Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems

FADEX - October 19th, 2023 - Rennes

Maxime Lanvin
Frédéric Majorczyk

Pierre-François Gimenez
Ludovic Mé

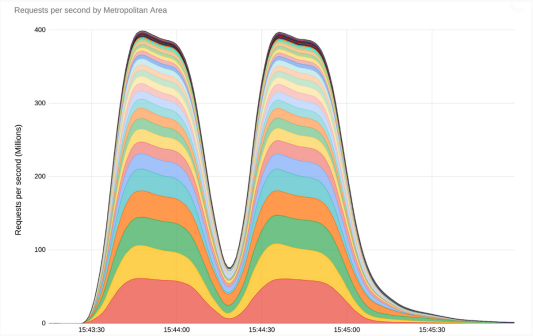
Yufei Han
Éric Totel

Introduction & motivation

Introduction

Systems are under attack

- Many untargeted, opportunistic attacks like password bruteforce
- Some targeted attacks with a huge power (e.g., DDoS attacks)
- Some very sophisticated attacks months or years in the making (SolarWinds, Stuxnet, TV5 Monde hack)



DDoS attacks against Google Cloud with 400 millions requests per second!

Introduction

How to protect?

- Prevention of attacks (password policy, updated systems, raising awareness against phishing, threat monitoring, etc.)
- Detection of attacks
- Reaction to attacks

Intrusion Detection

- Intrusion Detection Systems (**IDS**) offer a way to detect attacks and let operators react according to the alerts
- Two main paradigms: signature-based and anomaly-based detection
- We focus in this work on Network IDS (**NIDS**): we analyze network traffic

Comparison of the two paradigms

Signature based alert

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 alert

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 !

How to make ML models explainable?

Different techniques

- **Intrinsically explainable** models: decision tree, logistic regression, ...
- **Model-agnostic approaches**: local/global surrogate models: explain complex model using intrinsically explainable models: LIME, SHAP
- **Counterfactual analysis**: use examples around decision boundaries to explain decision

Most of these methods are adapted to supervised machine learning. Only one method works for anomaly detection (SHAP) but it's very slow

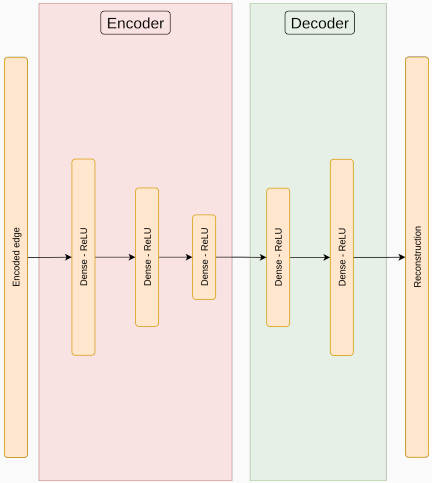
$\Rightarrow$ we introduce AE-pvalues, a new method faster and more accurate than SHAP, for explaining alerts

Summary

1. Introduction & motivation
2. AE-pvalues
3. Experiments with noise insertion
4. Experiment on CICIDS2017 dataset
5. Conclusion

AE-pvalues

Unsupervised anomaly detection: Autoencoder (AE)



Learning

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

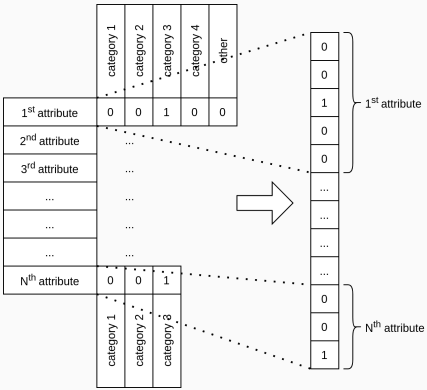
Detection

Raise an alert when the reconstruction error is above a threshold.

Goal

In our context, the explanations are an *ordered list of the network attributes* ranked from the most abnormal to the least abnormal.

