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

Ínnía -







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 (Solar-Winds, 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



Pierre-François Gimenez

Introduction & motivation

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

- 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



AE-pvalues

Limitations



Key Idea

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

AE-pvalues

Principle





Our approach

This area is called the p-value:

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

Pierre-François Gimenez

Experiments with noise insertion

Sec2graph: An anomaly detection NIDS



Autoencoder

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)



Benchmark results

0.95 XAI methods AE-pvalues_corr AE-pvalues 0.8 ---- AE-abs 10 12 14 8 к

Top-K Accuracy for network features

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

$$1 + 1 = 0$$

$$1 + 1 = 0$$

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

• 0 should be 2

$$1 + 1 = 0$$

- 0 should be 2
- $\bullet~+$ should be -

$$1 + 1 = 0$$

- 0 should be 2
- \bullet + should be -
- 1 should be -1

$$1 + 1 = 0$$

- 0 should be 2
- \bullet + should be -
- 1 should be -1
- $\bullet \ = {\rm should} \ {\rm be} >$

$$1 + 1 = 0$$

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

$$1 + 1 = 0$$

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

Benchmark results



Top-K Accuracy for network features

A more realistic evaluation

Evaluation modification: accepting correlated features as correct explanations

Benchmark results

Vocabulary reminder



	Mean rank of the	Mean rank of the	Mean rank of the
explaining method	perturbed to 0 dimension	perturbed to 1 dimension	network feature \downarrow
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.

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





Experiment on CICIDS2017 dataset

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

Applications - Clustering

Principle

Clustering of the alerts based on the explanations



Pierre-François Gimenez

Experiment on CICIDS2017 dataset

Applications - Clustering

Principle

Clustering of the alerts based on the explanations



Pierre-François Gimenez

Experiment on CICIDS2017 dataset

	http://						ht					filetra,	DSF.				4.	,		AZZ			Sec		
	p trans d	tatus	ada m_ his	ress v	Port V	hi sei	to ver	atus c	p me	40 40	a brow	dura	conn s	nime	eird no	veird	weird	info	info	sh key	st key	host	SSh C	ipher	à/-
		~~~~	159	2	100		, (C6	"On	-0 ₀	α 	05	967 -	"On	ule .	POe	ne	0er	991	ode .	159	919	919	ney .	Cht	" ⁹ 0
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - sql injection ·	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pierre-Franç	ois Gimenez							Expe	rime	nt o	n Cl(		2017	7 dat	aset										2

	http://						ht					filetra,	DSF.				4.	,		SSA			Sec		
	p trans d	tatus	ada his	ress v	PORT	hı ser	to ver	atus c	p_me	40 40	abrow	dura	conn s	nime	eird no	Veird	weird	info C	info	sh tet	st key	host	SSh C	Pher	à/-
		~~~~~	159	2	100	- <i>'U</i> e	, (C6	non	-0 ₀	10q	. o _S	Ser.	100	916	PDe	ne	er-	991	-0 ₀	159	919	919	ney.	Cht	'' ⁹ 0
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration ·	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pierre-Franç	ois Gimenez							Expe	rime	nt o	n Cl(2017	7 dat	aset										2

	http: http:						htt					filetra,	DSF.				4.	,		AZZ			Sec		
	p trans a	status	ada his	ess v	Port	h sei	tp ver	atus o	tp_me	4	a brow	dura	conn s	nime	eird n	Weird	weird	info	info	sh key	st key	host	SSh C	ipher	a.
		- 130-	159	2	-14e	"Ue	-'Ce	100	ode .	''0q	. os	367	10n	ale	Pe	ne	Cer	"OqI	ode -	, 15g	gla	alg -	ner .	CUL	*'90 '
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pierre-Franç	ois Gimenez							Expe	erime	ent o	n Cl(2017	7 dat	aset										2

	http_trap_	tar	ada	Tec	00	h	http_st	ar ha	0	4	'a ,	filetra,	nsfer I	nin 4	ein I	Wei.	http	http	o in s	SSh ho	St SS	2	SSh C	1DA	
	20	Pth	msg	tory	alue	ser Nue	Vice	sion C	ode	hog 40	os os	ser	tion	tate	Ype -	ame	Peer	adai	ode	msg	alg	alg	key -	lent	^a /90
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - sql injection ·	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pierre-Franç	ois Gimenez							Exne	rime	nt o	n Cl		2017	7 dat	aset										2

	http://						ht					filetra,	DSF.				4.	,		SSA			Sec		
	p trans d	tatus	ada his	ess v	Port V	hı ser	to ver	atus c	p me	1 40 1 40	a brow	dura	conn s	nime	eird no	veird	weird	info	info	sh tet	st key	host	SSh C	Pher	à/-
		~~~~	159	2	100	- <i>'U</i> e	, (C6	"On	-0 ₀	100	05	967 -	"On	ule .	POe	ne	0er	991	ode .	159	919	919	ney .	Cht	'' ⁹ 0
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - sql injection ·	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pierre-Franç	ois Gimenez							Expe	rime	ent o	n Cl(		2017	7 dat	aset										2

### Applications - True Positive analysis - Web attack: Brute Force



single connection graph

#### **Top 5 explanations**

network_feature

http method

http referrer

http request body len

http status code

http_status_msg http_trans_depth

user_agent_browser

user agent os

value

POST

130

302 Found

84

Mozilla/5.0 Linux x86 64

http://205.174.165.68/dv/login.php

## 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

Pierre-François Gimenez

Experiment on CICIDS2017 dataset

## Conclusion

#### Summary

- Explanation technique for alerts raised by AutoEncoderbased 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