Leveraging explainability to increase the usability of intrusion detection systems

Sci-Rennes - October 14th, 2024 - Rennes

Maxime Lanvin
Frédéric Majorczyk

Pierre-François Gimenez
Ludovic Mé

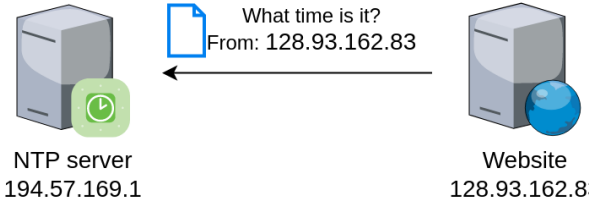
Yufei Han
Éric Totel

Introduction & motivation

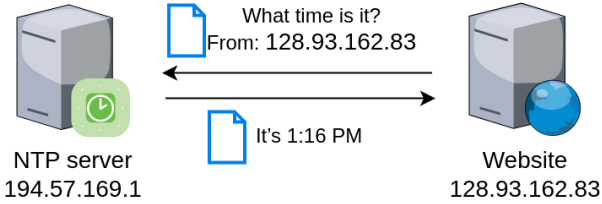
Simple denial of service attack

Website
128.93.162.83

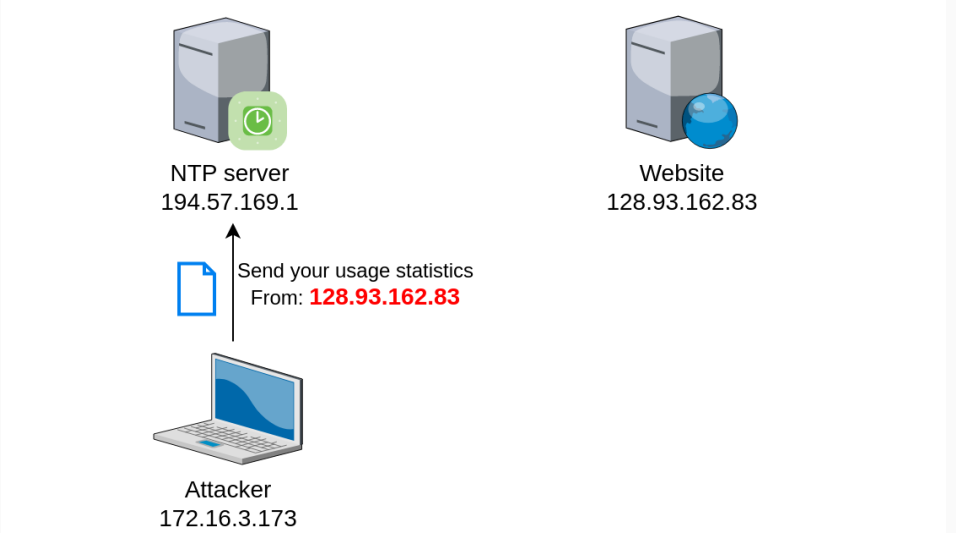
Simple denial of service attack



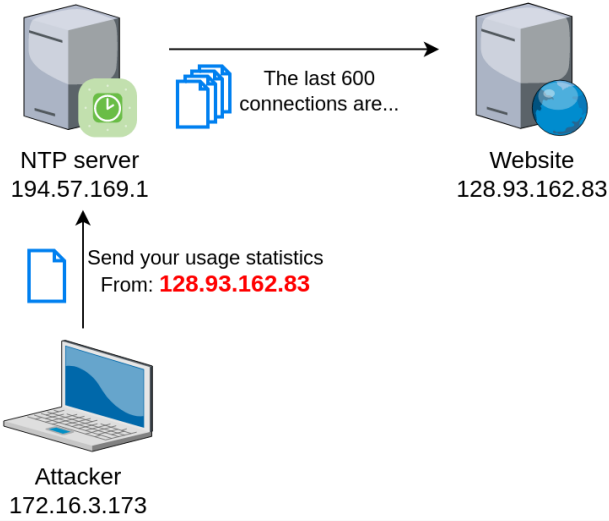
Simple denial of service attack



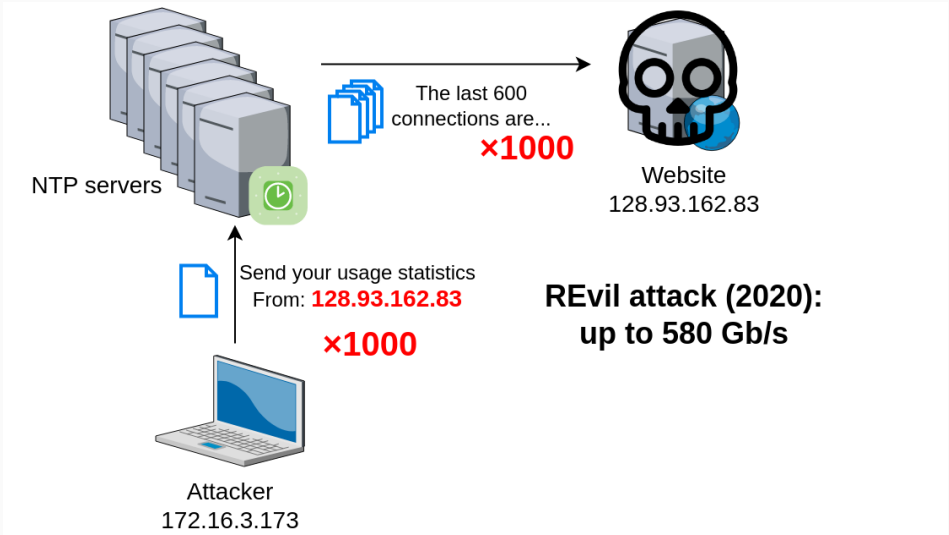
Simple denial of service attack



Simple denial of service attack



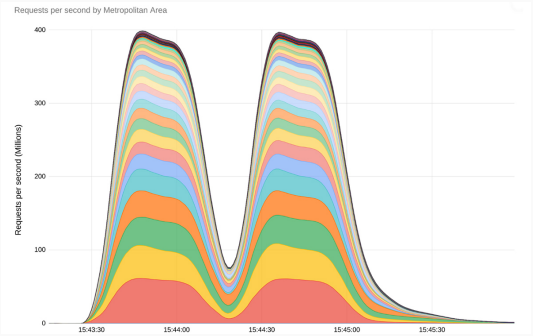
Simple denial of service attack



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!

Information system security

Information system security

- Prevent the attack, detect it, and react
- Detection with **IDS**: *Intrusion Detection System*

Detection relies on observation

- System : OS and applications logs
- Network : network communications

Constraints

- Partial and heterogeneous observations
- Adversarial context: the attacker hides!

```
2024-05-06T23:24:16.806598+02:00
stellar-sheep sshd[16039]:
Failed password for pfg from
192.168.1.36 port 48650 ssh2
```


```
"ts": 1591367999.305988,
"id.orig_h": "192.168.4.76",
