CasinoLimit: An Offensive Dataset Labeled with MITRE ATT&CK Techniques

Sébastien Kilian
V. Viet Triem Tong, J. Lalande , F. Majorczyk, A. Sanchez
N. Talon, P. Besson, H. Orsini, P. Lledo, P. Gimenez

sebastien.kilian@inria.fr
CentraleSupélec, Inria
FRANCE

Introduction

Datasets are essential in cybersecurity: Detection, CTI, forensic analysis ...

Data sources

- **Real-world data** - Logs from organizations under attack
- **Controlled environment** - attacks in a controlled environment

	Real-world (real attackers)	**Controlled env.** (attack agents)	**Botnet** (infected hosts)	**Malware** (malicious samples)
Accurate	✓	X	✓	X
Scalable	X	✓	X	✓
Available	X	✓	✓	✓

How to get attack data from human sources?

Problem statement

Cybersecurity datasets often lack **diversity** and does not reflect the **complexity** of real-world attacker behaviors.

Objectives:

1. **Collect** data with a large variety of attack techniques
2. **Label** a large amount of attack activity with fine-grained techniques
3. **Analyze** attacker behaviors efficiently

Introduction

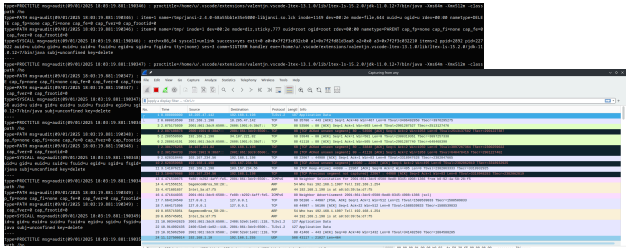
BreizhCTF

Largest in-person cybersecurity competition in France

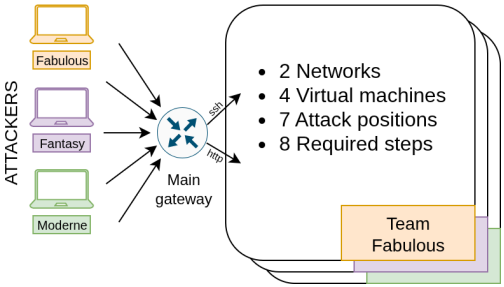
600 players | 120 teams | 10h

Collect:

- auditd events
- application logs
- network flows
- pcap files

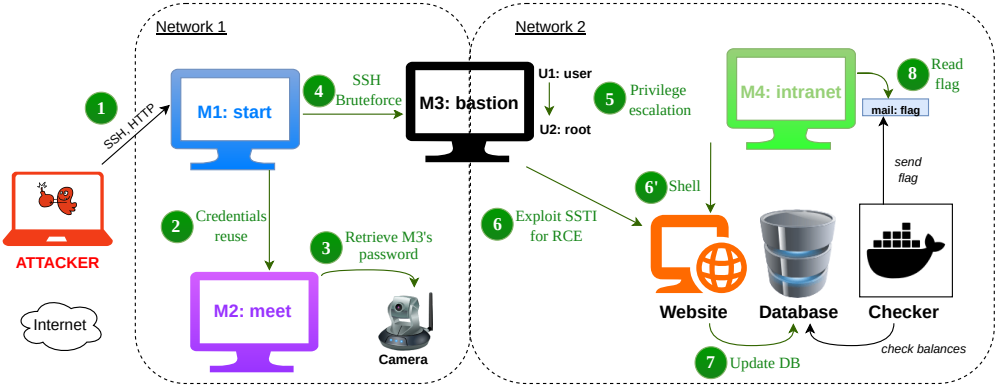


Infrastructure



The CasinoLimit Scenario

Goal: Progress in a casino's network to remove an entry in a database.



