Detecting Collaborative Scanners Using Clustering Methods

Andrei Ionescu (5492645)

TU Delft BSc Computer Science and Engineering

Objectives

- The main objective of this research is to discover to what degree it is possible to detect collaborative scanning groups using clustering methods.
- The secondary objective of this research is to evaluate and compare the performances of a multitude of clustering methods.

1. Introduction

- **Context**: Cybersecurity involves a continuous cat-and-mouse game between attackers and defenders.
- **Internet Scanning**:
  - Systematic process to gather information on networks and systems.
  - Can be benign (security assessments, research) or malicious (identifying vulnerable systems).
- **Evolution of Scanning Techniques**:
  - Initial scans were massive and detectable via intrusion detection systems (IDS) and firewalls.
  - Attackers evolved to distribute scans over multiple hosts to evade detection.
- **Current Challenge**: Detecting these distributed, collaborative scanning groups is now essential for cybersecurity specialists.
- **Impact of Vulnerabilities**: Exploiting vulnerabilities at scale can lead to sensitive data leaks, service disruptions, and financial damage.
- **Current Solutions**:
  - Gates' set cover technique identifies connections between IP addresses but is impractical for large data volumes due to its NP-complete nature.
  - Robertson et al. identify distributed scans provided that IP addresses are within the same subnet, which can be circumvented by using different subnets.
  - Yegneswaran et al. detect coordinated behavior by examining destination ports and IP addresses, but ignore other fields.
  - Griffioen et al. propose clustering source IP addresses based on scan probes and header values, effective but struggles with scanners that randomize packet fields and send low amounts of packets.

2. Research Questions

The main research question is:

- Is it possible to detect collaborative scanners using clustering methods?

The sub-questions it creates are:

- How do collaborative scans work and what assumptions does the proposed methodology make?
- What packet attributes should be considered for clustering?
- What clustering methods should be used?
- Once a cluster has been identified, how can we check that it is indeed a collaborative scanning group?
- What values should the hyperparameters have?
- If multiple clustering approaches can detect collaborative scanning groups, how do their performances compare?

Important Result

After applying the clustering methods to the unseen evaluation set, a score is computed for each one by taking the mean of the percentages achieved for every degree of certainty. DBSCAN stands out as the optimal clustering method by a considerable margin, as it performs about 247% better than CURE, 1057% better than BFR, 2467% better than K-Means and 7400% better than Hierarchical Clustering.

4. Methodology

- **Aggregating data**: packet data was aggregated into scanning sessions, defined as a stream of packets with no more than 3 hours between each other. If more than 3 hours pass, then that source IP address is considered to begin another scanning session.
- **Training the models:** create clusters based on the temporal structures of the obtained scanning sessions, using their start and end times, as well as the rate at which packets are sent within the sessions.
- **Evaluating the clusters and optimizing the hyperparameters:** train the models with different values and pick the ones which lead to the best clusters based on the degrees of certainty, then use these hyperparameters to apply the models to the test dataset.

3. Explored Clustering Methods

The clustering methods that have been considered are:

- K-Means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- CURE (Clustering Using Representatives)
- BFR (Bradley-Fayyad-Reina)

CURE and BFR differ from the other three alternatives in the fact that they are designed for handling large datasets that do not fit into memory. It should also be mentioned that CURE, BFR, and DBSCAN determine the number of clusters based on the data itself, making them particularly useful for this project since the actual number of clusters is unknown.

5. Results

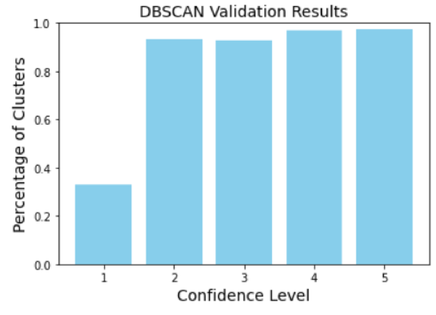

Figure 1:Results for applying DBSCAN using epsilon = 0.0001 and minimum samples = 2 to the validation dataset

The degrees of certainty are represented by levels 1 through 5: (1) same subnet, (2) same ISP, (3) same country, (4) covered by 3 subnets, (5) covered by 3 ISPs.

6. Conclusion

- DBSCAN is effective for detecting collaborative scanning activities by analyzing temporal patterns in packet transmissions.
- The method works without relying on packet content, ensuring adaptability across various tools and scenarios.
- DBSCAN outperformed the other clustering methods.
- Optimal hyperparameters identified: $min\_pts = 2$ and $epsilon = 0.0001$.

7. Future Research

The following points represent areas in which further research would prove beneficial:

- Test DBSCAN performance by injecting artificial collaborative scanner packets into network telescope datasets, using a wide array of scanning tools for realistic evaluation, and measure detection rates.
- Develop energy-efficient algorithms that maintain high performance while adhering to sustainability principles.

Responsible Research

The data used in the research is inherently anonymous, seeing as it is provided from a network telescope which is made up of only inactive IP addresses and all Internet backscatter was removed from it. Therefore, privacy concerns are a non-issue.

Contact Information

- Student Name: Andrei Ionescu
- Student Email: A.Ionescu-5@student.tudelft.nl
- Supervisors: George Smaragdakis, Harm Griffioen

TU Delft