RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité

Striking Back at Cobalt: Using Network Traffic Metadata to Detect Cobalt Strike Masquerading Command and Control Channels

Clément Parssegny, Johan Mazel, Olivier Levillain, Pierre Chifflier
ANSSI (French Cybersecurity Agency), Télécom Sud Paris

August 12th 2025

Outline

1. What is Cobalt Strike?

Off-the-shelf pentest frameworks features

Definition

Pentest framework can perform major kill-chain steps including:

- Payload generation
- Command and Control (C2)
- Many commands for reconnaissance, code injection, data exfiltration, etc.
- Audit report generation

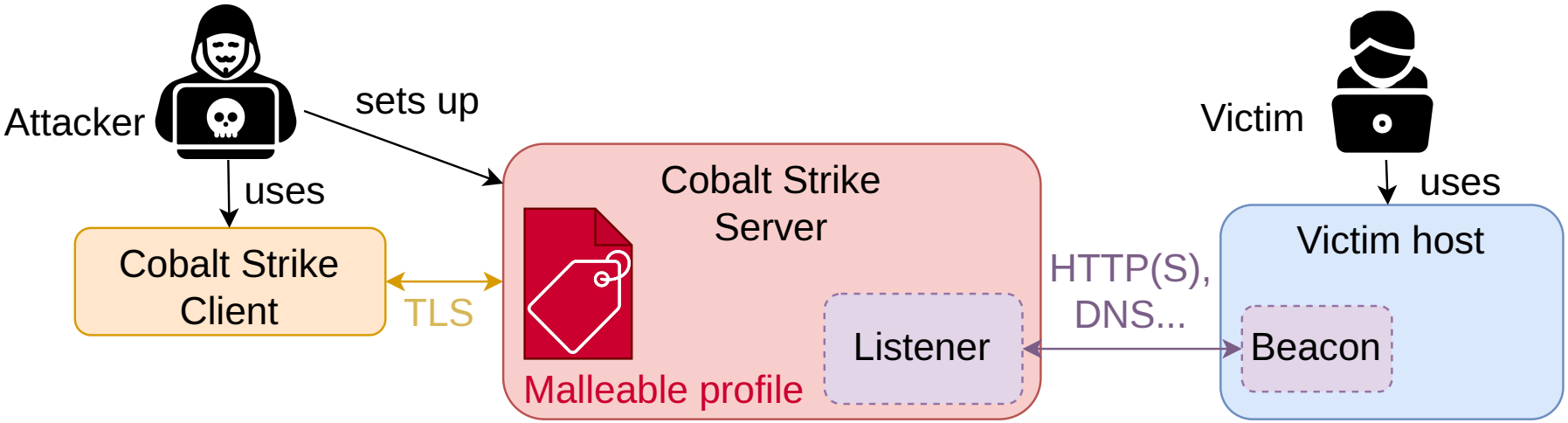
Examples: Brute Ratel, **Cobalt Strike**, Havoc, Sliver...

Related works: Detection of Cobalt Strike

- Motivation: well-known attackers use it (ex: Conti, Mustang Panda, Nobelium) [1]
- Several academic papers:
  - VanDerEijk et al. [2]
  - Ramos et al. [3, 4]
  - Yang et al. [5]
- Technical reports:
  - Talos [6]
  - BlackBerry [7]

Cobalt Strike C2: Architecture

Johan *Cobalt Strike C2: architecture ?* **Clément** *ok*



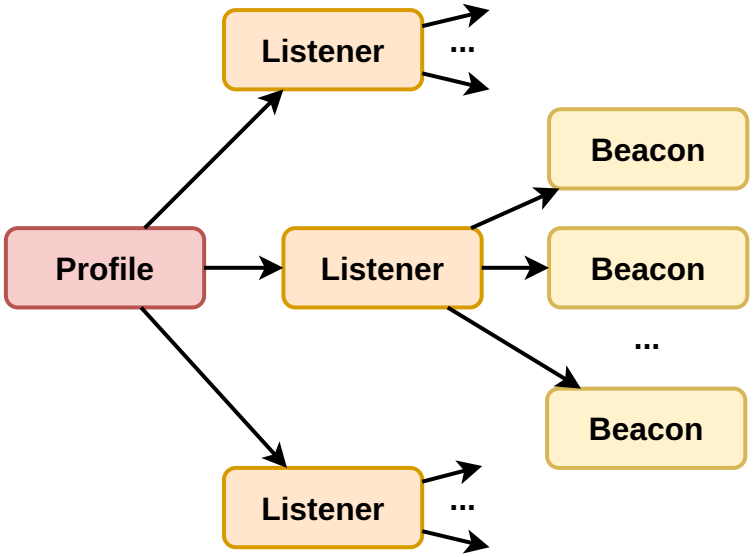
Cobalt Strike C2: Malleable profile

Johan *Cobalt Strike C2: malleable profile ?* **Clément** *ok*

Definition

Configuration file that contains rules for communication between the *Listener* and the *Beacons* (and more...).

Parameter examples: *sleeptime*, data obfuscation, URI, TLS certificate, post-exploitation behavior...



2. Detection of Cobalt Strike C2 traffic

Threat model

- The attacker establish C2 covert channels using HTTP, HTTPS or DNS
- We, defenders, passively monitor network traffic
- No application layer analysis

Objective and method

Research question

Can standard **Johan** *"or classic" ?* **Clément** *dans le papier on emploie standardnetwork features (duration, packet size and direction...) be used to differentiate Cobalt Strike C2 traffic from benign traffic ?* **Johan** *hum on utilise pas vraiment directement "packet size and direction" donc il faudrait reformuler, non ?* **Clément** *la direction est utilisée car c'est une feature par direction. Pour la durée, elle est dans les netflow donc à ce stade de la réflexion elle est incluse.*

Our approach needs:

- Data
- Machine learning tools (binary supervised classification)
- Features

Datasets

- NCC's collected beacons dataset [8]
- Cobalt Strike C2 network traffic generation platform
- Benign network traffic datasets from Stratosphere IPS [9], UPC [10] and UPN [11]
- Real world Cobalt Strike C2 traffic [12]

