



HAL
open science

A Multi-Level Framework to Identify HTTPS Services

Wazen M. Shbair, Thibault Cholez, Jérôme François, Isabelle Chrisment

► To cite this version:

Wazen M. Shbair, Thibault Cholez, Jérôme François, Isabelle Chrisment. A Multi-Level Framework to Identify HTTPS Services. IEEE/IFIP Network Operations and Management Symposium (NOMS 2016), IEEE/IFIP, Apr 2016, Istanbul, Turkey. p240-248, 10.1109/NOMS.2016.7502818 . hal-01273160

HAL Id: hal-01273160

<https://inria.hal.science/hal-01273160v1>

Submitted on 30 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Multi-Level Framework to Identify HTTPS Services

Wazen M. Shbair, Thibault Cholez, Jerome Francois, Isabelle Chrisment

Wazen M. Shbair

University of Lorraine, LORIA, France
shbair.wazen@loria.fr



NOMS2016 - Istanbul/Turkey

27 April, 2016

Outline

- 1 Introduction
- 2 Related work
- 3 The Multi-Level Framework
- 4 Evaluation Results
- 5 Conclusion & Future work

Security vs. Privacy

- HTTPS (HTTP-over-TLS) is a protocol for secure communication over internet.
- Content providers (Google, Facebook, ...) need securing contents over the web by moving to HTTPS.
- Based on French ISP, the amount of encrypted traffic represent almost 50% in 2015, compared with 5% in 2012.
- Despite SSL/TLS good intentions, it may be used for illegitimate purposes.

The Issue

An identification of HTTPS traffic without relying on decryption.

Practical solutions

- Legacy solutions: Port Based, DNS, IP, DPI → (Don't work).
- Decryption methods: HTTPS proxy ¹, Crack encryption algorithm → (**Privacy issues** & Computation complexity)
- Recent solutions: SSL certificate, SNI [1]→(Reliability issues).

Research work: flow-based statistical method

- + Applicable to encrypted traffic.
- - Low accuracy and computation overhead issues.
- - Hard to get precise information from general statistics.

¹Used by commercial solution like FireEye & Forefront



Flow-Based Statistical improvements

- One way is to combined it with algorithms from different fields like Machine Learning (ML) [2].
- Used to identifying the Type of Applications [3]
 - such as (HTTPS, SSH, P2P, Skype, VOIP, etc.)
- Used by Website Fingerprinting technique:
 - Identify accessed HTTPS web pages base on static object size parsed from unencrypted pages [4].



Flow-Based Statistical improvements

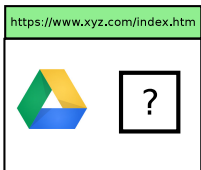
- One way is to combined it with algorithms from different fields like Machine Learning (ML) [2].
- Used to identifying the Type of Applications [3]
 - such as (HTTPS, SSH, P2P, Skype, VOIP, etc.)
- Used by Website Fingerprinting technique:
 - Identify accessed HTTPS web pages base on static object size parsed from unencrypted pages [4].

What is the proper level of identification?

Application type level OR Web pages Level



Flow-Based Statistical Method

Applications type
Inspector

HTTPS

encrypted traffic

Internet

```

5c c5 e0 5b 1e ...Ar].v ...X...[
00 00 20 c0 2b E...m.R...c...+
07 c0 11 00 33 /.../.../.../.../
05 00 04 01 00 .2.9././5
64 72 69 76 65 .W...drive
ff 01 00 01 00 .google. com...
19 00 0b 00 02 .....#...3t
05 00 05 01 00 .....#...3t
05 01 02 01 04 .....#...3t

3 05 7a 47 07 c0 .....U...
5b 74 97 e0 64 -.h./...pl[t..d
7 00 00 16 c0 .....h...h...
3 33 00 30 00 f./...3.9./
3 12 00 10 00 0.5...K...
5 2e 67 6e ff 1..lk7in7g 7bf.gn...
7 00 18 00 19 0 .....#...
1 00 16 00 14 04 .....#...
3 06 03 02 03 04 .....#...

i 03 56 5d ff c2 20 ...X... T..V]...
i 60 ea ab d8 05 01 E.B...
i b0 00 c0 2b 00 00 .....S...
i 00 23 00 00 33 74 .....#...3t
i 2f 33 2e 31 08 68 ...h2.sp dy/3.1.h
i 02 01 00 16 03 03 ttp/1.1.

33 03 56 5f 27 73 8f ...?... .V'*.s.
3b 31 03 b7 81 4b ca K.f...;1...K.
37 ab 00 c0 2f 00 00 .Tf...n] l7.../.
30 08 23 00 00 00 0b .....#...
3b 00 27 7d 00 27 7a .....#...}.'z
1e e0 a0 03 02 01 02 .....#...0

```

???

xyz
index pageWebsite
Fingerprinting
Inspector

Figure : An example of suspicious HTTPS traffic

- Application Type Level (Too generic)
- Website Fingerprinting Level (Too fine-grained)



A Multi-Level Framework to Identify HTTPS Services

The motivation

- An intermediate identification level **Service-Level**.
- Identify the HTTPS services without relying on header fields.
- Do not decrypt the HTTPS traffic.