"id.resp_h": "192.168.4.1",
"id.resp_p": 53, "proto":
"udp", "service":
"dns", "duration":
0.066851, "orig_bytes":
62, "resp_bytes":
141, "conn_state":
"SF", "orig_pkts":
2, "orig_ip_bytes":
118, "resp_pkts": 2,
"resp_ip_bytes": 197
```

Two categories of detectors

Signature-based detection

Date: 2024-04-25 10:24:52+02:00
Source IP: 194.57.169.1
Destination IP: 128.93.162.83

Signature : alert udp any any -> any 123 (content:"|00 02 2A|"; offset:1; depth:3; byte_test:1,!&,128,0; byte_test:1,&,4,0; byte_test:1,&,2,0; byte_test:1,&,1,0; threshold: type both, track by_dst,count 2, seconds 60);

Potential attack using NTP!

Signatures database

- \+ quick, clear
- − regular updates, only documented attacks

Anomaly detection

Date: 2024-04-25 10:24:52+02:00
Source IP: 194.57.169.1
Destination IP: 128.93.162.83

Anomaly score: 7,6

Normal behavior model

- \+ can detect undocumented attacks
- − false positives, **no description of the alert**

Two categories of detectors

Signature-based detection

Date: 2024-04-25 10:24:52+02:00
Source IP: 194.57.169.1
Destination IP: 128.93.162.83

Signature : alert udp any any -> any 123 (content:"|00 02 2A|"; offset:1; depth:3; byte_test:1,!&,128,0; byte_test:1,&,4,0; byte_test:1,&,2,0; byte_test:1,&,1,0; threshold: type both, track by_dst,count 2, seconds 60);

Potential attack using NTP!