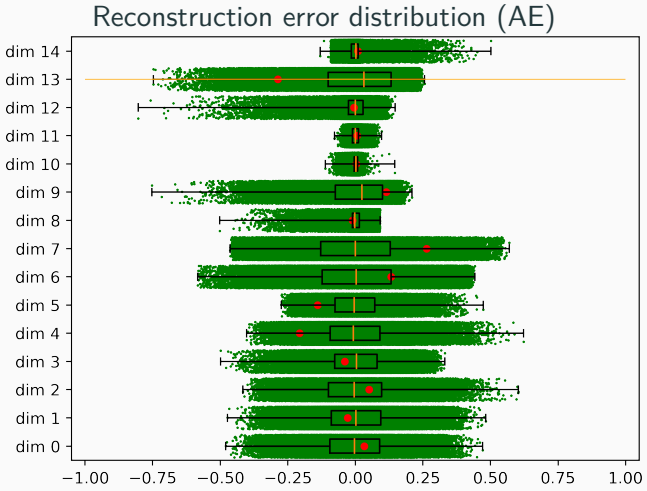
One Hot Encoding - Meaning of the vectors



Intuitive idea

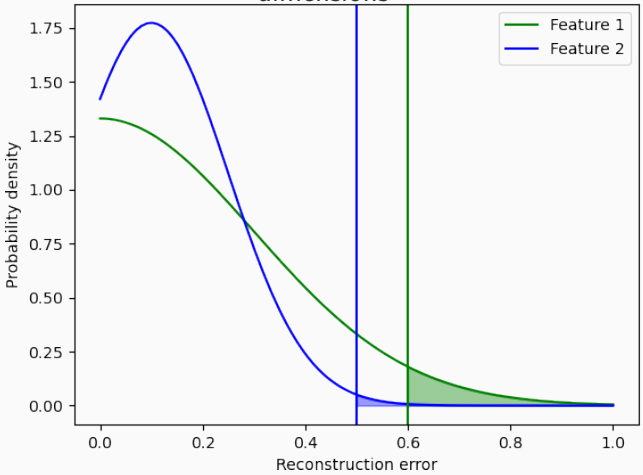
- When the reconstruction error is large, check the error dimension by dimension
- The higher the error of a dimension, the highest in the explanation list
- We call this method "AE-abs" and it has been proposed in the literature