The core techniques

- 1 Machine Learning techniques.
- 2 Novel multi-level classification approach.
- 3 Well tuned set of features.



Machine Learning Techniques

Overview

- Machine learning (ML) is a type of artificial intelligence (AI).
- The basic requirements:
 - 1 Training dataset and Labelling
 - 2 Statistical Features and ML algorithms.
 - 3 Evaluation techniques.
- ML phases: Training → Classification → Validation

Dataset Collection

- We build our own dataset in a well controlled environment with volunteer users of our lab.
- We use the SNI for HTTPS dataset Labelling.

What is SNI ?

SNI indicates the actual destination hostname a client is attempting to access over HTTPS.

The Ground-Truth

Since no SNI filtering is applied in our lab, so we utilized it as Ground Truth.

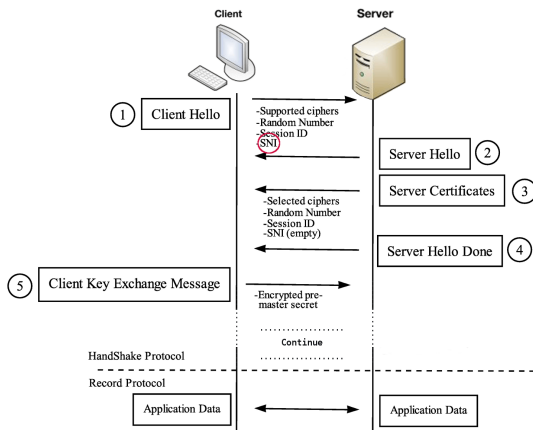


Figure : TLS handshake



Statistical features and ML algorithms

Statistical features

- A set of **42 features** over the TLS connections is used.
- Classical **30 features** from previous work [5, 6].
- New **12 features** are proposed over the encrypted payload.

ML algorithms

- The ML algorithms use them to build the classification model.
- Based on a preliminary experiments **C4.5** and **RandomForest** algorithms are selected.

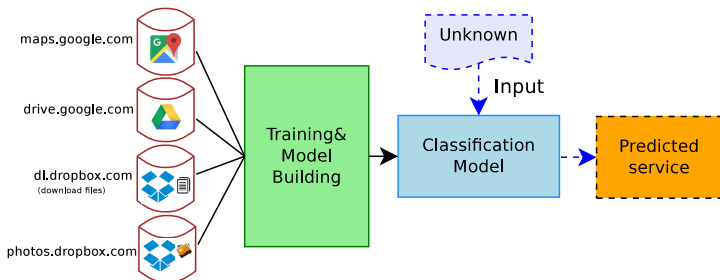


Figure : Flat classification view

Legacy machine learning flat classification

- Identifying the websites and applications directly.
- Drawbacks: low scalability, low accuracy and high error rate.

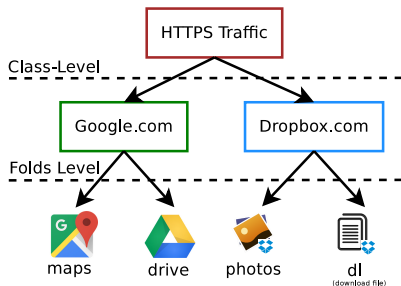


Figure : Multi-level presentation (inspired from Biology field)

A Novel Multi-Level Classification Approach

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

Common evaluation techniques

- A K-fold cross-validation, Precision, Recall, F-Measure.
- Receiver Operating Characteristics (ROC) analysis.
- The classification errors over time and the Confidence-Score.

