Master's Thesis

Analyzing Dataset Quality and Machine Learning Models Efficiency for Intrusion Detection Systems

Presented by

Mohammed **SGHIOURI**

Internship advisors

Pr. Gérard **CHALHOUB**

Mr. Maxime **PUYS**

Journées Scientifiques de l'Axe SIC

MAIN PROBLEM

Main Problem

How to build a suitable Anomaly-based Intrusion Detection System for various use cases?

HOW TO SOLVE THIS?

How to choose the well-suited dataset?

- KDD CUP99
- NSL-KDD
- Kyoto 2006+
- UNSW-NB15
- CIC-IDS17
- CSE-CIC-IDS18
- ToN-IoT18
- BoT-IoT

How to solve?

Data Preparation for Training

- By analysing the **PCAP** files, and the generated **csv** files of the benshmark datasets.
- Understand differents behaviors genrated/simulted within a network.
- Understand the use and the meaning of each feature generated from Pcap file.

CSV

Destination Port	Flow Duration	Total Fwd Packets	.....	Label
88	640	7		BENIGN
88	900	9		BENIGN
88	1205	7		BENIGN
88	511	7		BENIGN
88	773	9		BENIGN
88	986	9		BENIGN
88	935	9		FTP-Patator
389	572849	15		FTP-Patator
49193	1	2		BENIGN
88	1075	9		BENIGN
88	2687	9		BENIGN
135	14263712	21		BENIGN
49671	14257993	11		BENIGN
135	68	1		BENIGN
49671	58	1		BENIGN
80	26641766	4		BENIGN
49666	34256029	39		BENIGN

PCAP

No.	Time	Source	Destination	Protocol	Length	Info
130	82.263004	192.168.10.50	192.168.10.3	LDAP	469	SAS
131	82.263007	192.168.10.50	192.168.10.3	TCP	469	[TC
132	82.263314	192.168.10.3	192.168.10.50	LDAP	382	SAS
133	82.263318	192.168.10.3	192.168.10.50	TCP	382	[TC
134	82.263421	192.168.10.50	192.168.10.3	TCP	66	338
135	82.263424	192.168.10.50	192.168.10.3	TCP	66	[TC
136	82.264009	192.168.10.50	192.168.10.3	LDAP	469	SAS
137	82.264013	192.168.10.50	192.168.10.3	TCP	469	[TC

THE GOAL OF THIS STUDY

Conclusion

Being able to build an anomaly-based IDS specific to each use case.

- A better feature selection approach
- Customization of the dataset's behavior
- Build a dataset with a specific network behavior from the available datasets.

THANK YOU FOR LISTENING