# MalPaCa: What sequential features and combination thereof describe a malicious network behavior best?

Best = not privacy intrusive, easy to extract, most general

## 1 Background

- MalPaCa is a clustering tool. Its main objective is to classify network behaviors
- It employs sequential features instead of statistical ones
- Prior to the results of this research there were 4 features considered as its inputs (Packet size, Time interval, Dest & Source ports)
- The applied clustering algorithm is HDBSCAN

## 2 Methodology

### Data Extraction and Code Reformating

- Data was trimmed and subsequently uniformly distributed
- Labels are assigneby the Zeek network analyzer
- Refactor code

### Feature Extraction

- Understanding the malware within the IoT-23 dataset and determining the behaviors of each malware
- Evaluation of best features describing given malware

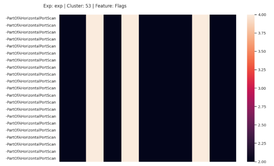
### Feature combination

- Given the set of candidate features, the most generalizable subset of those features was created (Grid search)

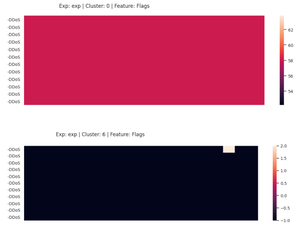
## 3 Result

### TCP Flags

SYN-scan cluster

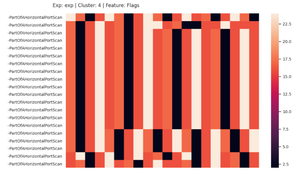


DDoS flood clusters



Okiru-attack cluster

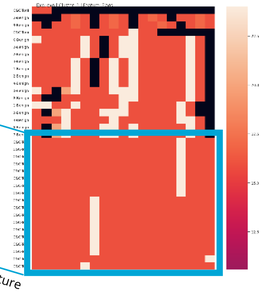
XMas scan cluster



Pure C&C-Torii cluster

C&C Torii + HTTPS cluster

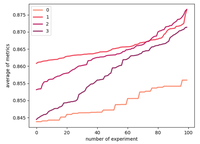
(clustering combination without TCP flags)



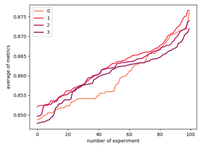


### Grid Search

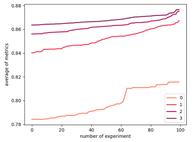
Each graph represents the top 100 combination scores of averaged metrics, grouped by weight (each line represents a weight) and sorted in ascending order:



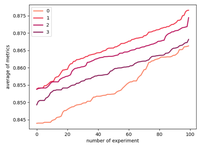
Time interval



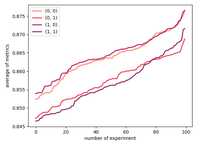
Packet size



Source & Destination Ports



TCP Flags



Variants of time feature (Normalization, 2TD)

Combination based on the average weight of the first 50 scores of the grid-search

The weights are the scores of the metric, rounded

| | Cluster Purity | Malware Purity | Clustered Data Points | Average of Metrics |
|---|---|---|---|---|
| Flag weight | 2 | 1 | 1 | 1 |
| Source Port weight | 2 | 2 | 3 | 3 |
| Destination Port Weight | 2 | 2 | 3 | 3 |
| Time Interval weight | 1 | 0 | 2 | 2 |
| Packet Size Weight | 2 | 1 | 2 | 1 |
| Second Time Difference | 0 | 0 | 0 | 0 |
| Time feature normalization | 1 | 1 | 0 | 1 |

## 4 Conclusion

- TCP Flag s feature was extracted and showed significant results (mainly in purity metrics)
- Normalization of the time feature influenced positively the purity metric
- A single best combination for all metrics was not chosen. However, it is possible to select a combination for single metrics.

**Author: Jonathan Garack** CSE3000

**Supervisors: Sicco Verwer, Azqa Nadeem**

TU Delft