Detecting Collaborative Scanners based on Shared Behavioral Features

Author
Andrei-Iulian Vișoiu - a.i.visoiu@student.tudelft.nl

Supervisor
Dr. Ir. Harm Griffoen

TU Delft

01 Introduction

Background

- **Port Scanning** - sending probe packets to the public IPv4 address to find open ports (services).
- Scanning is often the first stage of a potential cyber attack. ("**Reconaissance**"). Therefore, important to detect and take action.
- **Distributed Scanning** - distributing port scanning on multiple machines, to avoid detection.

Research Gap

- No established method to detect distributed (collaborative) scanning.

02 Research Question

"How can we detect collaborative scanners in network telescope data using clustering algorithms, based on behavioral features?"

Hypothesis: machines working together will exhibit similar scanning patterns.

Subquestions:

- What are the most relevant behavioral features to identify a group?
- What does the group composition look like?
- How can the performance of such an approach be measured?

03 Methodology

1. Data: unsolicited TCP packets captured by TU Delft network telescope, 1-20 Feb 2024. After sanitisation, 921,846 unique source IPs.
2. Aggregate behavioral features: ports, tool used, IP generation algorithm, inter-packet time, total hits, distinct IPs hit
3. Train HDBSCAN (minPts=2,10) and DBSCAN on 2 feature sets, with and without hits and distinct IPs.
4. Evaluate:
   a. Calinski-Harabasz Index
   b. Heavy Hitters - group that sent more than 30.000 packets/day, on average.
   c. Partial Cover - groups that hit more than 32,000 distinct IPs without overlap.
5. Post-Processing: greedy algorithm to find sub-groups with no overlap.

04.3 Results - Post-Processing

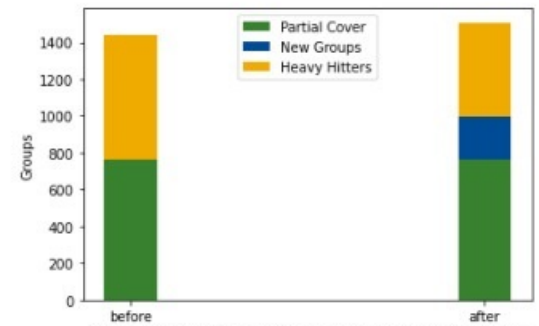

Figure 7: Effect of post-processing on the number of groups with partial covers. Significant increase in the groups satisfying the criteria - 30.32%

04.1 Results - HDBSCAN - minPts=2

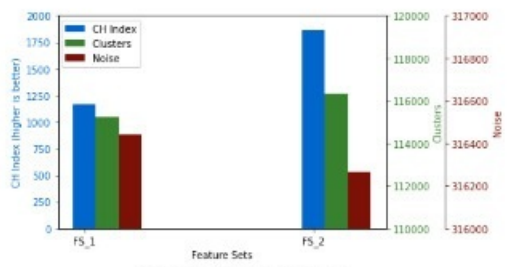

Figure 1: HDBSCAN metrics, minPts=2. Eliminating total hits and distinct IPs shows a positive impact.

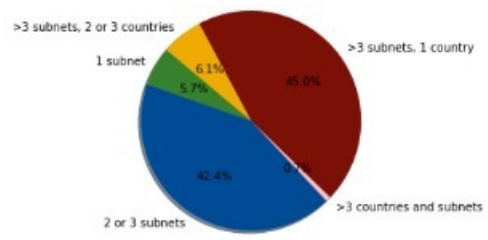

Figure 2: HDBSCAN geographic distribution, FS_2. The majority of the scanning groups are not exceeding country boundaries.

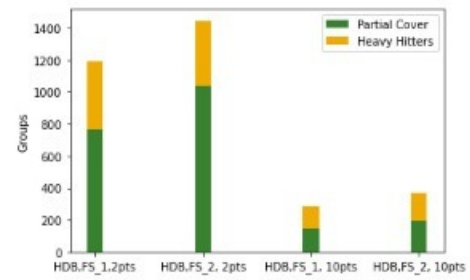

Figure 3:Adress Range Coverage, HDBSCAN Setting the minimum points to 2 reveals more heavy hitters.

04.2 Results - DBSCAN

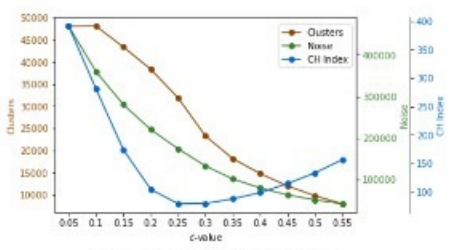

Figure 4: Effect of ε variation, DBSCAN, FS_2 Chosen value for ε is 0.1

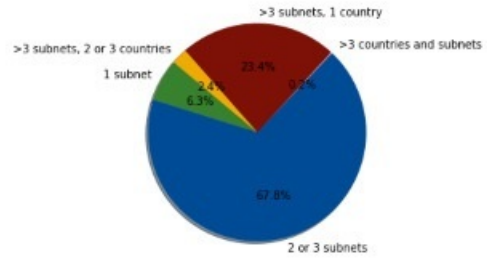

Figure 5: DBSCAN geographic distribution, FS_2, ε=0.1 Change from HDBSCAN. More group with 2 or 3 subnets.

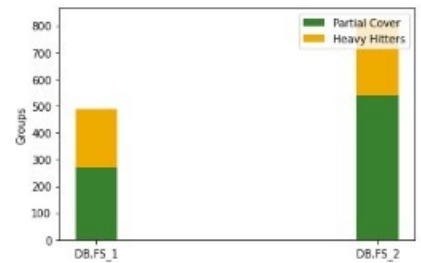

Figure 6: Address Range Coverage, DBSCAN, FS_2, ε=0.1 Lower than HDBSCAN, with comparable noise and lower cluster number.

05 Discussion

Manual group analysis reveals that the overlap criterium might not be the best for all groups.

Group #1 - Label 3025 - 6 IPs in HK, 1 in Tokyo. High overlap of 57028 with 62173 distinct IPs. Overlap happens between HK IPs, although they share very similar patterns. They scan port 22 (SSH), 3389 (Remote Desktop), 3306 (MySQL) and 443 (HTTPS). Overlap might happen because of different credentials per IP.

Group #2 - Label 38 - 4 IPs from China, in 2 different subnets with 2 IPs per subnet. IPs in the same subnet overlap, but combining any IP from the first subnet with any IP from the other subnet results in no overlap and the same hit IPs. Adversary could use machines in multiple subnets to avoid detection and program machines in each subnet to scan a set part of the internet.

06 Conclusions & Future Work

Technique is effective in discovering overall scanning patterns, with HDBSCAN generally more effective.

Main limitation is in the inability to always find full groups (hitting most of the network range). This happens as there is no way to enforce/encourage this behavior. Secondarily, some clusters also contain groups with overlaps.

An indication has been found that scanning groups could be scanning the same IPs multiple times, and this is an interesting point for new research.

Another future research direction - using clustering algorithms that support adding constraints, such as C-DBSCAN to integrate domain knowledge.

Moreover, further analysis could be conducted to design feature sets tailored to each tool and have different models trained.