Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems

Creach Lab - October 10^{th}, 2023 - Inria Rennes

Maxime Lanvin[2] Pierre-François Gimenez[2] Yufei Han[1] Frédéric Majorczyk[3] Ludovic Mé[1] Éric Totel[4]

[1]Inria, Univ. Rennes, IRISA, {firstname.lastname}@inria.com

[2]CentraleSupélec, Univ. Rennes, IRISA, {firstname.lastname}@centralesupelec.fr

[3]DGA-MI, Univ. Rennes, IRISA, frederic.majorczyk@intradef.gouv.fr

[4]Samovar, Télécom SudParis, Institut Polytechnique de Paris, eric.totel@telecom-sudparis.eu

Inria CentraleSupélec DGA TELECOM SudParis

Introduction on NIDS & motivation

Two paradigms for NIDS

Intrusion Detection

- Intrusion Detection Systems (**IDS**) offer a way to detect attacks and let operators react according to the alerts
- We focus in this work on Network IDS (**NIDS**)

Paradigms

- **Signature**-based : detection of signature associated with known attacks
- **Anomaly**-based : detection of deviation from a normal behavior

Comparison of the two paradigms

Signature based

Supervision

- ts: 2023-01-19T14:02:46.143Z

- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389

- signature: "ET ATTACK_RESPONSE Possible CVE-2021-44228 Payload via LDAPv3 Response"

- category: "Attempted Administrator Privilege Gain"
- severity: 1

- CVE: **CVE_2021_4422**

Anomaly based

Supervision

- ts: 2023-01-19T14:02:46.143Z

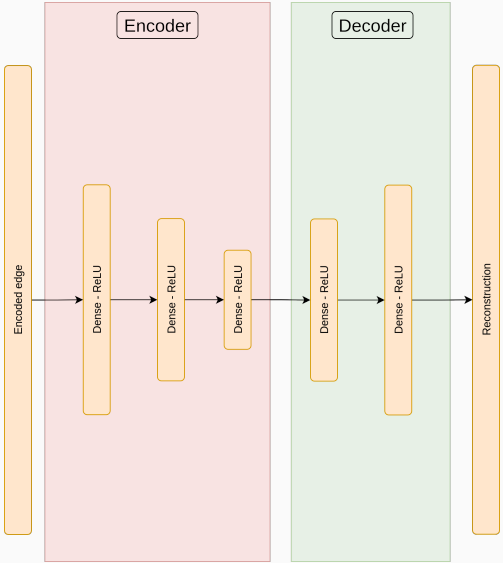
- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389

Good Luck ! Enjoy !

Summary

1. Introduction on NIDS & motivation
2. AE-pvalues
3. Benchmark XAI techniques
4. Using explanations on CICIDS2017 dataset
5. Conclusion

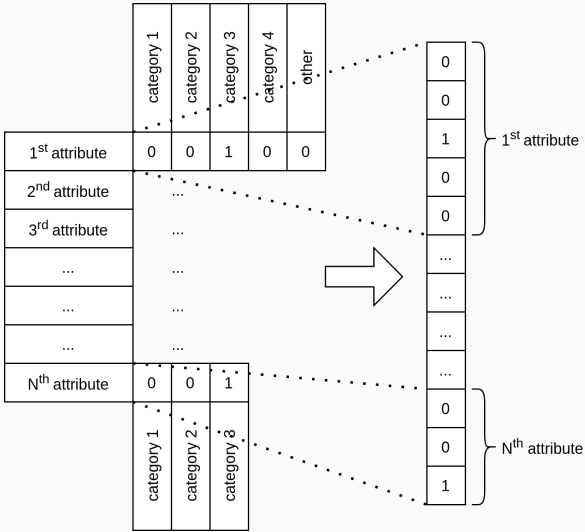
Unsupervised anomaly detection : Autoencoder



Learning

Minimisation of the reconstruction error between the input vector and its reconstructed version.