Signatures database

- \+ quick, clear
- – regular updates, only documented attacks

Anomaly detection

Date: 2024-04-25 10:24:52+02:00
Source IP: 194.57.169.1
Destination IP: 128.93.162.83

Anomaly score: 7,6

Normal behavior model

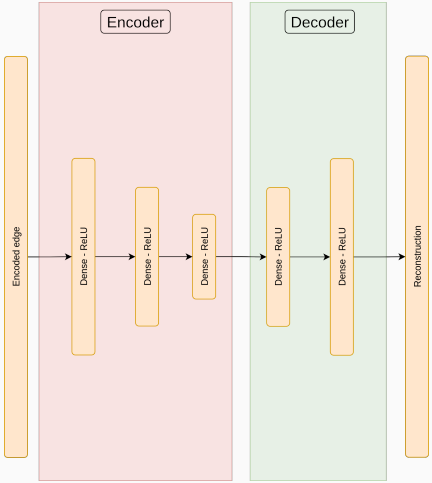
- \+ can detect undocumented attacks
- – false positives, **no description of the alert**

Summary

1. Introduction & motivation
2. AE-pvalues
3. Experiments with noise insertion
4. Conclusion

AE-pvalues

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

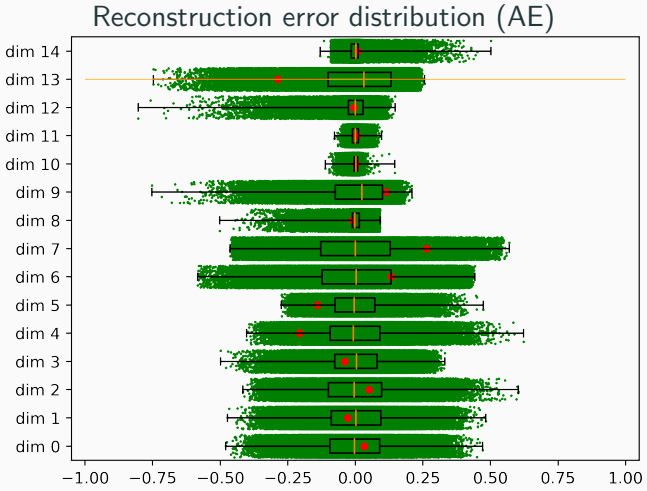
In our context, the explanations are an *ordered list of the network attributes* ranked from the most abnormal to the least abnormal.

AE-abs explanation method

Intuitive idea

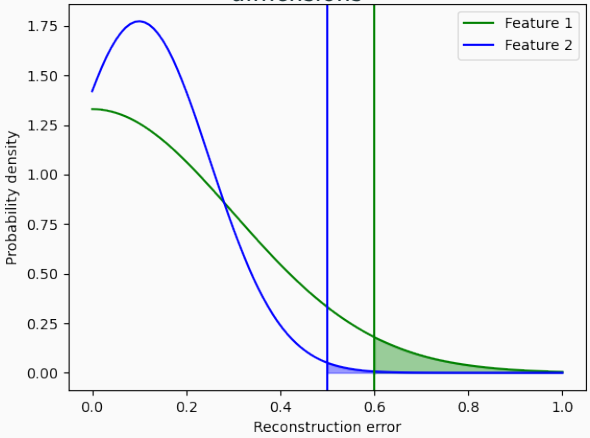
- When the reconstruction error is large, check the error dimension by dimension
- The higher the error of a dimension, the highest in the explanation list
- We call this method "AE-abs" and it has been proposed in the literature

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.

