Are Outliers Just Noise? Unpacking Anomalies in Malware Clustering

Bharath Kumar Swargam, Prajna Bhandary, Charles Nicholas
CSEE, UMBC

Abstract

- Outliers in malware clustering are often discarded as noise - but could they hold value?
- We investigate outliers produced by two clustering algorithms **(K-Means LTS and HDBSCAN)** on the MOTIF and SOREL-20M malware datasets using DLL import features.
- We evaluate clustering performance using internal metrics and analyze the overlap, structure, and family distribution of flagged outliers.
- Our findings: outliers are not merely noise, but sometimes unique or misclassified malware that deserve closer inspection.

Motivation

- Outliers in malware clustering are often discarded—but could they reveal mislabeled, novel, or evasive threats?
- We question the assumption that outliers = noise.
- **Goal:** Analyze and interpret clustering outliers using real-world malware datasets.

Methodology

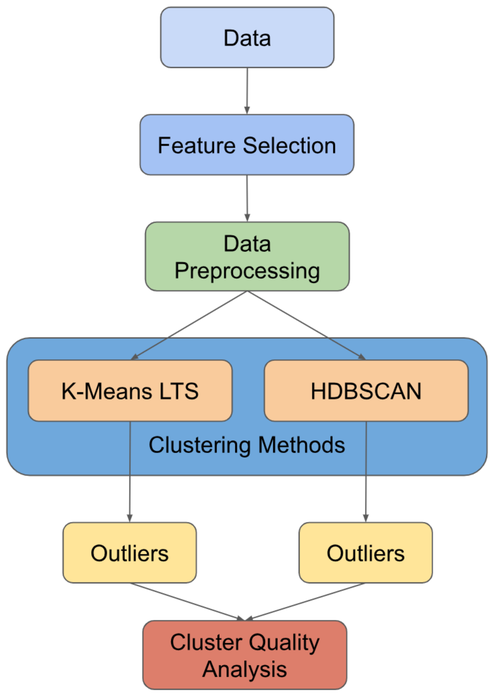
- **MOTIF**: 3,090 well-labeled malware samples (454 families)
- **SOREL-20M (subset)**: 627,298 ransomware samples after filtering
- **Features:** DLL imports (binary vector: presence/absence)
- **Metrics:** Silhouette Score (cohesion/separation), Davies–Bouldin Index (DBI – lower is better)
- **Analysis Focus:** Overlap, function count, family distribution

Presented at MTEM '25, July 15-17, The Aerospace Corporation

Experiment & Results:

Metric	MOTIF		SOREL	
	K-Means LTS	HDBSCAN	K-Means LTS	HDBSCAN
Silhouette Score	0.1981	0.551	**0.2498 (↑10.47%)**	0.6109
Davies–Bouldin Index	1.4320	**0.910(↓3.93%)**	1.6612	1.7235
Outliers / Noise Points	235	1464	15783	17892
Total Samples	2380	2380	157848	157848
Outlier Percentage	9.9%	61.5%	10.0%	11.3%
Overlap Count	157	157	5716	5716
Overlap % (K-Means perspective)	**66.8%**	–	**36.2%**	–
Overlap % (HDBSCAN perspective)	–	**10.7%**	–	**31.9%**
Function Count Stats				
Mean (Outliers)	228.99	**103.75(↓)**	1339.20	560.54(↓)
Mean (Inliers)	108.06	**145.97(↑)**	647.40	736.52 (↑)
Median (Outliers)	195	90	1096	326
Median (Inliers)	86	92	438	514
Min (Outliers)	**120**	**11**	**436**	**44**
Max (Outliers)	798	798	5900	5739

Process Flow



DLL Import Results

Table 1:Top 4 Malware Families Among Outliers (MOTIF and SOREL)

Dataset	Method	Malware Family	%
MOTIF	K-Means LTS	**zerot (15)**	6.4%
		flawedammyy (11)	4.7%
		xmrig (8)	3.4%
		qbot (7)	3.0%
	HDBSCAN	**icedid (112)**	7.7%
		phorpiex (33)	2.3%
		gandcrab (30)	2.0%
		maze (26)	1.8%
SOREL	K-Means LTS	**cerber (4984)**	31.6%
		bunitu (1103)	7.0%
		expiro (658)	4.2%
		cryptxxx (610)	3.9%
	HDBSCAN	**cerber (5098)**	28.5%
		expiro (995)	5.6%
		tofsee (684)	3.8%
		zbot (671)	3.8%

Results

- **K-Means LTS (MOTIF):** More rare or underrepresented families. (e.g.,`zerot`, `xmrig`)
- **HDBSCAN (MOTIF):** Captures more popular/overlapping families. `icedid`, `gandcrab`)
- **Top family in both (SOREL):** `cerber` (most common ransomware)
- **K-Means Outliers:** 1855 unique families
- **HDBSCAN Noise:** 1760 unique families

Conclusion

- Outliers span a broad range of families—possibly mislabeled or evasive
- Outliers are not always noise — they reflect structural or behavioral anomalies
- Clustering method influences what gets flagged as outlier
- Richer outlier analysis could improve early detection of novel threats

Future Work

- Deeper static and dynamic analysis of preserved outlier samples.
- Apply explainable AI tools to understand why specific samples are marked as outliers by clustering algorithms.
- Validate outlier generalizability across other malware datasets to assess whether flagged anomalies are dataset-specific or universally rare.

Contact Information

- zk18813@umbc.edu
- prajnab1@umbc.edu
- nicholas@umbc.edu