What it looks like



Limitations

Comparison of the reconstruction errors of two dimensions

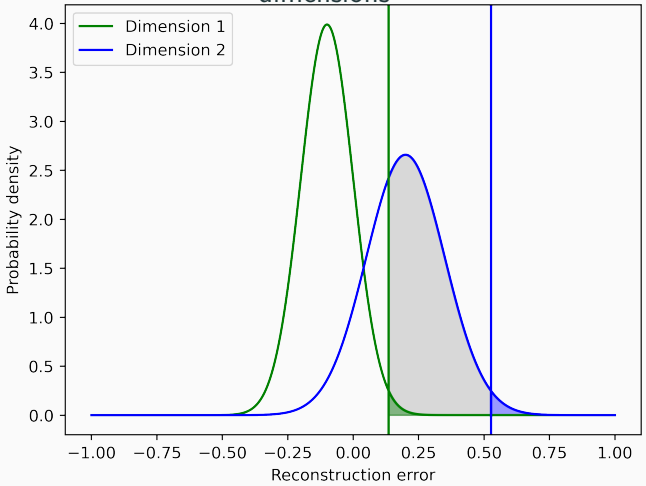


Key Idea

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

Principle

Comparison of the reconstruction errors of two dimensions



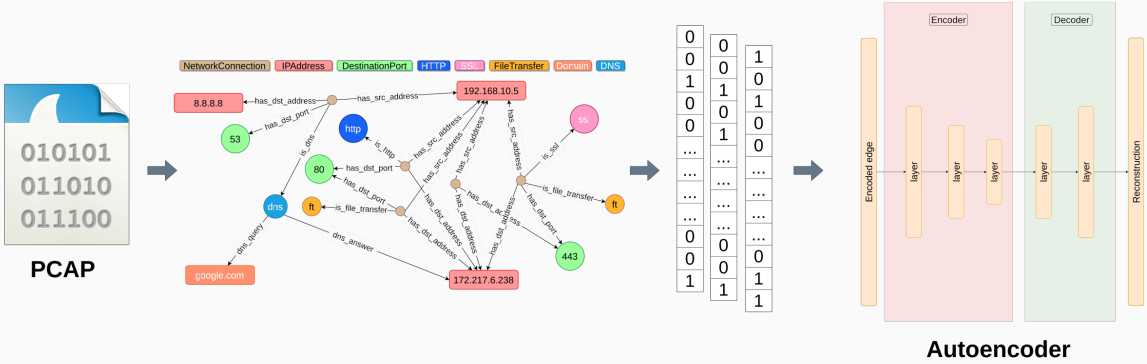
Our approach

This area is called the p-value:

$$p_i = \frac{\#\{r_i \geq e_i\}}{\#\{r_i\}}$$

Experiments with noise insertion

Sec2graph: An anomaly detection NIDS



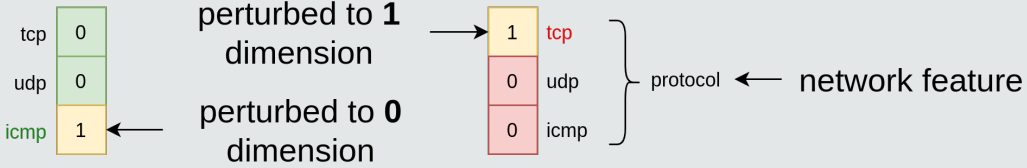
Experimental protocol

Protocol

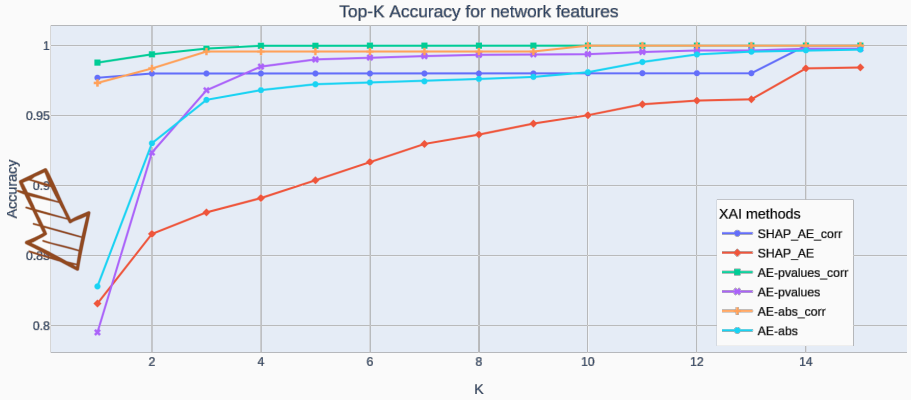
- Inject noise in a known network characteristic of vectors
- Assess ability of XAI methods to find the noisy network characteristic

Experiment with AE-abs (intuitive method), SHAP_AE (state of the art), AE-pvalues (our method)

Example of noise insertion in the protocol characteristic



Benchmark results



Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations. But sometimes several explanations are correct...

Several correct explanations

$$1 + 1 = 0$$

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2
- $+$ should be $-$

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2
- $+$ should be $-$
- 1 should be -1

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2
- $+$ should be $-$
- 1 should be -1
- $=$ should be $>$

Several correct explanations

$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2
- $+$ should be $-$
- 1 should be -1
- $=$ should be $>$
- "(mod 2)" is missing