Dataset: An example of a privilege escalation technique

Team Fantasy

```
02:35:24   tbenedict@start   scp poc tocean@10.35.128.20
02:35:39   tocean@bastion    ./poc
02:35:39   root@bastion      /bin/bash
```

Dataset: An example of a privilege escalation technique

Team Fantasy

02:35:24	tbenedict@start	scp poc tocean@10.35.128.20
02:35:39	tocean@bastion	./poc
02:35:39	root@bastion	/bin/bash

Team Moderne

03:31:34	tbenedict@start	scp CVE.zip tbenedict@meetingcam:/home/tbenedict/CVE.zip
03:32:26	tbenedict@meet	scp CVE.zip tocean@bastion:/home/tocean/CVE.zip
03:40:21	tocean@bastion	untar
03:40:33	tocean@bastion	tar --help
03:40:53	tocean@bastion	tar tf CVE.tar
03:40:56	tocean@bastion	tar -xvf CVE.tar
03:41:54	tocean@bastion	./exp
03:41:54	root@bastion	/bin/bash

Dataset content

Log volume

- 25 Million auditd events (3.4 Million processes)
- 1.1 Billion packets (99 Million network flows)
- **Total:** 540 GB of data

Dataset content

Log volume

- 25 Million auditd events (3.4 Million processes)
- 1.1 Billion packets (99 Million network flows)
- **Total:** 540 GB of data

Specificities

- Multiple attacker profiles
- Erratic behaviours

Dataset content

Log volume

- 25 Million auditd events (3.4 Million processes)
- 1.1 Billion packets (99 Million network flows)
- **Total:** 540 GB of data

Specificities

- Multiple attacker profiles
- Erratic behaviours

→ **Labels are needed to understand and analyze the attack activity.**

Labeling system logs

MITRE ATT&CK: Tactics and techniques

- Most of the labels can be put **automatically**
- Some need human intervention and validation

Example:

00:48:33.613 **[tocean@bastion]$** ping 8.8.8.8

Labeling system logs

MITRE ATT&CK: Tactics and techniques

- Most of the labels can be put **automatically**
- Some need human intervention and validation

Example:

00:48:33.613 **[tocean@bastion]$** ping 8.8.8.8

- Naive label: T1018: Remote System Discovery

Labeling system logs

MITRE ATT&CK: Tactics and techniques

- Most of the labels can be put **automatically**
- Some need human intervention and validation

Example:

00:48:33.613 **[tocean@bastion]$** ping 8.8.8.8

- Naive label: T1018: Remote System Discovery
- Correct label: T1016: System Network Configuration Discovery

Network logs labeling

Methodology

Find the **network consequences** of each process and propagate the labels.

Labeling:

- Build requests to match flows
- Most can also be done automatically

Example:

00:48:33.613 **[tocean@bastion]$** ping 8.8.8.8

Network logs labeling

Methodology

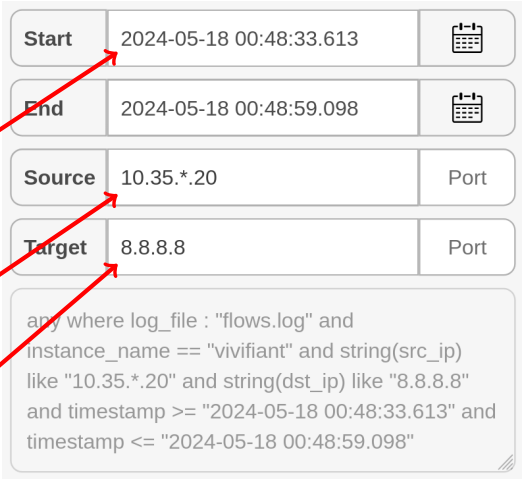
Find the **network consequences** of each process and propagate the labels.

Labeling:

- Build requests to match flows
- Most can also be done automatically

Example:

00:48:33.613 **[tocean@bastion]$** ping 8.8.8.8



Matching example

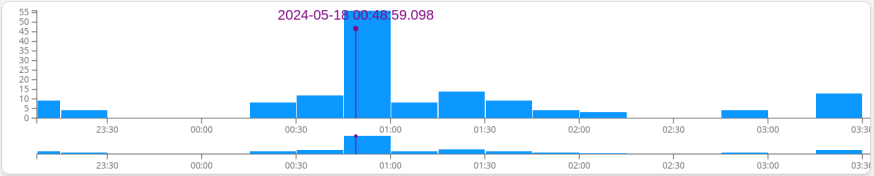
System logs:

Time	user@host	Command
00:48:33	tbenedict@meet	ssh
00:48:34	tocean@bastion	cat /etc/passwd
00:48:34	tocean@bastion	ls -la /home
00:48:35	tocean@bastion	chmod +x exploit.sh
00:48:35	tocean@bastion	ping 8.8.8.8
00:48:35	tocean@bastion	./exploit.sh
00:48:36	tocean@bastion	rm -rf /tmp/file.txt
00:48:37	tocean@bastion	sudo su

Network flows:

Time	Source	Destination
00:48:35	10.35.229.20:3285	203.0.113.1:25
00:48:35	10.35.229.20:1029	198.51.100.2:110
00:48:35	10.35.229.20:4390	192.0.2.3:443
00:48:35	10.35.229.20:21398	10.35.229.50:3306
00:48:35	10.35.229.20:0	8.8.8.8:8
00:48:35	10.35.229.20:67926	10.35.229.10:22
00:48:35	10.35.229.20:0	8.8.8.8:8
00:48:36	10.35.229.10:22	10.35.229.20:67926
00:48:35	10.35.229.20:12345	192.168.1.1:80
00:48:35	10.35.229.20:54321	172.16.0.5:443
00:48:36	10.35.229.20:31093	8.8.4.4:53
00:48:36	10.35.229.20:1341	10.35.229.30:8080
00:48:36	10.35.229.20:34154	10.35.229.40:21

vivifiant (150) | Search | Expand All | Validated



	T1033: System Owner/User Discovery	00:39:31.080	[tbenedict@start]$ id -u
		00:45:50.659	[tocean@bastion]$ /bin/sh /usr/bin/lesspipe
	T1049: System Network Connections Discovery	00:46:24.861	[tocean@bastion]$ ss -ntlp
	More (2) T1114: Email Collection	00:47:01.031	[tocean@bastion]$ sh -c /usr/lib/mutt/source-muttrc.d
	T1114: Email Collection	00:47:05.392	[tocean@bastion]$ mutt
☑	T1016: System Network Configuration Discovery Network	00:48:33.613	[tocean@bastion]$ ping 8.8.8.8
	T1105: Ingress Tool Transfer	00:48:59.098	[tocean@bastion]$ git clone https://github.com/sxlmnwb/CVE-2023-03
	More (8) T1068: Exploitation for Privilege Escalation	00:49:05.587	[tocean@bastion]$ /usr/bin/ld -plugin /usr/lib/gcc/x86_64-linux-gnu/11/
	More (1) T1068: Exploitation for Privilege Escalation	00:49:23.872	[tocean@bastion]$ screen
	More (3) T1068: Exploitation for Privilege Escalation	00:49:33.088	[tocean@bastion]$ screen
	More (10) T1068: Exploitation for Privilege Escalation	00:49:52.990	[tocean@bastion]$ sh -c rm -rf './ovlcap/upper/*'
	T1083: File and Directory Discovery	00:49:56.294	[root@bastion]$ ls --color=auto
	More (8) T1114: Email Collection	00:50:03.442	[root@bastion]$ mutt
	T1114: Email Collection	00:50:06.398	[root@bastion]$ mutt
		00:50:24.787	[admin@bastion]$ /bin/sh /usr/bin/lesspipe

Info

Event UID: 44272-bastion
Timestamp: 2024-05-18 00:48:33.613
Position: tocean@bastion
Proctitle: ping 8.8.8.8

Network

Advanced mode

Start	2024-05-18 00:48:33.613	
End	2024-05-18 00:48:59.098	
Source	10.35.*.20	Port
Target	8.8.8.8	Port

any where log_file : "flows.log" and instance_name == "vivifiant" and string(src_ip) like "10.35.*.20" and string(dst_ip) like "8.8.8.8" and timestamp >= "2024-05-18 00:48:33.613" and timestamp <= "2024-05-18 00:48:59.098"

