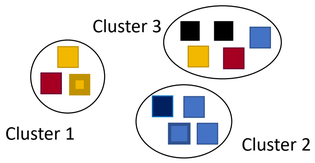
The influence of sequence length on the clustering performance of MalPaCA

Author: **Johannes Hagspiel**
Supervisors: **Azqa Nadeem, Sicco Verwer**
CSE3000 Research Project
26 June 2021

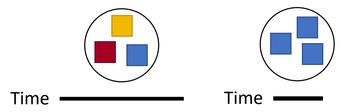
1 Background



MalPaCA is a novel, unsupervised clustering approach for classifying software behaviour based on packets send.

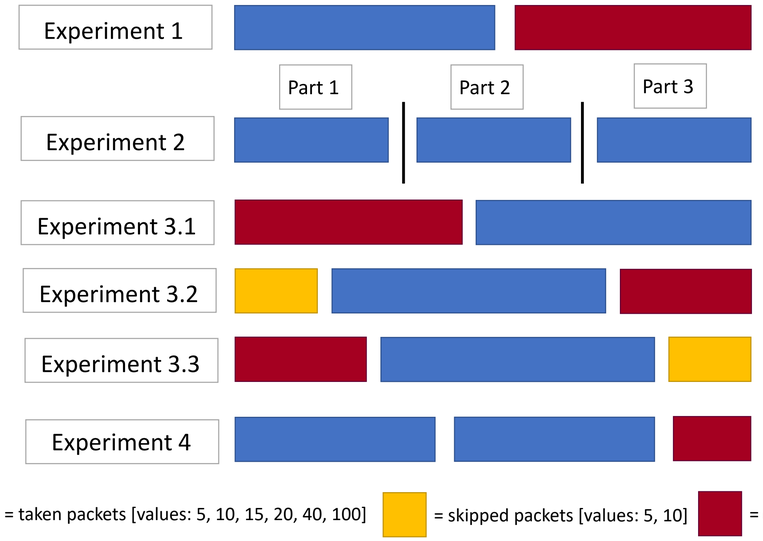
MalPaCA defines behavior as a unidirectional flow of packets based on IP source and destination

The current default is analyzing per connection the first 20 packets.



Which and how many packets MalPaCa analyzes influences its clustering effectiveness and efficiency.

2 Experiments



= taken packets [values: 5, 10, 15, 20, 40, 100] = skipped packets [values: 5, 10] = discarded packets

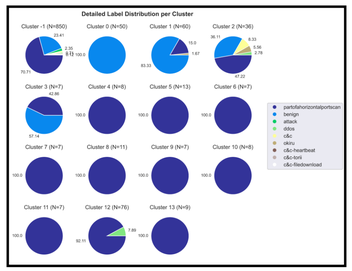
Experiment 5 – Redo previous experiments with “Netflow v5” behavior definition

= ingress interface
= IP source address
= IP protocol
= IP destination address
= source port
= IP type of service
= destination port

3 Results

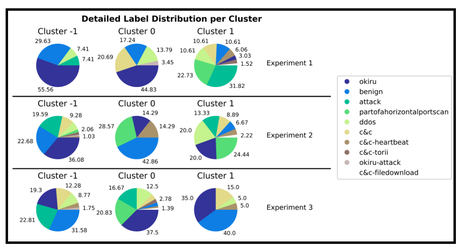
Experiment 1

Taking **8** packets from the beginning leads to the best clustering result. The current default of **20** does not lead to successful clusters.



Experiment 2

Different segments from a connection **do** represent different behaviors and they **do** lead to different clusters.



Experiment 3

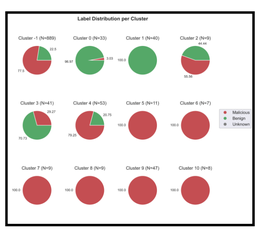
No sub-experiment has performed better than the threshold 8 from experiment 1. **No** benefit can seemingly be obtained from analysing a connection from the end or skipping packets.

Experiment 4

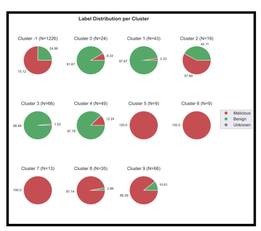
The experiment could **not be finished** for many values due to lack of computing power. The results for the remaining values are **worse** than their equivalent from experiment 1.

Experiment 5

No systematic differences in terms of clustering results can be observed. This is likely because both the current MalPaCA definition and the "Netflow v5" definition are quite **similar**.



MalPaCA behavior definition

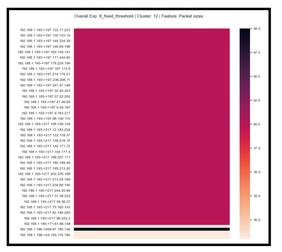


“Netflow v5” behavior definition

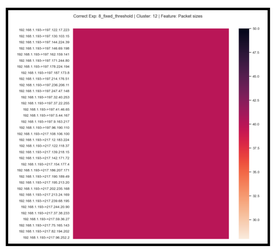
4 MalPaCA Improvements

The “clustering error” metric from the original paper has been improved in that it is now **automatically** generated and the user sees what the **correctly and incorrectly** clustered connections are.

Total Cluster



Correctly Clustered



Incorrectly Clustered