One Hot Encoding - Meaning of the vectors



Providing explanations

Goal

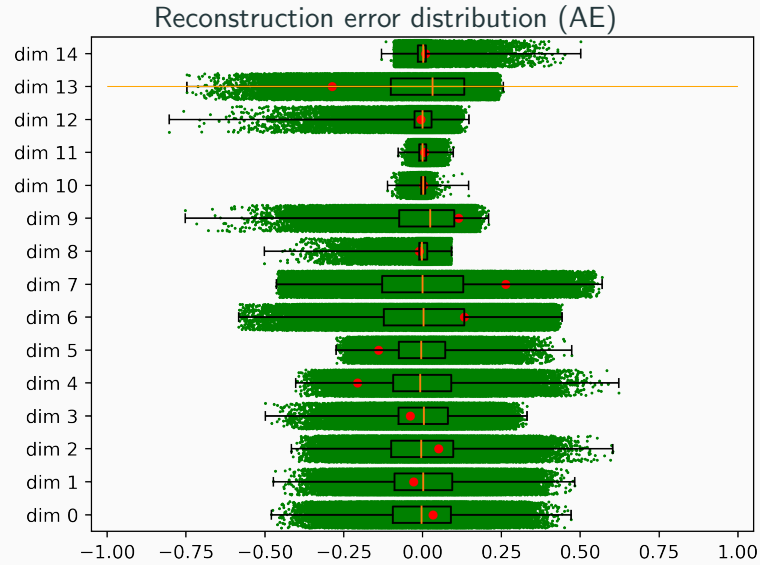
In our context providing explanations consists in providing the list of the **network attributes ranked** from the most abnormal to the least abnormal.

Example

[connection_duration, user_agent, ..., http_method, ..., destination_port]

AE-pvalues

XAI techniques for Autoencoders



Possible methods

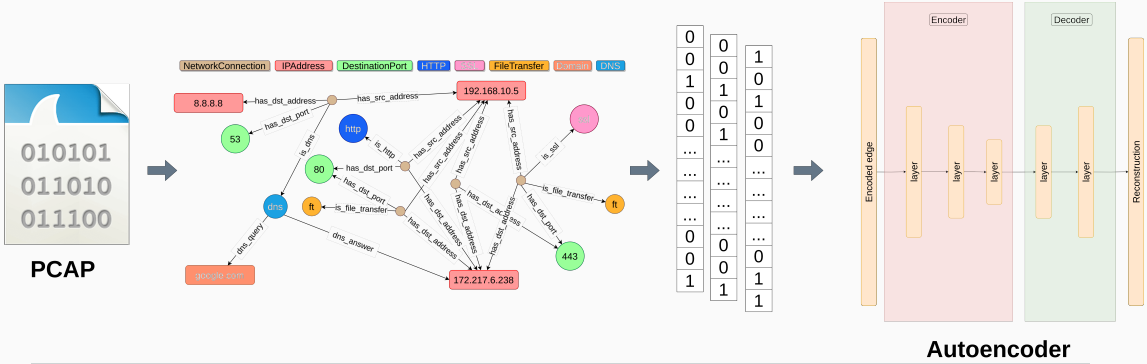
- Ranking by **absolute** values
- Ranking by **shapley** values
- Ranking by **p-values**

Obervation

The highest reconstruction error is not always an indication of the most abnormal dimension.