A novel method more suitable for multi-level approach

- If service provider and the service name are predicted correctly
→ **Perfect identification.**
- If service provider is predicted but not the service name
→ **Partial identification.**
- If neither service provider nor the service name are predicted
→ **Invalid identification**

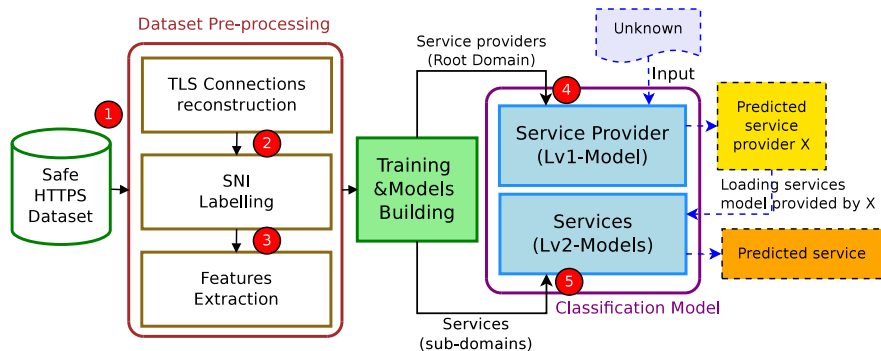


Figure : The work-flow of the HTTPS traffic identification framework



Evaluation Results

Overview

The evaluation of the proposed solution contains 3 parts:

- Evaluation of the collected dataset.
- Evaluation of the proposed features set.
- Evaluation of the multi-level classification approach.

Evaluation of the collected dataset

- Contains more than 288,901 HTTPS connections.
- Pre-processed to be suitable for multi-level approach.
- Processed to determine a reasonable threshold for the minimum number of labelled connections per service.

Optimized by Features Selection technique

- 18 features are highly relevant: 10 out of 12 from our proposed set and 8 out of 30 from the classical ones.
- This validates the rationale of the proposed features for identifying HTTPS services.

Table : The 18 selected features

Client ↔ Server
Inter Arrival Time (75th percentile)
Client → Server
Packet size (75th percentile, Maximum), Inter Arrival Time (75th percentile), Encrypted Payload(Mean, 25th, 50th percentile, Variance, maximum)
Server → Client
Packet size (50th percentile, Maximum), Inter Arrival Time (25th, 75th percentile), Encrypted payload(25th, 50th, 75th percentile, variance, maximum)



The proposed features set performance

By using WEKA ² tool the features set are tested by C4.5 and RandomForest algorithm:

- **Classical 30-features:**

C4.5 achieves $83.4\% \pm 1.0$ Precision,
RandomForest achieves **$85.7\% \pm 0.4$** Precision.

- **Full 42-features:**

C4.5 achieves $86.65\% \pm 0.7$ Precision,
RandomForest achieves **$87.82\% \pm 0.68$** Precision.

- **Selected 18-features:**

C4.5 achieves $85.87\% \pm 0.64$ Precision,
RandomForest achieves **$87.60\% \pm 0.10$** Precision.

²www.cs.waikato.ac.nz



HTTPS Identification Framework

- The framework has been evaluated in two steps:
 - Evaluate each level separately, to measure the performance of each classification model.
 - Evaluate the whole framework as one black box.
- Evaluation conditions:
 - Full features set (42 features).
 - RandomForest as ML algorithm.
 - At least 100 connections number per service.
 - K-Fold cross validation with $k=10$.

The Multi-level Classification Approach Evaluation

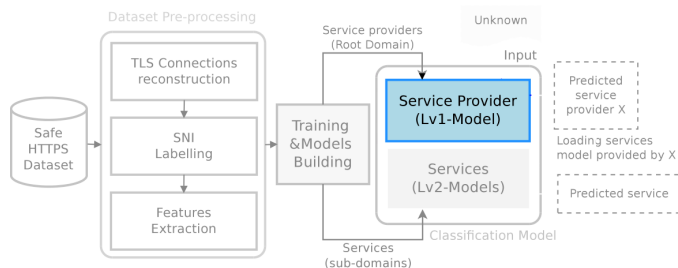


Figure : Top Level of the framework

Top level evaluation

Experiments show that we can identifying the service provider of HTTPS traffic with 93.6% overall accuracy.

The Multi-level Classification Approach Evaluation

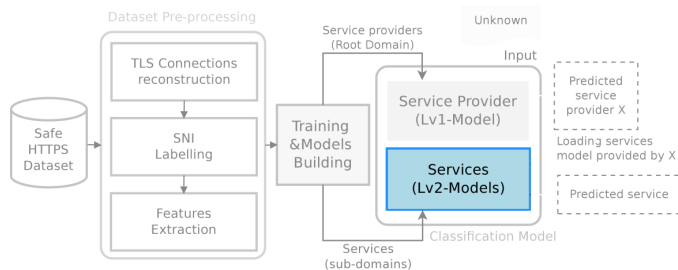


Figure : Second Level of the framework

Second level evaluation

A separate classification models are built and evaluated for each service provider with the same approach used in the Top-level.