Several correct explanations

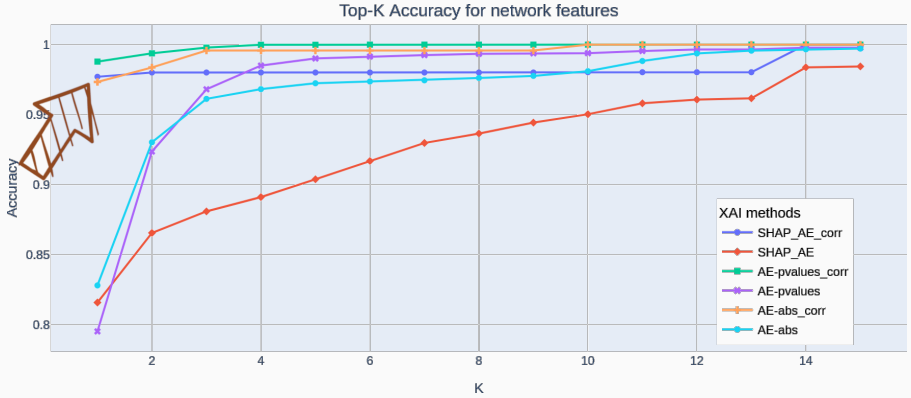
$$1 + 1 = 0$$

Where is the error?

We can all agree there is an error. But where do you think it is?

- 0 should be 2
- $+$ should be $-$
- 1 should be -1
- $=$ should be $>$
- "(mod 2)" is missing
- "is false" is missing

Benchmark results

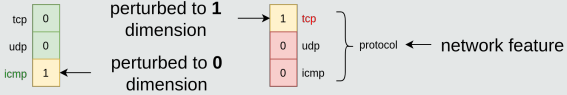


A more realistic evaluation

Evaluation modification: accepting correlated features as correct explanations

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

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

Conclusion

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

Comparison of the two paradigms

Signature based alert

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

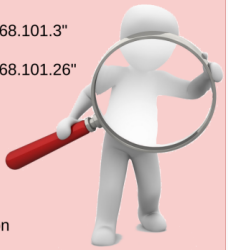
- CVE: **CVE_2021_4422**

Anomaly based alert

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



Abnormal features:
- connection_duration
- user_agent
- http_method
- http_trans_depth
- http_status_code
- ...

Experiment on CICIDS2017 dataset

Experiments

CICIDS2017 dataset

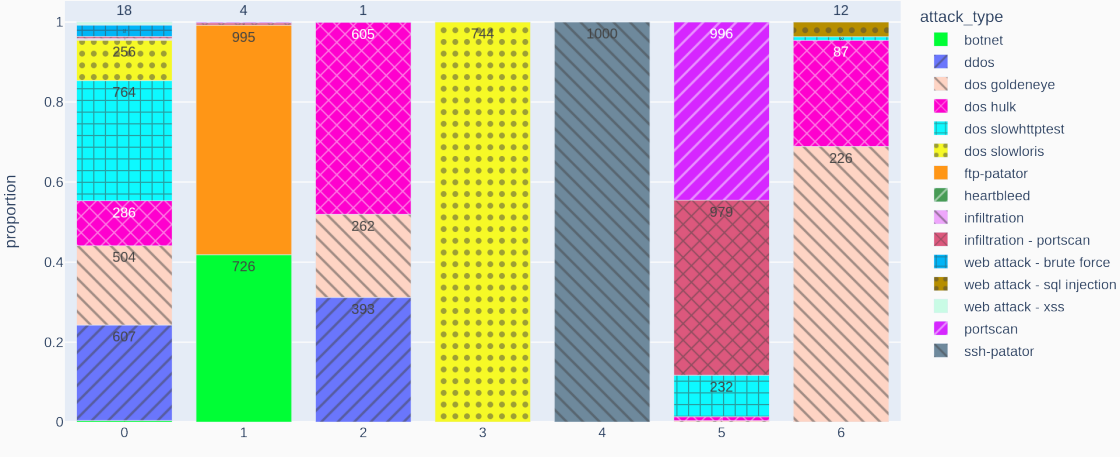
- A dataset of packets from a simulated network (no real users) with 12 machines
- Five days of recording: Monday without attack, Tuesday to Friday with attacks
- Attacks: port scan, DoS, web attacks, botnet, bruteforce, CVE exploit, etc.