Benchmark XAI techniques

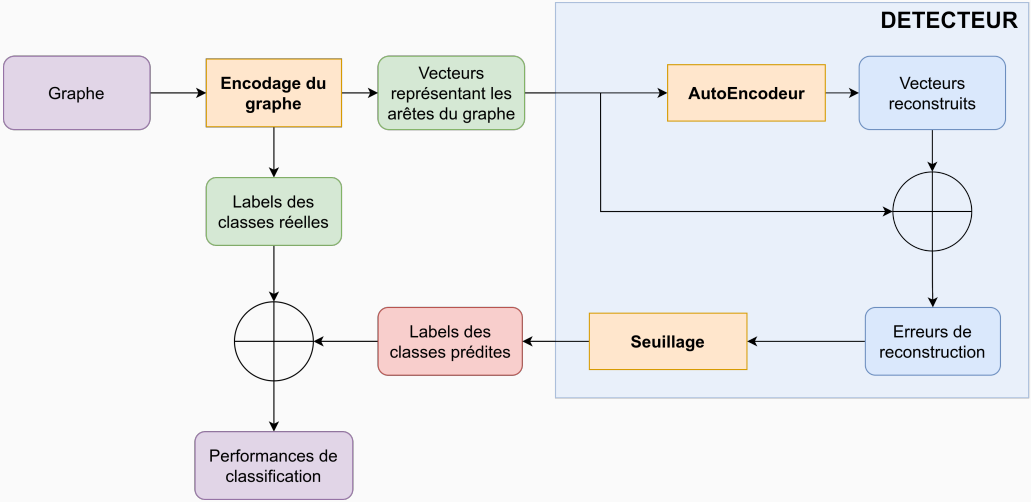
Sec2graph : An anomaly detection NIDS



Assumptions

- **Unsupervised** : no attacks used for the training
- **Anomaly-based** NIDS : detect drift from normal behaviours
- **The graph** helps to create a meaningful neighborhood

Anomaly Detection - Autoencoder

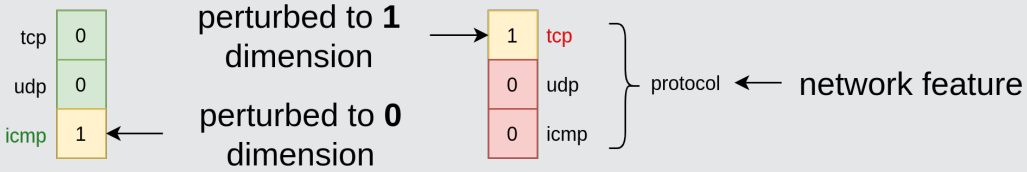


Methodology for the comparison

Methods

- Inject noise in a known network characteristic of vectors
- Assess ability of XAI methods to find the noisy network characteristic

Exemple of noise insertion in the protocol characteristic

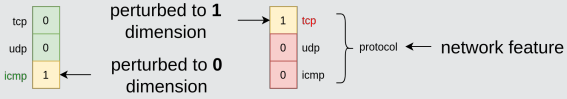


Evaluation : correlated attributes

http_status_code = 200 is equivalent to http_status_msg = OK

Benchmark results

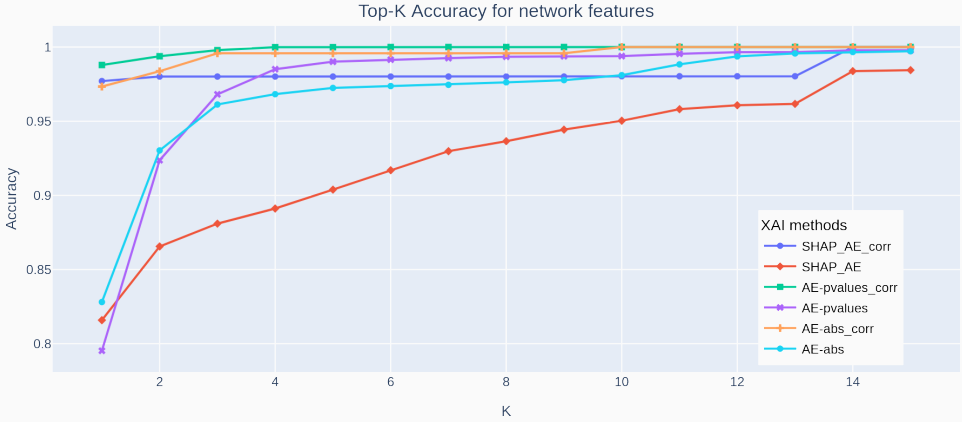
Vocabulary reminder



explaining method	Mean rank of the perturbed to 0 dimension	Mean rank of the perturbed to 1 dimension	Mean rank of the network feature ↓
AE-pvalues_corr	**2.96**	**1.63**	**1.02**
AE-abs_corr	3.89	1.61	1.07
SHAP_AE_corr	4.71	4.44	1.26
Random_corr	5.68	16.3	1.85
AE-pvalues	**4.61**	**3.07**	**1.39**
AE-abs	5.78	4.78	1.49
SHAP_AE	18.96	7.18	2.15
Random	26.93	27.13	7.8

Table of mean ranks of the perturbed to 0 or 1 dimensions, and the network feature where the noise is inserted.