Second level evaluation

- From 68 distinct service providers, 51 service providers have more than 95% of good classification of their own different services.
- For example, we can differentiate between 19 services run under Google.com, with 93% of Perfect identification.

Table : The second level models accuracy

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

The Multi-level Classification Approach Evaluation

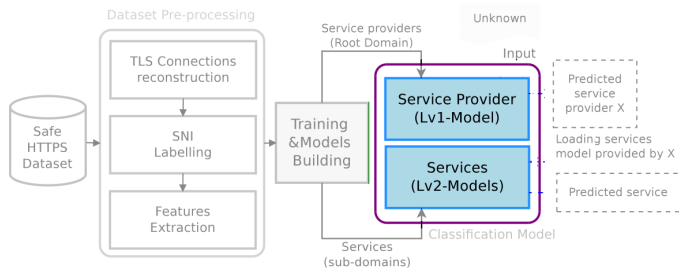


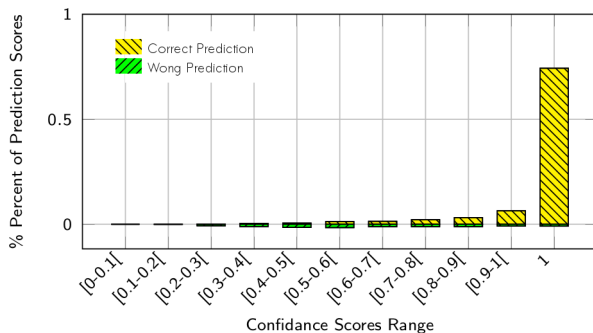
Figure : The complete classification model

Evaluate the framework as black-box (Level1&2)

Results show that we achieve 93.10% of Perfect identification and 2.9% of Partial identification.



The Multi-level Classification Approach Evaluation



The confidence score

- Measures the level of agreement between decision trees.
- Results shows that 86.68% of the predictions are in the sub-ranges $[0.8-0.9[$, $[0.9,1[$ and 1.

The classification errors over time

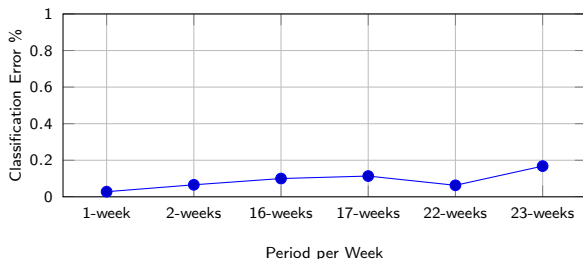


Figure : Effect upon classification error over time

Result

We can notice that even after 23 weeks without new learning phase, we still identify 80% (error <20%) of HTTPS services.



Conclusion & Future work

Conclusion

- A complete framework to identify the HTTPS services with several innovations (Multi-level classification, SNI-labelling, new set of features, without decryption).
- Based on real traffic, the results show that despite the challenging task, a high level of accuracy of 93.10% achieved.

Future Work

- To adapt and extend our current framework for real-time analysis identification of HTTPS services.
- Improve the global security of networks especially by developing a HTTPS firewall.

References

- [1] W. M. Shbair, T. Cholez, A. Goichot, and I. Chrisment, "Efficiently bypassing sni-based https filtering," in *Integrated Network Management(IM),2015 IFIP/IEEE International Symposium on*, pp. 990–995, IEEE, 2015.
- [2] W. de Donato, A. Pescape, and A. Dainotti, "Traffic identification engine: an open platform for traffic classification," *Network, IEEE*, vol. 28, no. 2, pp. 56–64, 2014.
- [3] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management*, vol. 25, no. 5, pp. 355–374, 2015.
- [4] B. Miller, L. Huang, A. D. Joseph, and J. D. Tygar, "I know why you went to the clinic: Risks and realization of https traffic analysis," in *Privacy Enhancing Technologies*, pp. 143–163, Springer, 2014.
- [5] Y. Okada, S. Ata, N. Nakamura, Y. Nakahira, and I. Oka, "Comparisons of machine learning algorithms for application identification of encrypted traffic," in *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, vol. 2, pp. 358–361, IEEE, 2011.
- [6] Y. Kumano, S. Ata, N. Nakamura, Y. Nakahira, and I. Oka, "Towards real-time processing for application identification of encrypted traffic," in *Computing, Networking and Communications (ICNC), 2014 International Conference on*, pp. 136–140, IEEE, 2014.