Machine learning

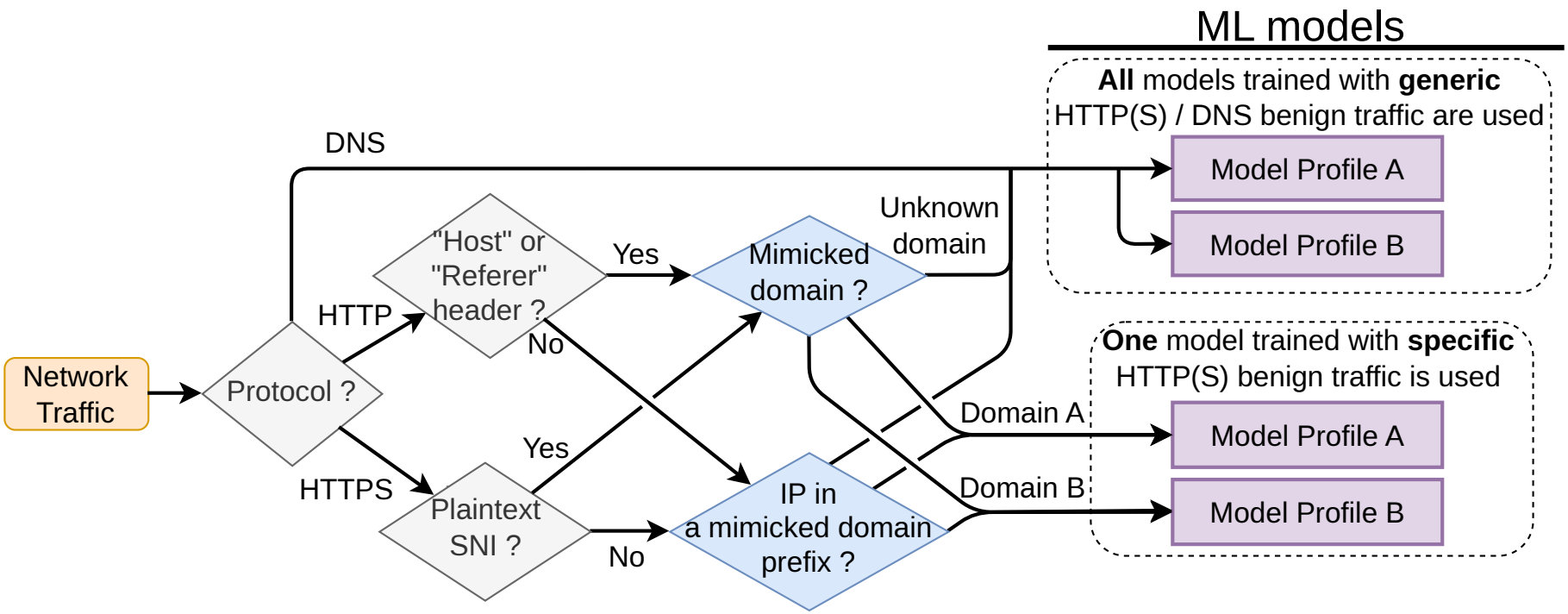
- A baseline to compare with: "*A machine learning based approach to detect stealthy cobalt strike C&C activities from encrypted network traffic*", Ramos et al. [3].
- An algorithm: Random Forest with hyperparameter tunning
- A validation method: stratified k-fold cross-validation
- A performance metric: F_1 score
- A feature importance metric: Mean Decreasing Impurity (MDI)

Features

Feature group name	Flow information							
	Nb packets	Packet size	OSI layer used for size	Duration	TCP flags	TCP history	Protocol	Service
Netflow v5	$\mathcal{B}$	total ($\mathcal{B}$)	L3	$\surd$	$\surd$	-	-	-
Netflow v9	$\mathcal{U}$	total ($\mathcal{U}$) minimum & maximum (S)	L3	$\surd$	$\surd$	-	-	-
Ramos et al.	$\mathcal{B}$	total ($\mathcal{B}$)	L4	-	-	$\surd$	$\surd$	$\surd$

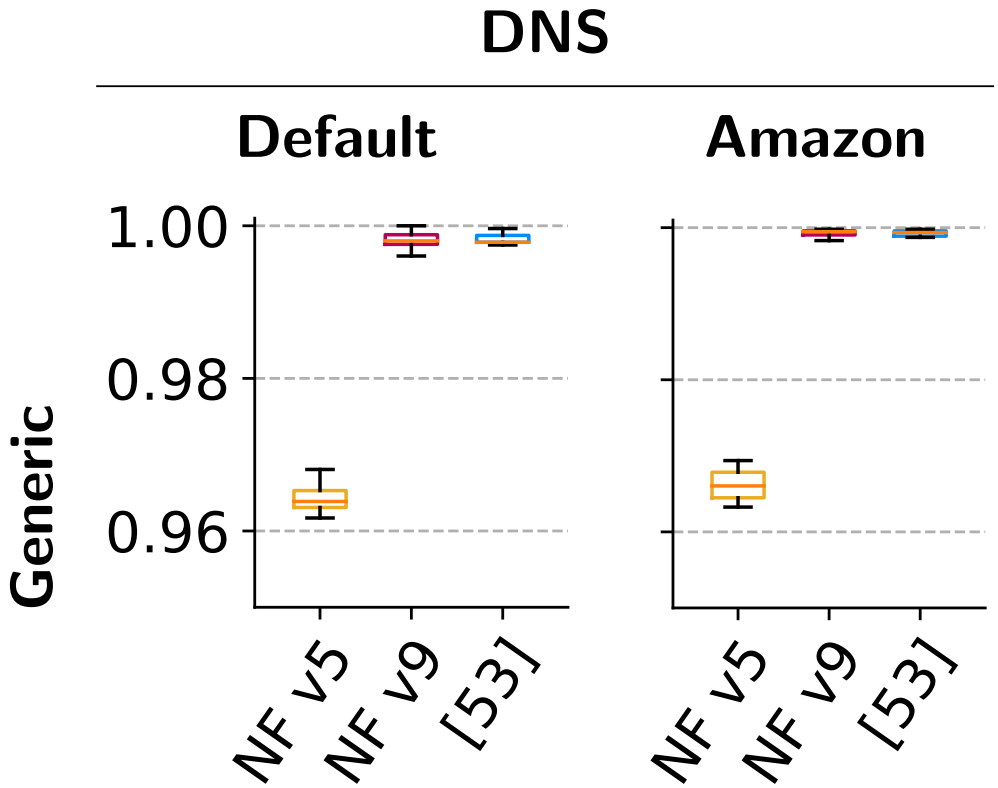
Table: $\mathcal{B}$: bi-direction $\mathcal{U}$: uni-directional S: sent packets only (Beacon to Listener) L3/4: Layer 3/4. Byte counts are based on the layer payload