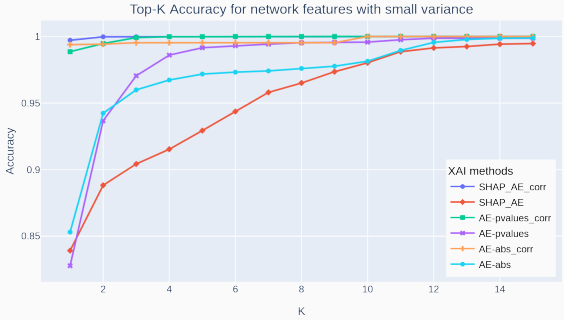
Benchmark results

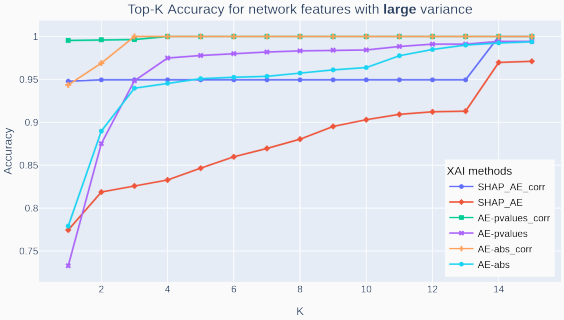


Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations.

Benchmark results





Benchmark results

Method	Processing time per sample
SHAP_AE	28 s
AE-pvalues	1.9 ms
AE-abs	1.0 ms

Processing time for one sample for each explaining method

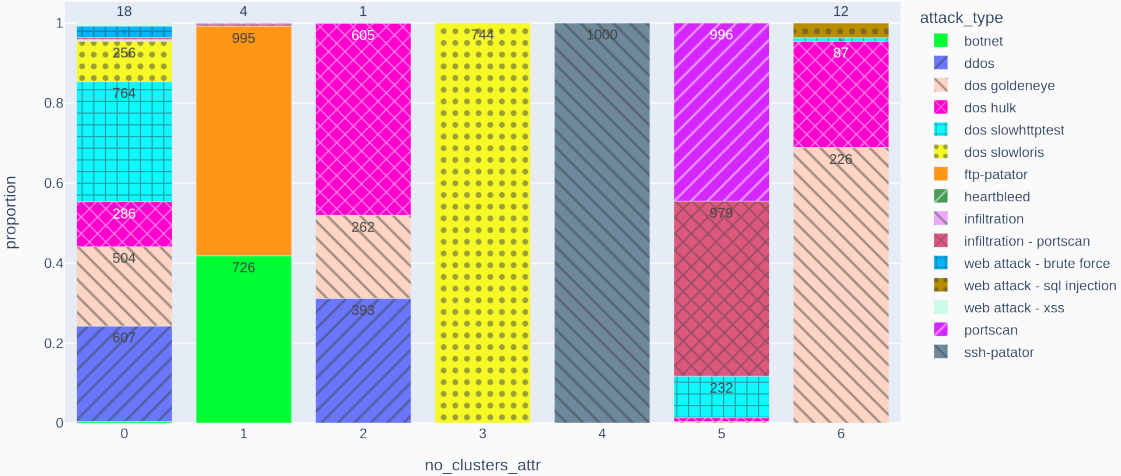
Conclusion

AE-pvalues is approximately $10,000$ faster than the SHAP_AE method.