Experimental protocol

- Learn a model on Monday (it does not know what kind of attacks exist!)
- Analyze the remaining days with the model to identify anomalies
- Generate explanations for these alerts
- Check whether the explanations match the attacks

Principle

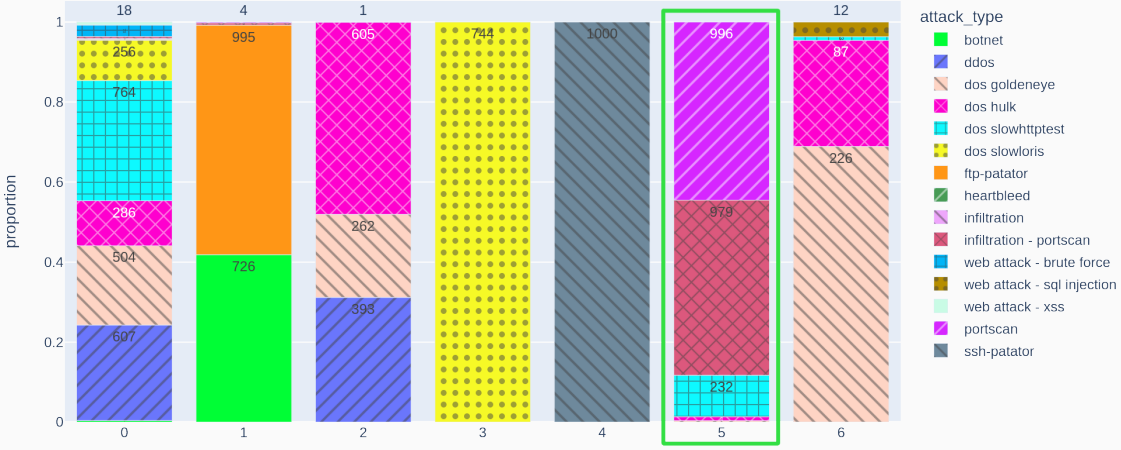
Clustering of the alerts based on the explanations