Proposed method



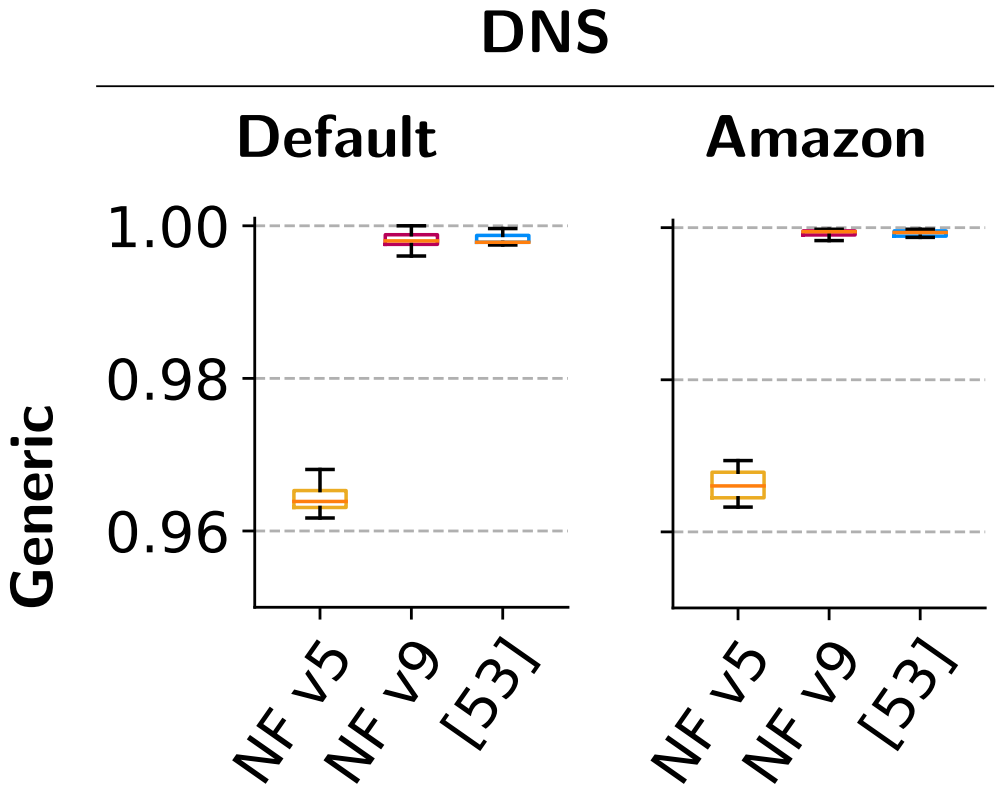
Results

Johan *tu mets les takeaway de ces slide en haut, mais en fait, c'est une conclusion que tu déduis de la figure, il faut le mettre en bas de la slide, avec "⇒" ? ou carément un bloc takeaway ? possiblement, ne pas l'afficher de suite ?* **Clément** *modifié*

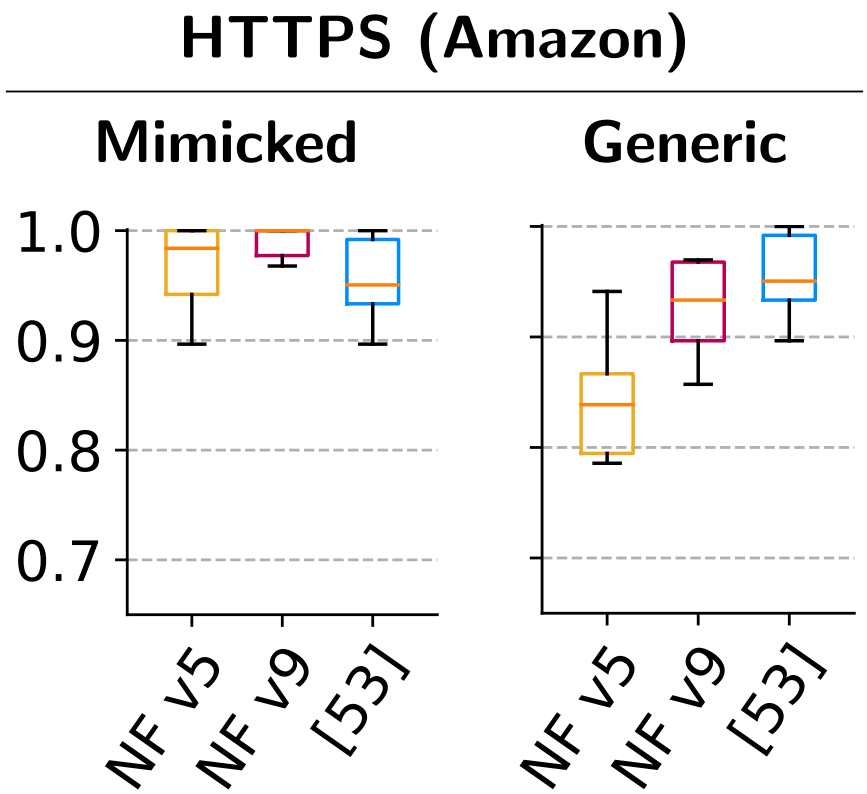


Results

Johan *tu mets les takeaway de ces slide en haut, mais en fait, c'est une conclusion que tu déduis de la figure, il faut le mettre en bas de la slide, avec "⇒" ? ou carément un bloc takeaway ? possiblement, ne pas l'afficher de suite ?* **Clément** *modifié*

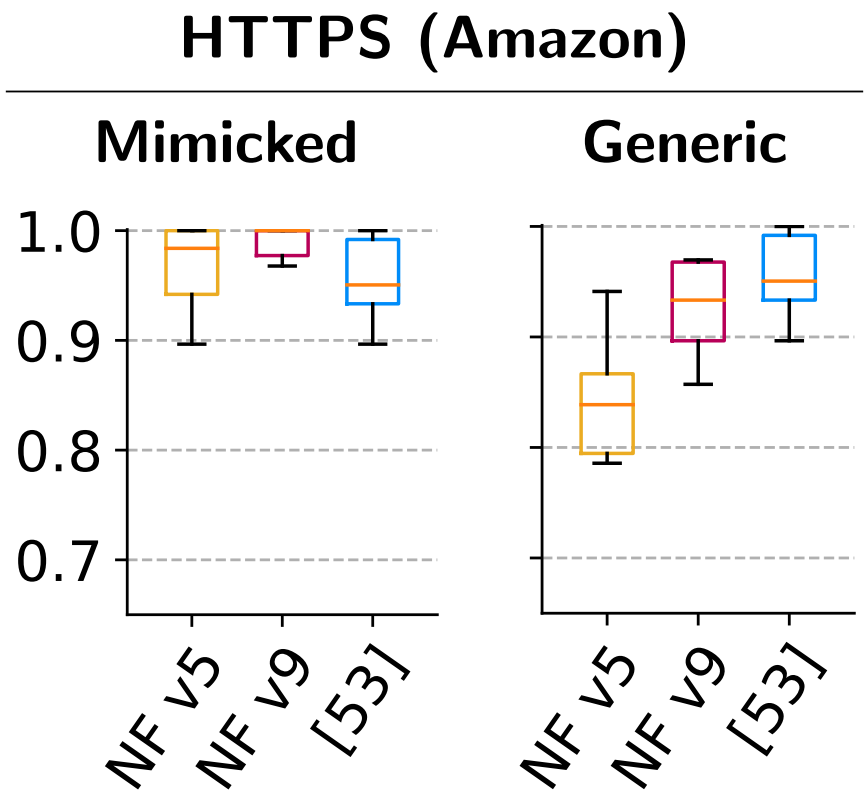


Results



[53] refers to the baseline.