Using explanations on CICIDS2017 dataset

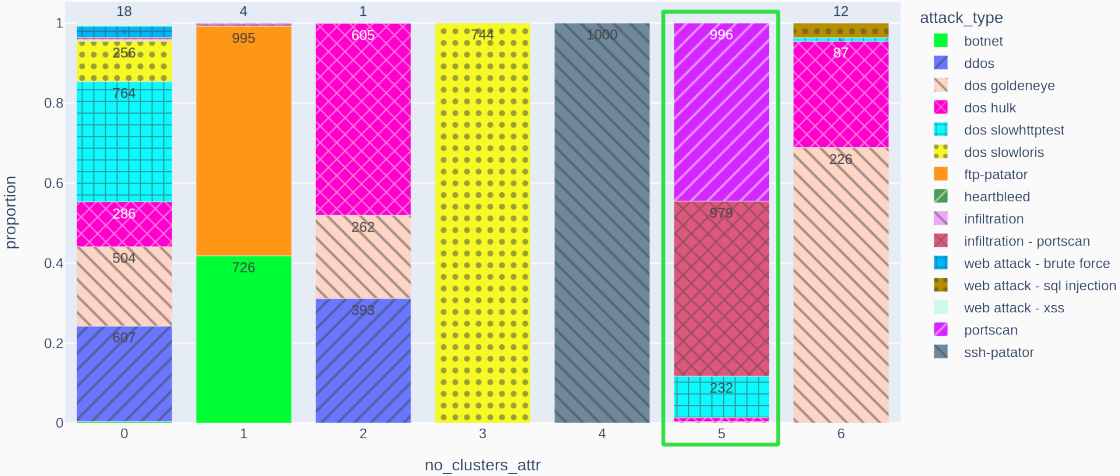
Principle

Clustering of the alerts based on the explanations

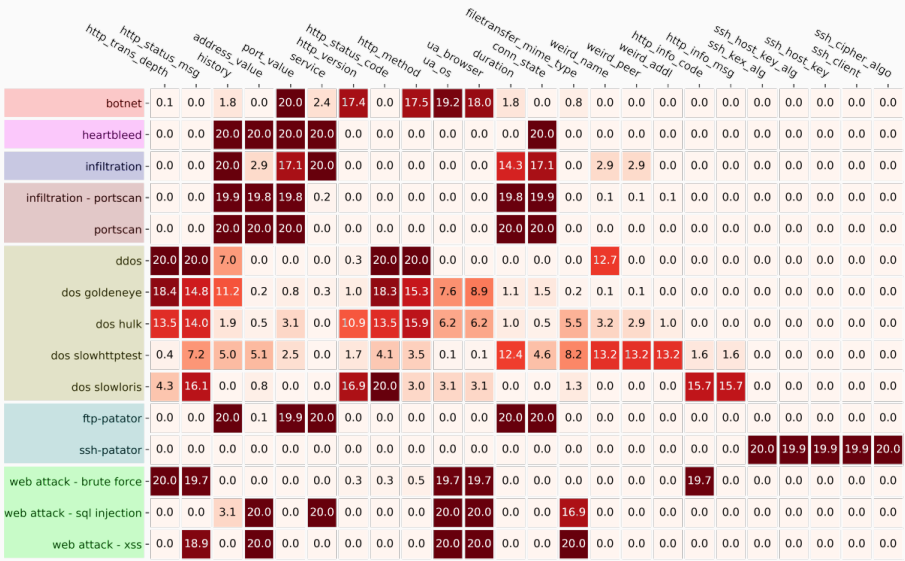


Principle

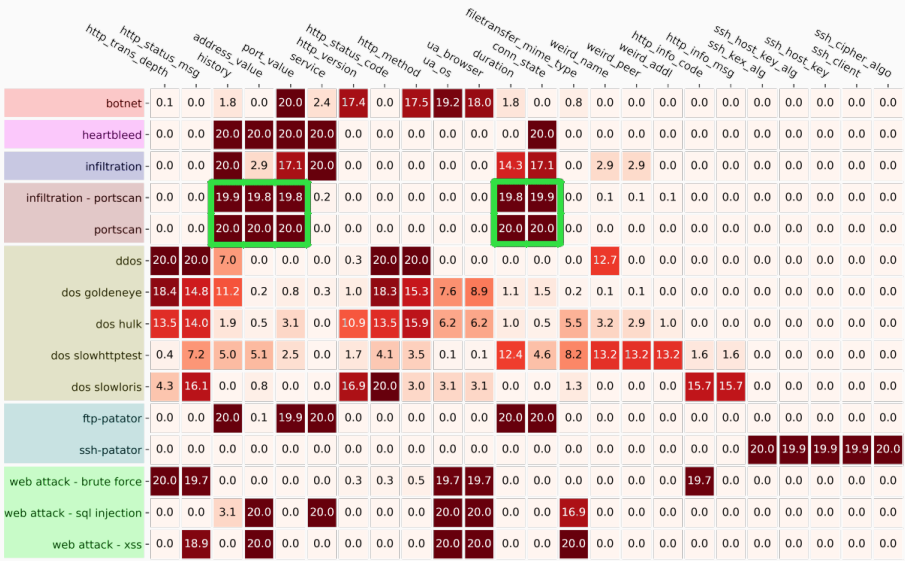