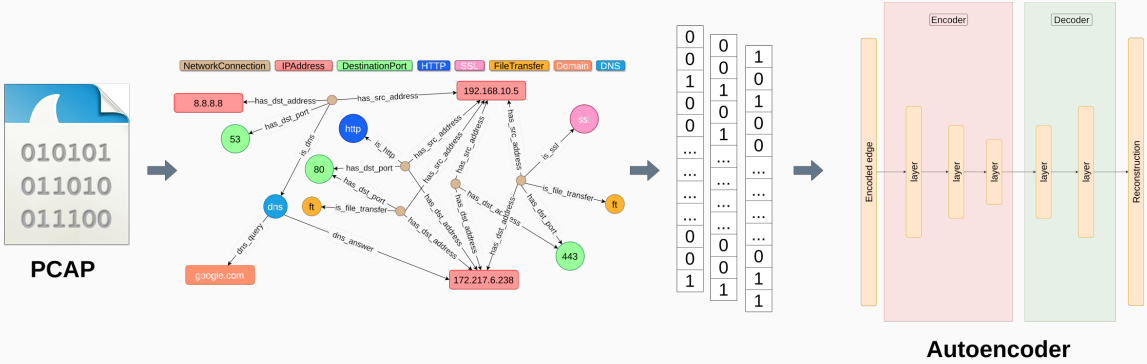
Our approach

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



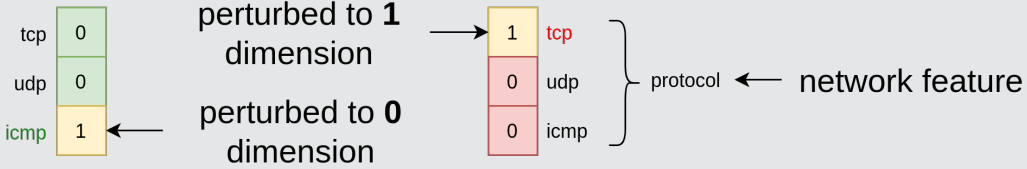
Experimental protocol

Protocol

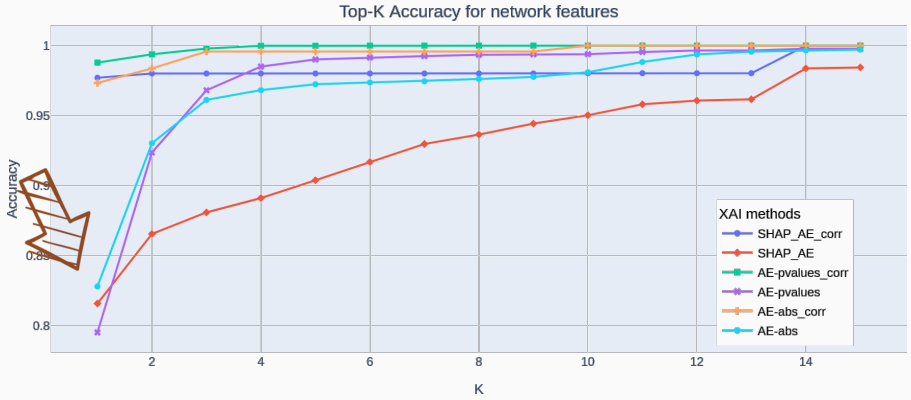
- Inject noise in a known network characteristic of vectors
- Assess ability of XAI methods to find the noisy network characteristic

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

$$1 + 1 = 0$$

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

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

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

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

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

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

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

Several correct explanations

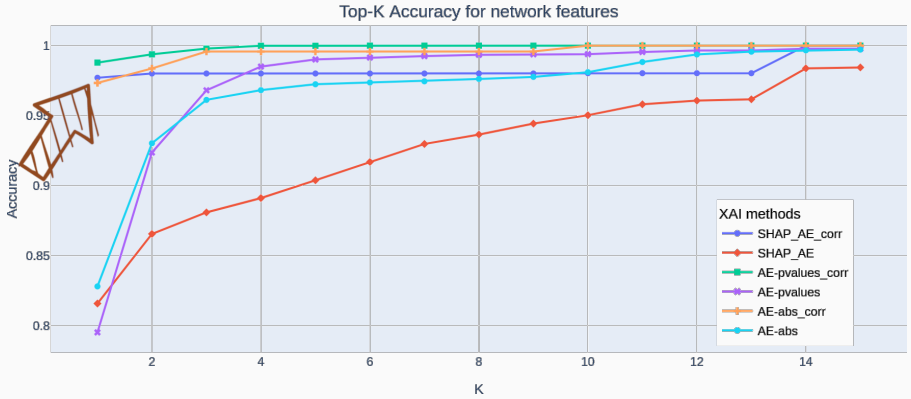
$$1 + 1 = 0$$

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

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

Conclusion

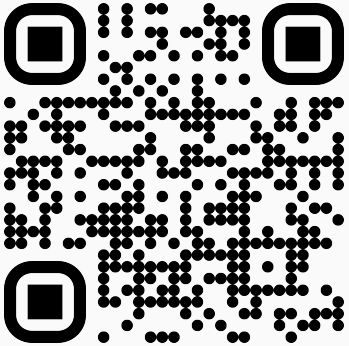
AE-pvalues is approximately $10,000$ faster than the SHAP_AE method.

Conclusion

Conclusion

Summary

- Alerts from anomaly detection are historically difficult to investigate
- Contribution: an explanation technique for alerts raised by AutoEncoder-based NIDS
- Very good accuracy results on noise injection and on actual attacks
- Our method is also much faster than the state of the art
- You want to know more? Maxime defends his PhD on December 17th



gitlab code for *AE-pvalues*

gitlab.inria.fr/mlanvin/ae-pvalues