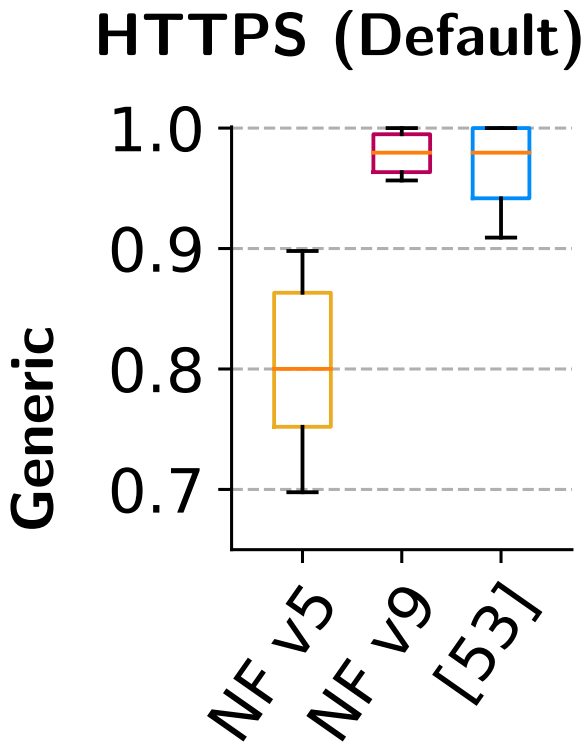
Results



⇒ Better performance if mimicked domain identified

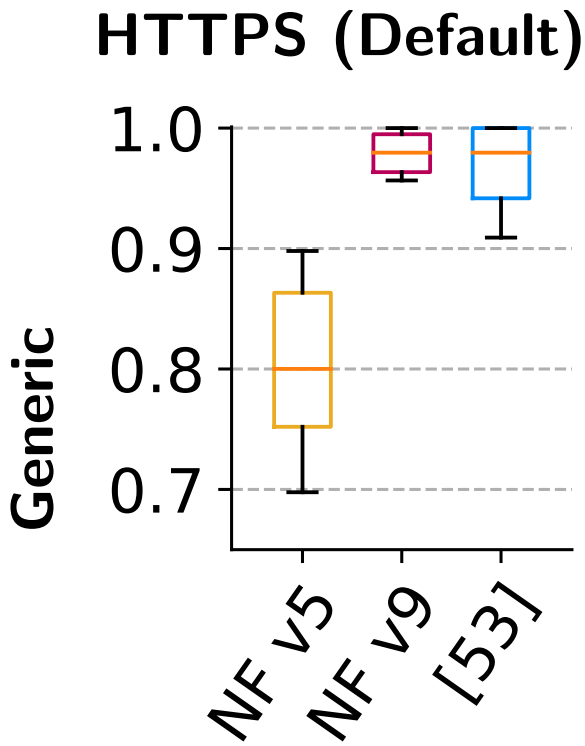
[53] refers to the baseline.

Results



[53] refers to the baseline.