Clustering of the alerts based on the explanations



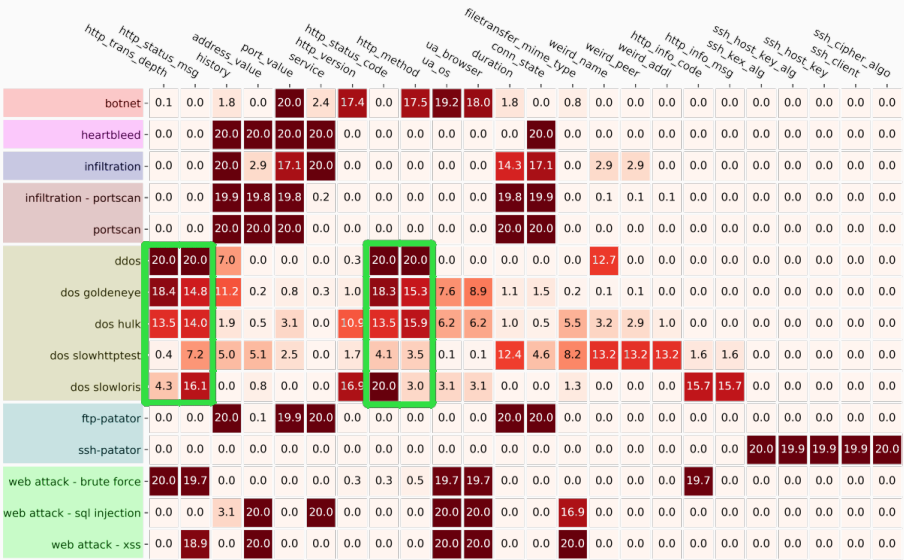
Applications - Feature contribution to attack types



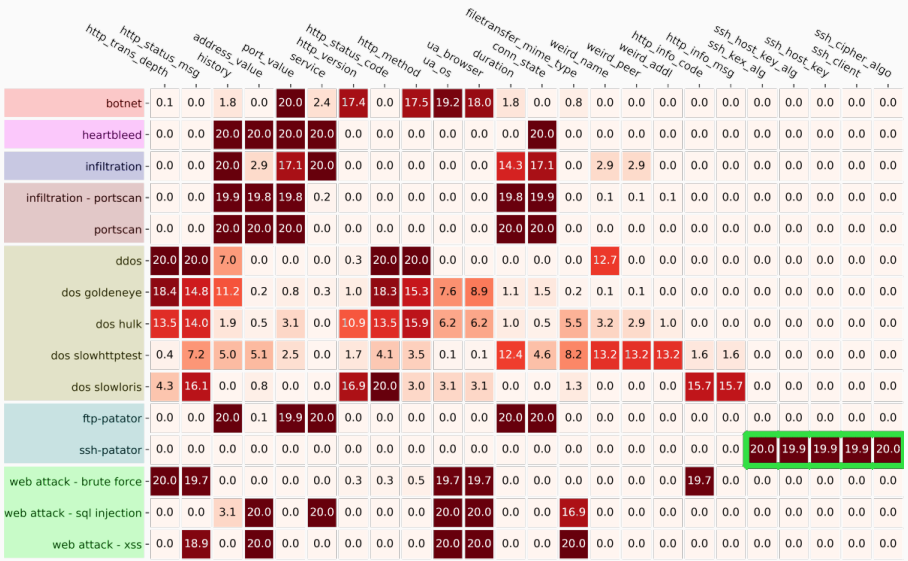
Applications - Feature contribution to attack types