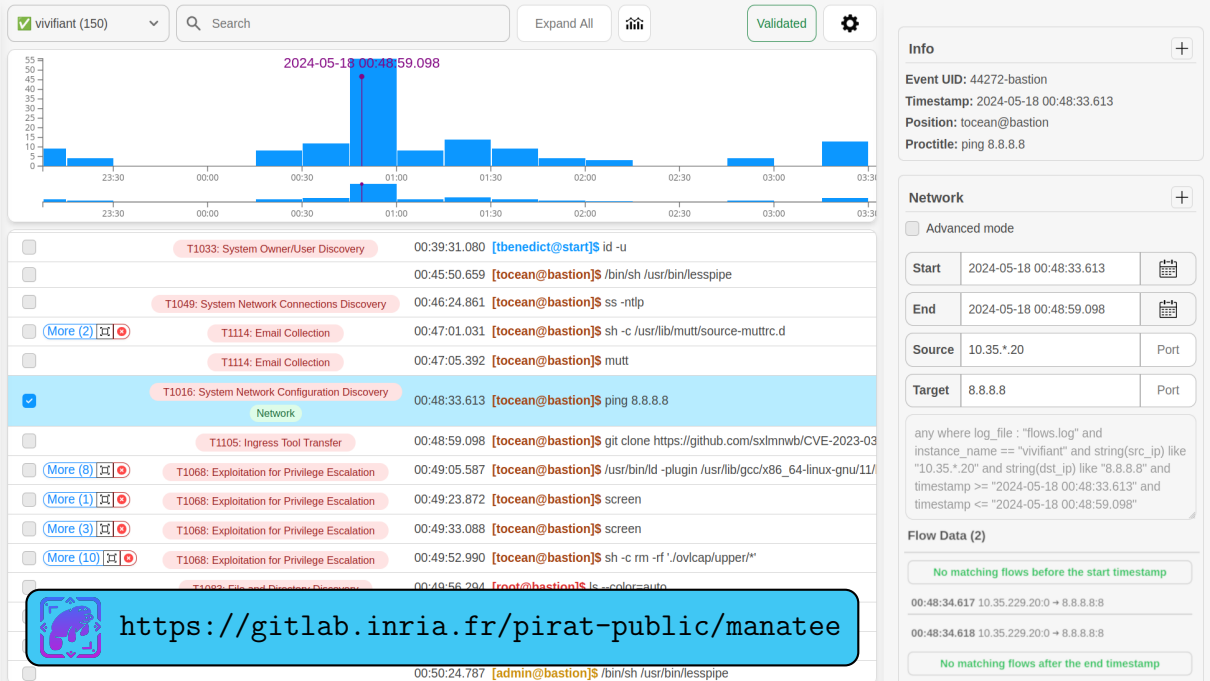
Flow Data (2)

No matching flows before the start timestamp

00:48:34.617 10.35.229.20:0 → 8.8.8.8:8

00:48:34.618 10.35.229.20:0 → 8.8.8.8:8

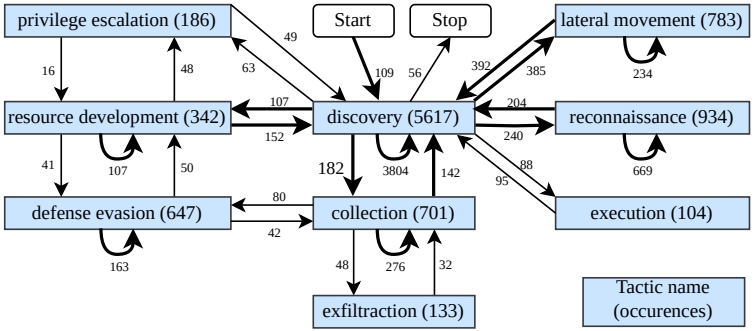
No matching flows after the end timestamp