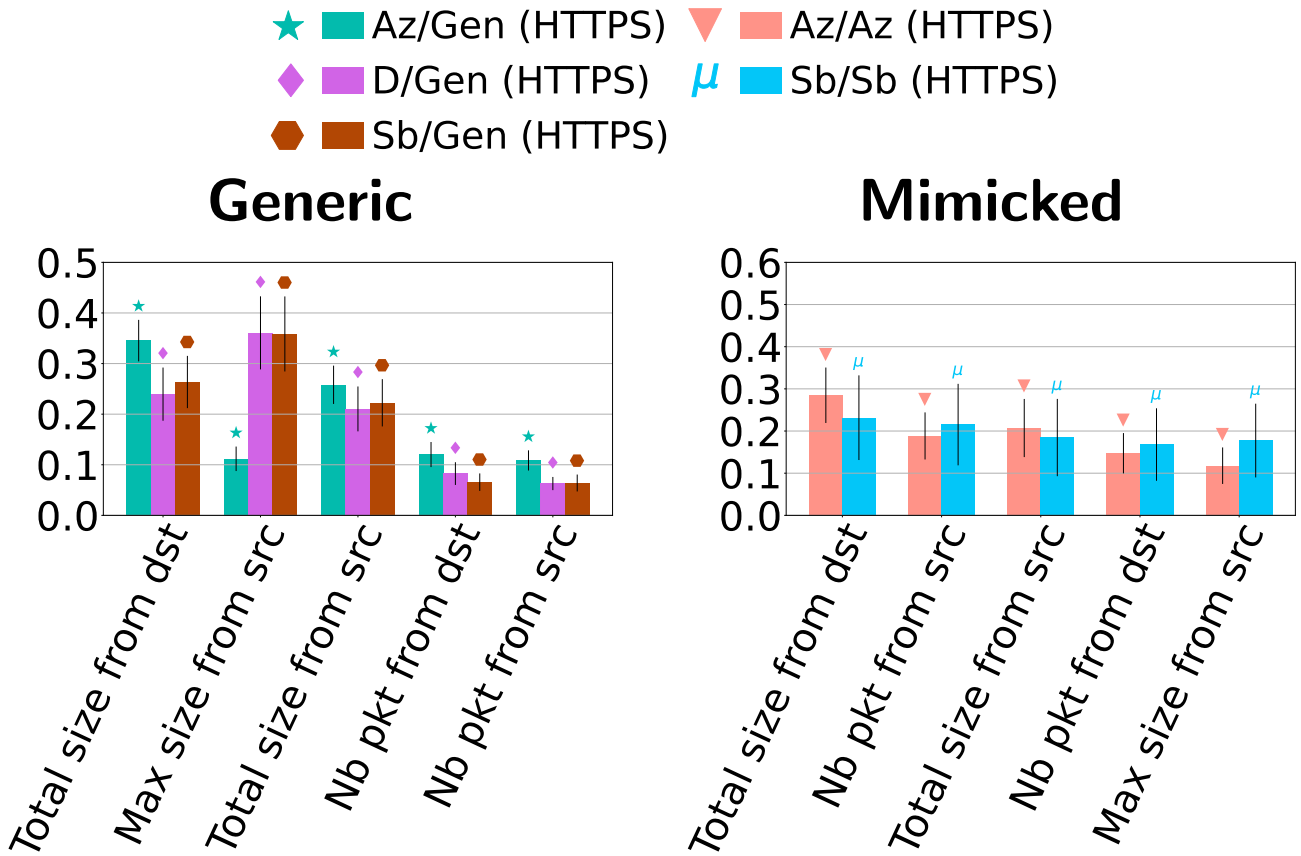
Results



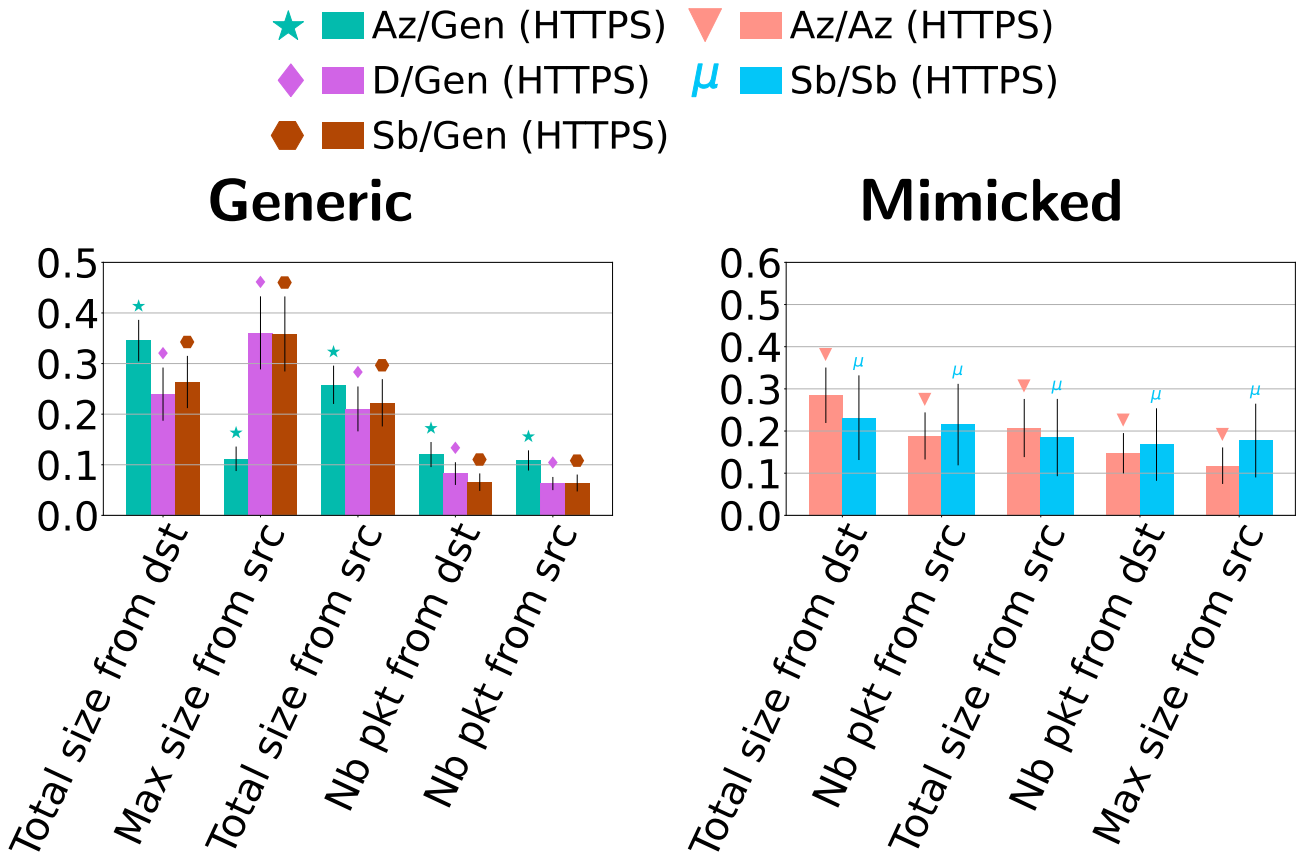
⇒ Netflow v5 is not that effective but Netflow v9 performs equally or better than the Ramos et al. baseline

[53] refers to the baseline.

Feature importance: Netflow v9 (HTTPS)



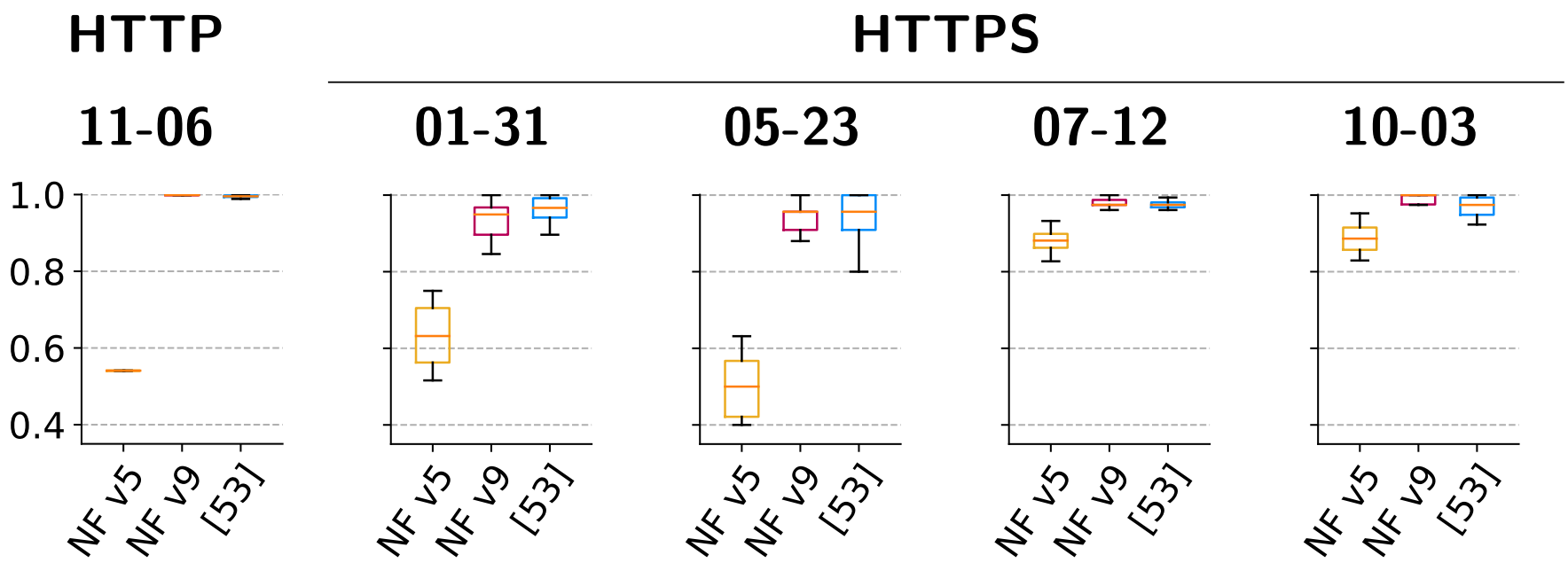
Feature importance: Netflow v9 (HTTPS)