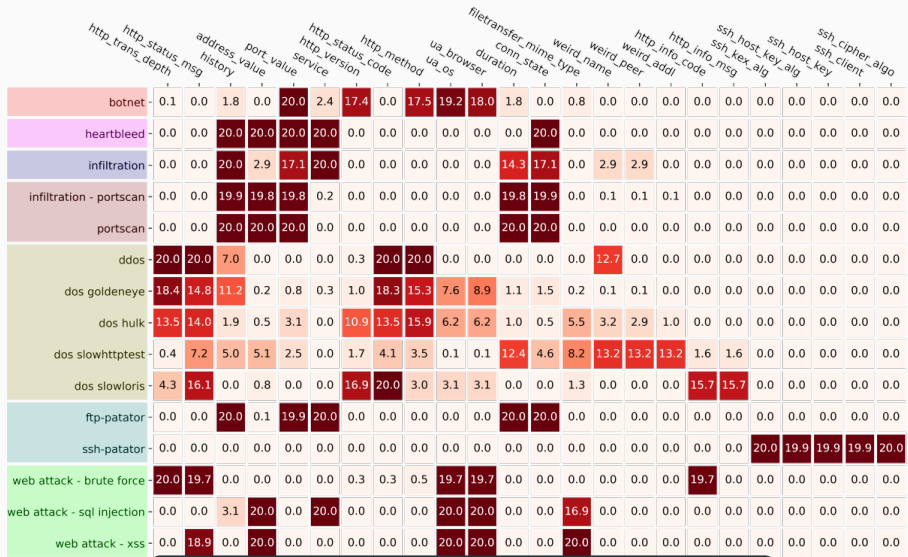
Applications - Clustering

Principle

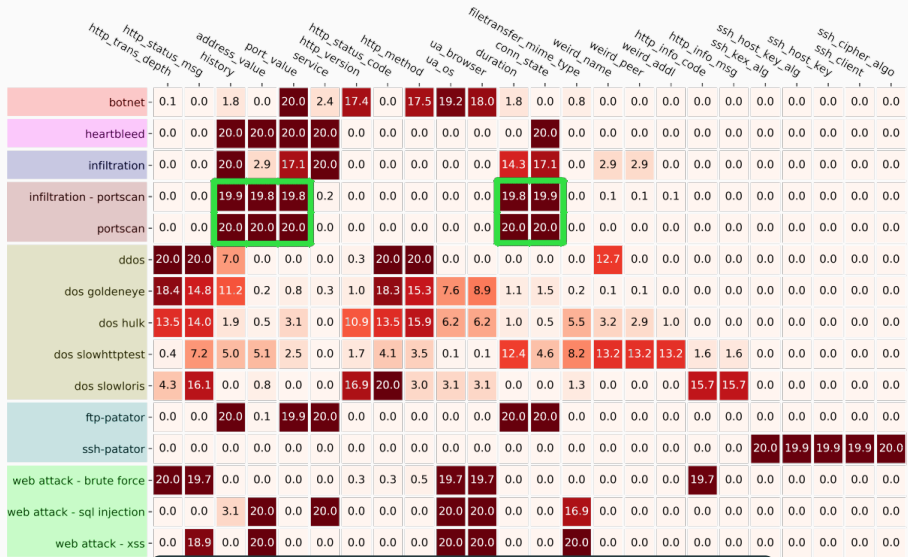
Clustering of the alerts based on the explanations



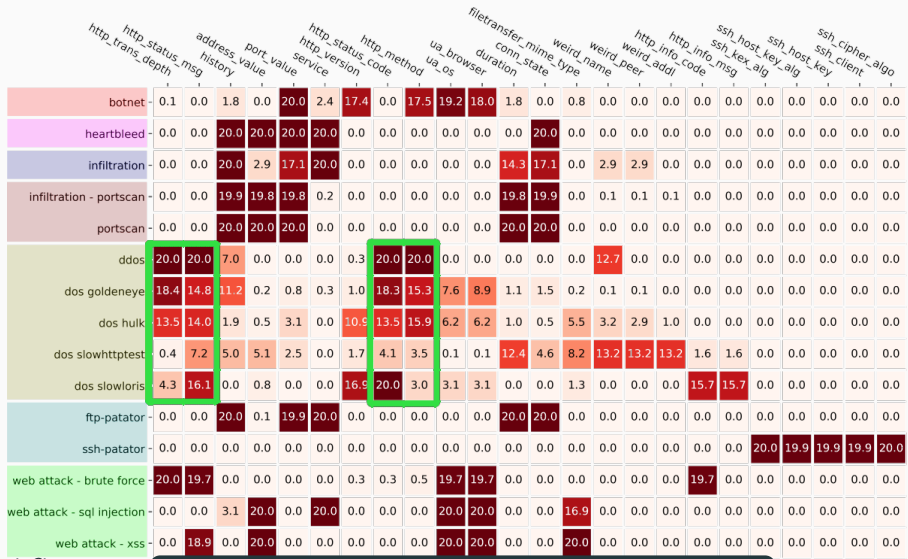
Applications - Top 5 explanations: feature contribution to attack types



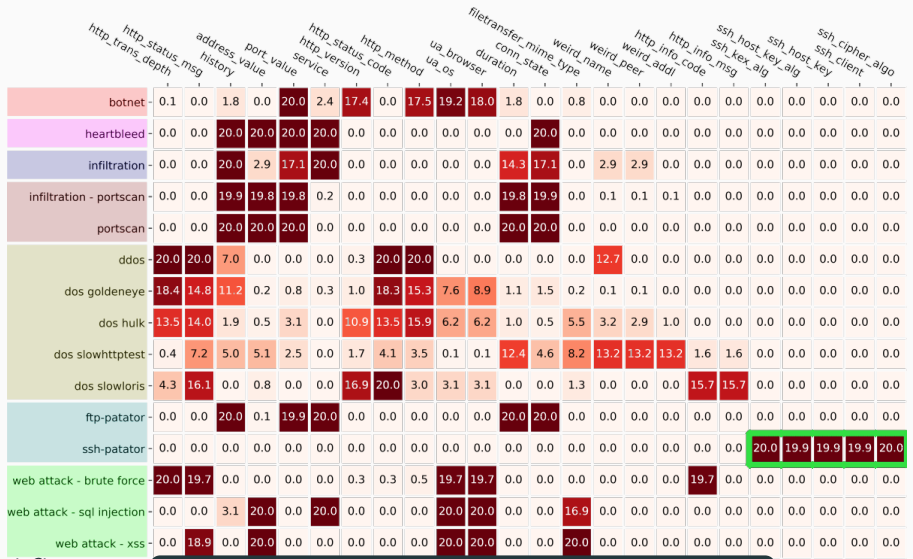
Applications - Top 5 explanations: feature contribution to attack types