vivifiant (150)
Search
Expand All
Validated
2024-05-18 00:48:59.098
Info
Event UID: 44272-bastion
Timestamp: 2024-05-18 00:48:33.613
Position: tocean@bastion
Proctile: ping 8.8.8.8
Network
Advanced mode
Start
2024-05-18 00:48:33.613
End
2024-05-18 00:48:59.098
Source
10.35.*.20
Port
Target
8.8.8.8
Port
any where log_file : "flows.log" and instance_name == "vivifiant" and string(src_ip) like "10.35.*.20" and string(dst_ip) like "8.8.8.8" and timestamp >= "2024-05-18 00:48:33.613" and timestamp <= "2024-05-18 00:48:59.098"
Flow Data (2)
No matching flows before the start timestamp
00:48:34.617 10.35.229.20:0 → 8.8.8.8:8
00:48:34.618 10.35.229.20:0 → 8.8.8.8:8
No matching flows after the end timestamp
T1033: System Owner/User Discovery
00:39:31.080 [tbenedict@start]$ id -u
00:45:50.659 [tocean@bastion]$ /bin/sh /usr/bin/lesspipe
T1049: System Network Connections Discovery
00:46:24.861 [tocean@bastion]$ ss -ntlp
More (2)
T1114: Email Collection
00:47:01.031 [tocean@bastion]$ sh -c /usr/lib/mutt/source-muttrc.d
T1114: Email Collection
00:47:05.392 [tocean@bastion]$ mutt
T1016: System Network Configuration Discovery
Network
00:48:33.613 [tocean@bastion]$ ping 8.8.8.8
T1105: Ingress Tool Transfer
00:48:59.098 [tocean@bastion]$ git clone https://github.com/sxlmnwb/CVE-2023-03
More (8)
T1068: Exploitation for Privilege Escalation
00:49:05.587 [tocean@bastion]$ /usr/bin/ld -plugin /usr/lib/gcc/x86_64-linux-gnu/11/
More (1)
T1068: Exploitation for Privilege Escalation
00:49:23.872 [tocean@bastion]$ screen
More (3)
T1068: Exploitation for Privilege Escalation
00:49:33.088 [tocean@bastion]$ screen
More (10)
T1068: Exploitation for Privilege Escalation
00:49:52.990 [tocean@bastion]$ sh -c rm -rf './ovlcap/upper/*'
00:50:24.787 [admin@bastion]$ /bin/sh /usr/bin/lesspipe
https://gitlab.inria.fr/pirat-public/manatee

Applications: Empirical kill chain

Tactics transitions

A typical behaviour stand out from the sequences of tactics used by the attacker.



Applications: Attribution

Personal identification information removed
but some habits allow to differentiate players.

Identification indicators

- Text editor (vi, vim, nano . . .)
- Options of a few commands (nc, ss, nmap . . .)
- Aliases of commands (ip addr show)

Example:

Player A	Player B
`ss -tuln`	`ss -nalup`

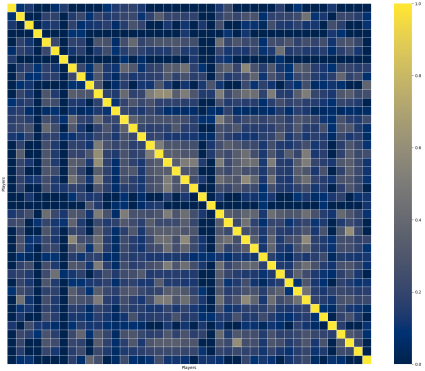


Figure: Similarity matrix between players (Jaccard index)

Key takeaways: Methodology

Labeling Methodology

- Suitable for a large amount of **attack activity** and high **diversity of techniques**
- Network traffic labels are **deduced** from system logs

MANATEE

- Dedicated tool for the labeling process
- Available at: gitlab.inria.fr/pirat-public/manatee

Key takeaways: Dataset

CasinoLimit - Dataset of attacks

- You can **replay** the challenge easily
- 540 GB of attack activity (system **and** network)
- 114 attacker profiles