⇒ Total size exchanged from the server has more importance in the decision

Detection on real-world Cobalt Strike traffic

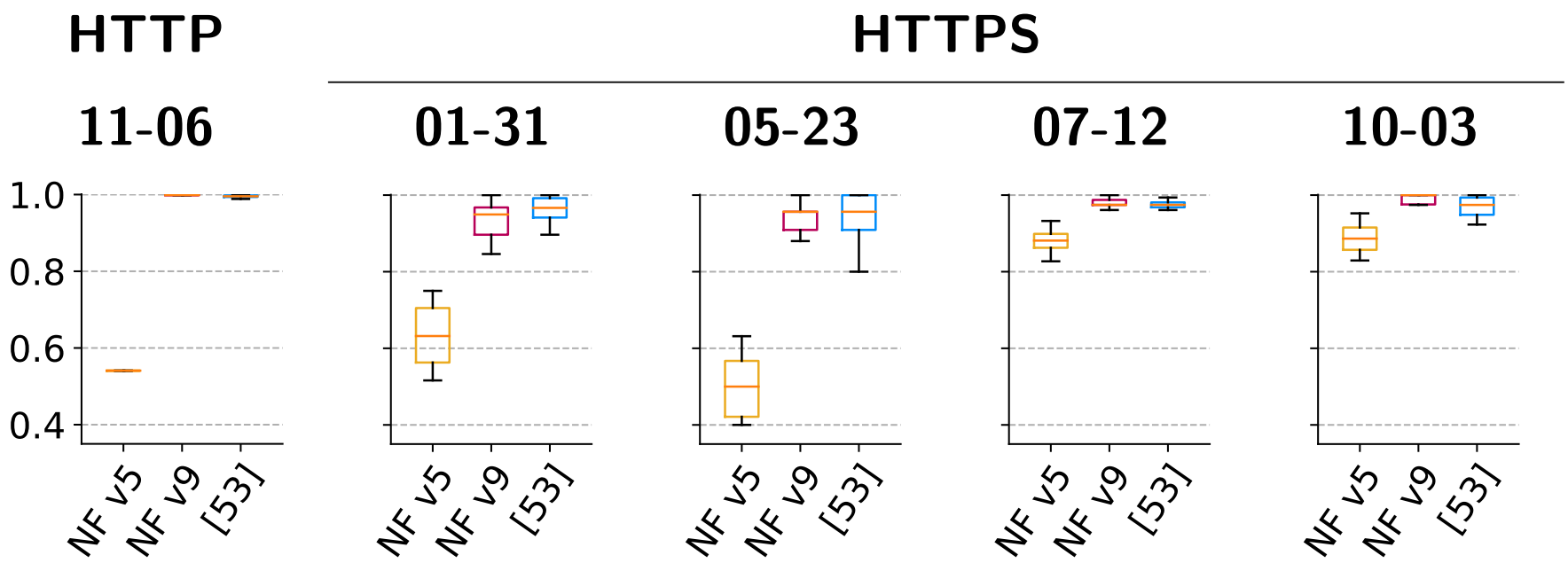
- 1 trace with HTTP and known domain
- 4 traces with HTTPS and **Johan** *and* **Clément** *okunknown* **Johan** *unknown ?* **Clément** *ok* domain



[53] refers to the baseline.

Detection on real-world Cobalt Strike traffic

- 1 trace with HTTP and known domain
- 4 traces with HTTPS and **Johan** *and* **Clément** *ok**unknown*** **Johan** *unknown ?* **Clément** *ok* domain



⇒ Encouraging results but not enough traces to be confident

[53] refers to the baseline.

3. Conclusion

Conclusion

- Detection of Cobalt Strike and other pentest frameworks is a challenge
- Standard network metadata features produce encouraging results
- Better performance is observed when the mimicked traffic is identified

Future work

- Evaluate the method on "data sending" profiles (Dropbox, videoconferencing, etc.)
- Evaluate the generalization capability of the method
- Extend the method to other C2 frameworks

Questions ?

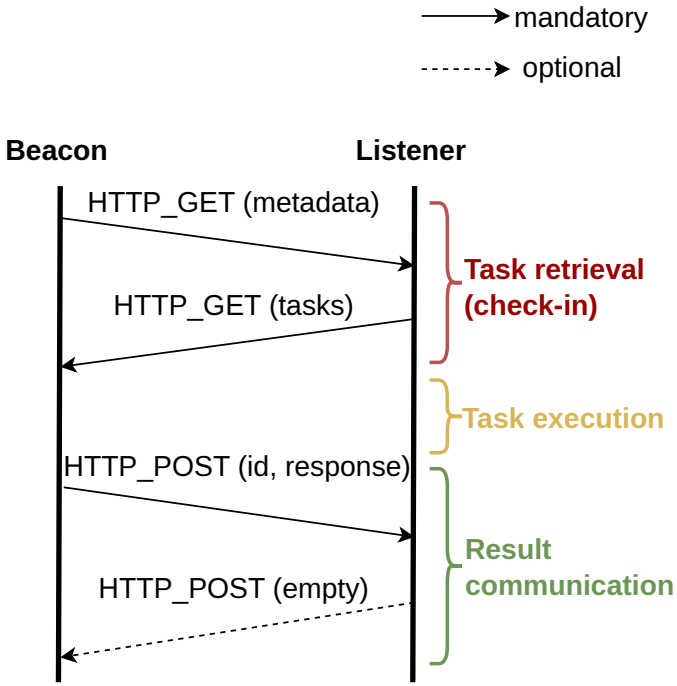
Johan *à merger avec conclusion ?* **Clément** *je parle pas dans cette slide.*
C'est juste pour afficher les points à retenir pendant les questions

Takeaway

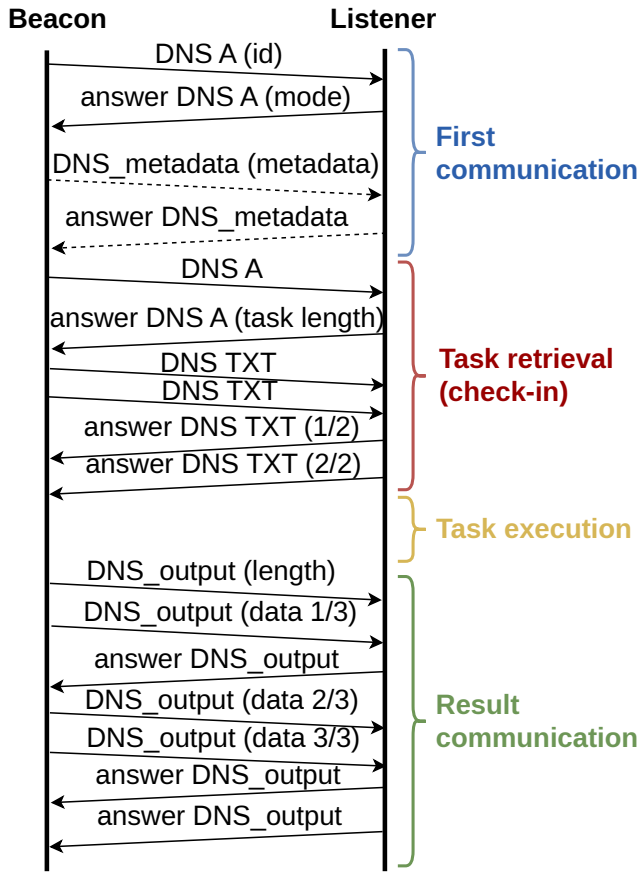
- Standard network features allows detection despite mimicking capability
- Determining the domain mimicked optimize the performance
- Method tested on real-world traffic
- Artifacts available at: `https://github.com/cp-tsp/ares2025-artifacts`