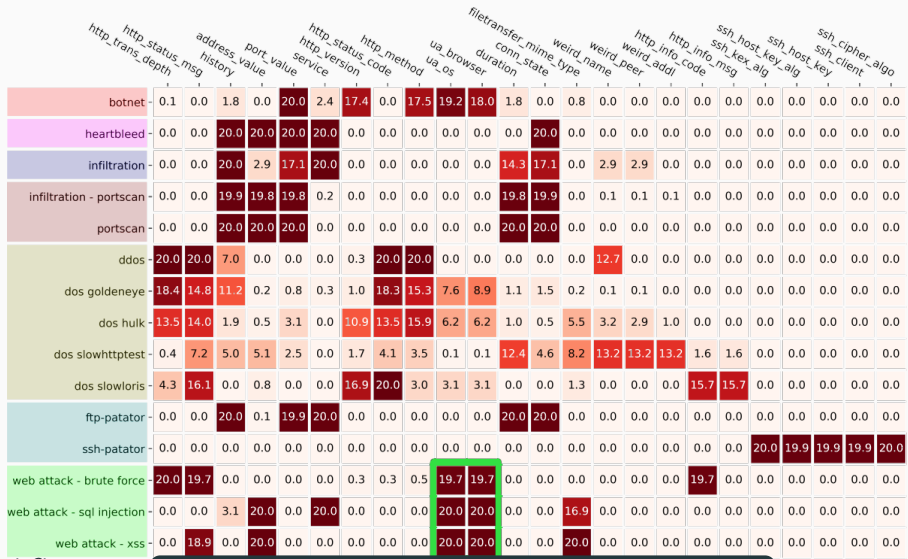
Applications - Top 5 explanations: feature contribution to attack types



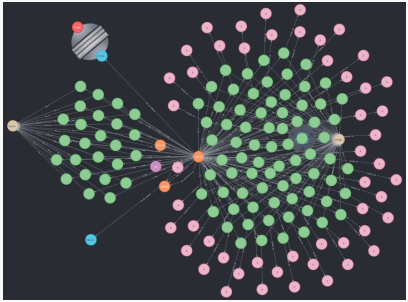
Applications - Top 5 explanations: feature contribution to attack types



Applications - Top 5 explanations: feature contribution to attack types



Applications - True Positive analysis - Web attack: Brute Force



single connection graph

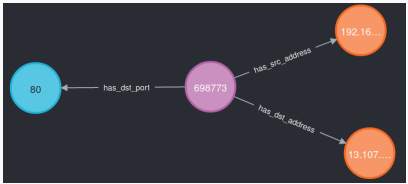
network_feature	value
http_method	POST
http_referrer	http://205.174.165.68/dv/login.php
http_request_body_len	130
http_status_code	302
http_status_msg	Found
http_trans_depth	84
user_agent_browser	Mozilla/5.0
user_agent_os	Linux x86_64

Top 5 explanations

user_agent_browser - user_agent_os - http_status_msg

http_status_code - http_trans_depth

Applications - Forensic analysis - A False Positive Analysis



single connection graph

network_feature	value
src_ip	192.168.10.15
dst_ip	13.107.4.50
src_port	49451
dst_port	80
proto	tcp
history	DadAttr
conn_state	RSTRH
orig_bytes	4226
resp_pkts	8884791

Top 5 explanations

port_value - history - conn_state - resp_pkts - orig_bytes

Conclusion

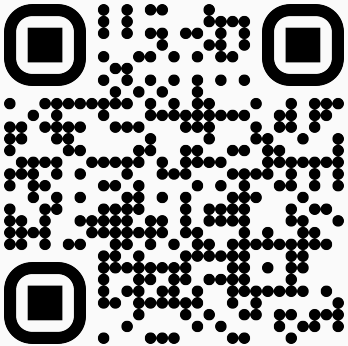
Conclusion

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