Results

- **14** / 14 tactics used
- **67** / 282 techniques
- **9243** labels placed

Paper and dataset →

https://casinolimit.inria.fr

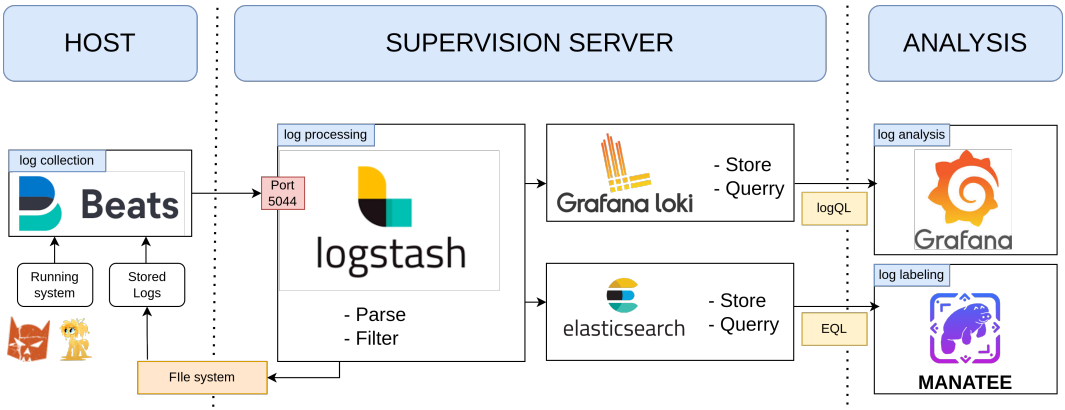
What's next?

Future work

- Enhanced modelling of attacker behaviors
- Compare with offensive models

Meet us in Rennes for the BreizhCTF 2026!

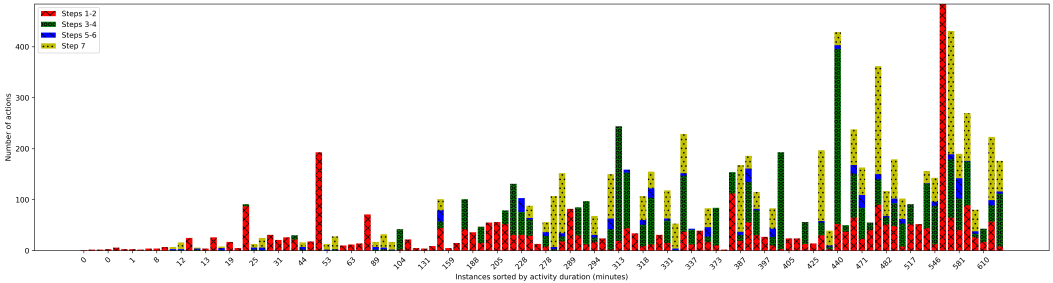
Data collection



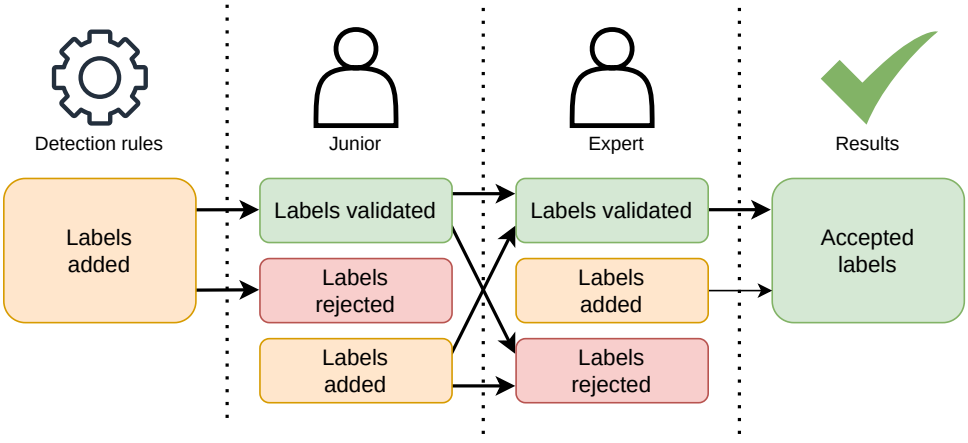
Application: Player efficiency

Player efficiency

The difference in efficiency between players is significant.



Labelling process



General overview

Results

- All **14 tactics** seen in the dataset
- **67 / 282** different techniques
- **9243** labels placed

→ 540 GB of data

	Input data	**Preprocessed**	**Labeled**
processes	3.4M	405K	399K (98%)
events	25M	5M	4.9M (98%)
flows	99M	52M	6.1M (12%)