
  - Medical service ports: **ohsc**: 18186 (Occupational Health SC), and **biimenu**: 18000 (Beckman Instruments, Inc)



Conclusion

Encryption can be overcame

- well-defined use case / target
- need to maintain signature databases

Remaining issues

- adversarial behaviors
- encrypted and *optimized* protocols (e.g. multiplexing)