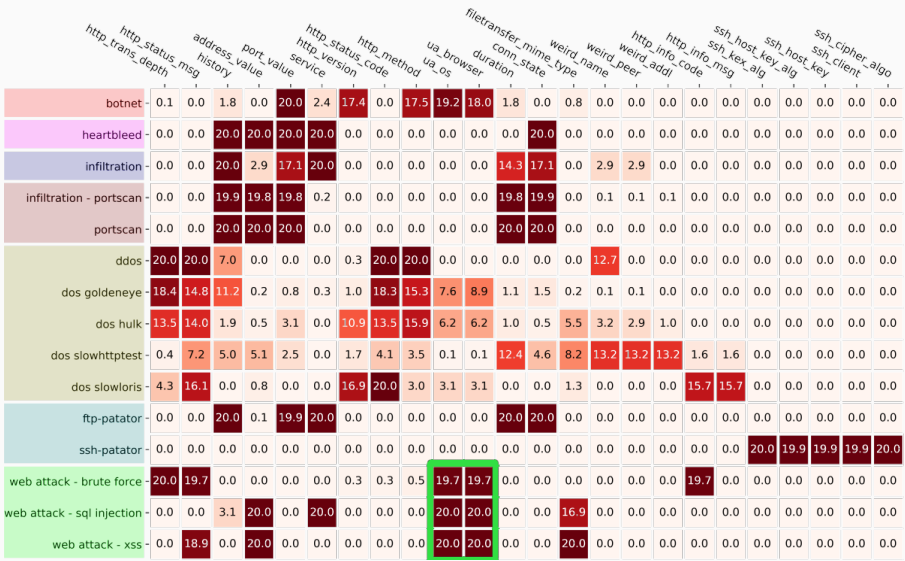
Applications - Feature contribution to attack types



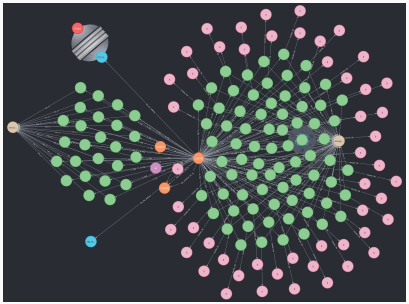
Applications - Feature contribution to attack types



Applications - Feature contribution to attack types



Applications - Forensic analysis - Web attack : Brute Force



single connection graph

network_feature	value
http_method	POST
http_referrer	http ://205.174.165.68/dv/login.php
http_request_body_len	130
http_status_code	302
http_status_msg	Found
http_trans_depth	84
user_agent_browser	Mozilla/5.0
user_agent_os	Linux x86_64

Top 5 explanations

user_agent_browser - user_agent_os - http_status_msg

http_status_code - http_trans_depth

Conclusion

Conclusion [1]

Summary :

- Explanation technique for alerts raised by AutoEncoder-based NIDS
- Clustering alerts based on explanations
- Help manual analysis

Future works

Leverage explanation techniques for the detection and alert triage

gitlab code for *AE-pvalues*

gitlab.inria.fr/mlanvin/ae-pvalues

1. Work accepted at RAID 2023 conference https://doi.org/10.1145/3607199.3607247