Appendix

Example of communication



(a) HTTP

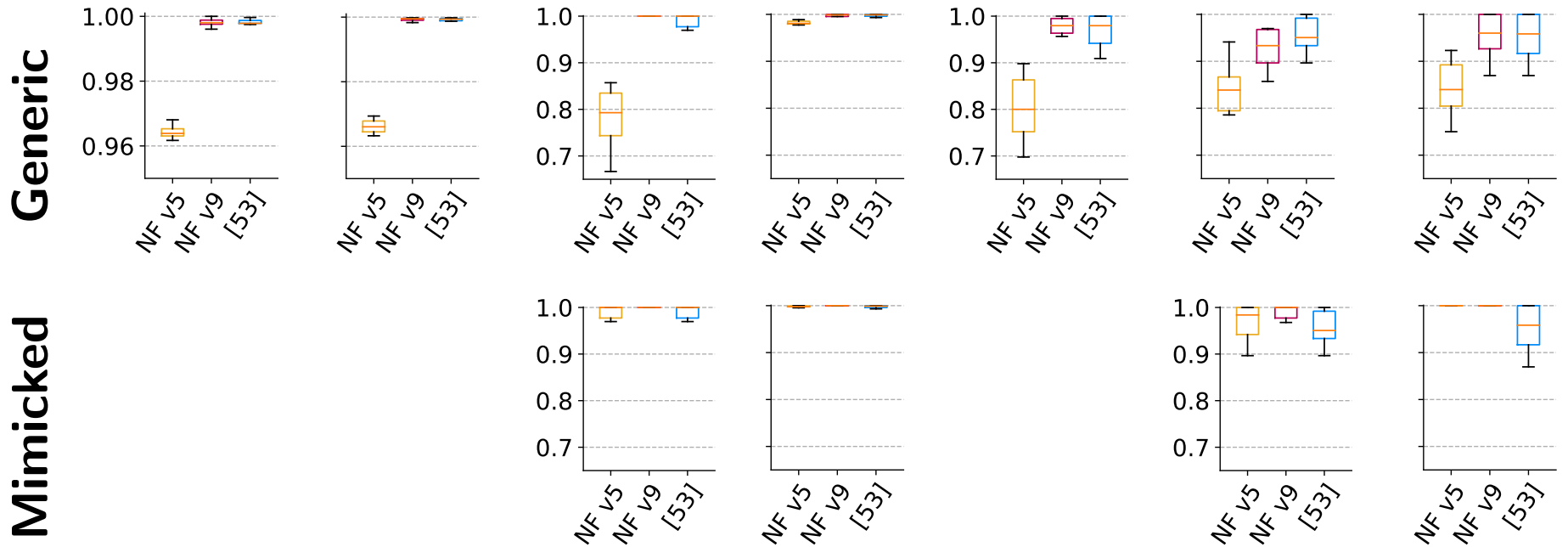


(b) DNS in TXT mode.

Most common domains mimicked in the NCC's dataset

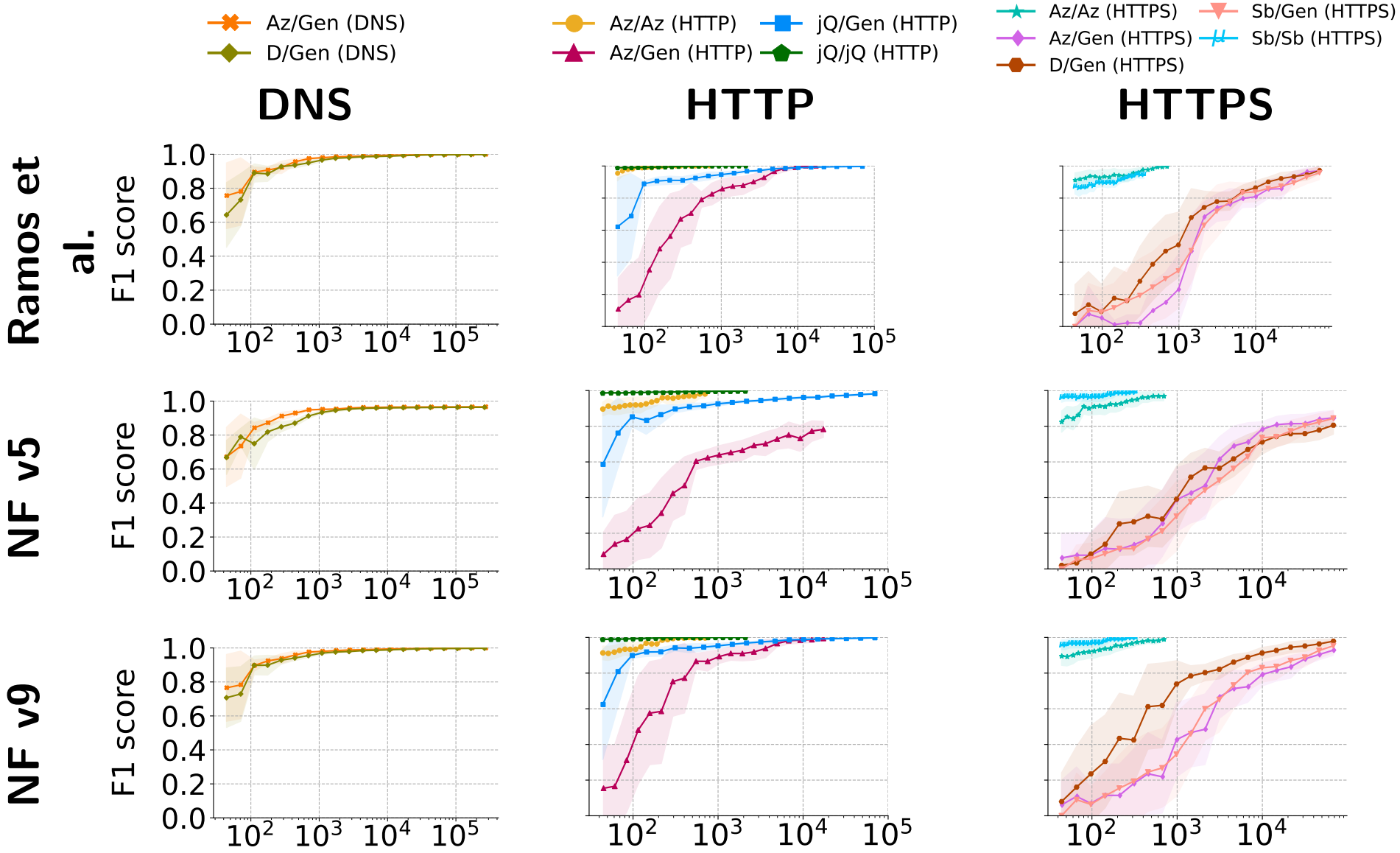
Spoofed domain	Nb of TLSH-based grouped profiles	Nb of instances	Nb of unique profiles
code.jquery.com	30	4 949	601
www.amazon.com	36	3 662	243
download.windowsupdate.com	42	993	145
locations.smashburger.com	2	827	3
www.google.com	40	722	195
www.bing.com	29	548	92
ocsp.verisign.com	24	520	54
www.microsoft.com	33	360	67
onedrive.live.com	15	316	63
audio-sv5-t1-3.pandora.com	7	281	38

Boxplots

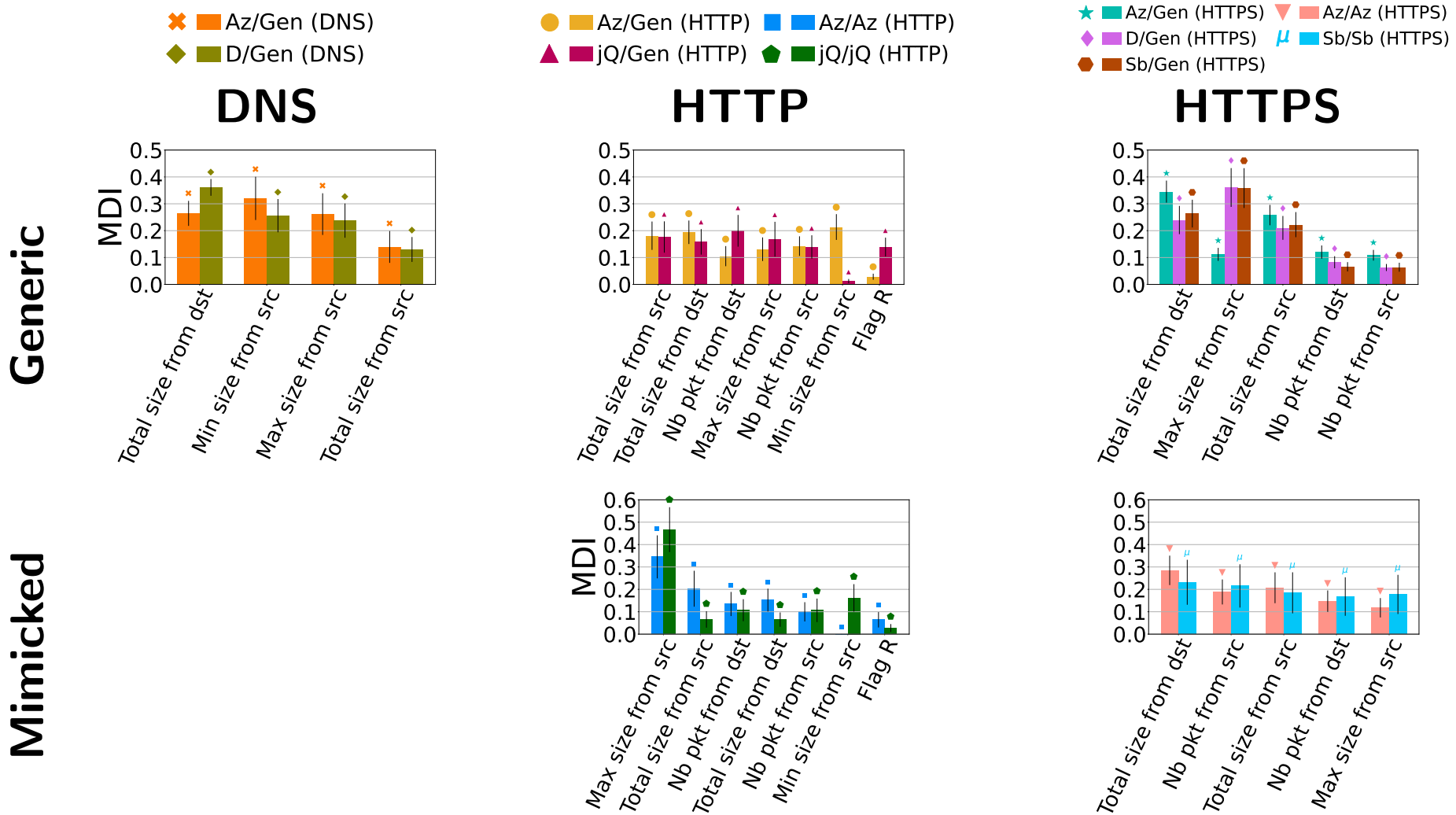


[53] refers to the baseline.

Learning curves



Feature importance: Netflow v9



Bibliography

[1] MitreAtt&ck. https://attack.mitre.org/software/S0154/.

[2] Coen Schuijt Vincent van der Eijk. *Detecting Cobalt Strike beacons in NetFlow data*. 2020.

[3] Fabian Martin Ramos and Xinyuan Wang. "A Machine Learning Based Approach to Detect Stealthy Cobalt Strike C&C Activities from Encrypted Network Traffic". In: *Machine Learning for Networking*. 2022.

[4] Fabian Martin Ramos and Xinyuan Wang. "Detecting Stealthy Cobalt Strike C&C Activities via Multi-Flow based Machine Learning". In: *ICMLA*. 2023.

[5] Xiaodu Yang et al. "PETNet: Plaintext-aware encrypted traffic detection network for identifying Cobalt Strike HTTPS traffics". In: *Comput. Nets.* (2024).

[6] N.Mavis. *The art and science of detecting Cobalt Strike*. 2020.

[7] T.J.O'Leary et al. *Finding Beacons In The Dark*. 2021.

[8] Yun Zheng Hu. *Mining data from Cobalt Strike beacons*. https://www.nccgroup.com/us/research-blog/mining-data-from-cobalt-strike-beacons. 2022.

Bibliography (cont.)

[9] Stratosphere. *Stratosphere Laboratory Datasets*. https://www.stratosphereips.org/datasets-overview. 2015.

[10] UPC. https://historic.cba.upc.edu/monitoring/traffic-classification.html. 2015.

[11] Víctor Labayen et al. "Network traffic and code for machine learning classification". In: *Machine Learning for Networking*. 2020.

[12] MTA. https://www.malware-traffic-analysis.net/2023/.