Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces

By Shankara Narayanan

Main ideas

1. HTTP-based malware is becoming more prevalent
2. Obfuscation techniques cannot change the malicious activity being performed when executed
3. Hence analyzing the malware network traffic and extracting a generic behavioral signature can be used to detect future malware variants with low FP and FN
4. Two-stage clustering as direct clustering on a large dataset will prove computationally infeasible
5. After clustering, a generic HTTP signature is extracted which is then provided to an IDS at the edge of the network which will monitor the traffic for malware activities

Steps involved

1. Coarse grain clustering based on the statistical features extracted from their malicious HTTP traffic (Total GET, POST, the average length of URLs, etc.)
2. Fine grain clustering based on the structural similarity between the HTTP traffic generated by each sample
3. Cluster merging and meta-clustering to get a representative signature for a cluster

Techniques analyzed

1. Single linkage clustering + Davies-Boulding (DB) cluster validity index
2. Ward Linkage Hierarchical Clustering
3. K – Means clustering

Cleaning the dataset (Deceptor + Downloader dataset)

1. When analyzing the columns in the dataset, it is clear that there are a lot of empty cells filled with 0, -1, nan, or []. This needs to be cleaned by checking the percentage of presence of such entries in each column. If a column contains mostly of these values then they need to be dropped. A threshold of 89% was chosen and any column which contains more than 89% of these values will be dropped from the data frame.

After dropping, the percentage of the presence of such default values can be found in the below table.

2. Let these 24 columns be called “good_cols” but it contains target labels in various columns and they need to be removed. With this, the data frame is ready for feature engineering

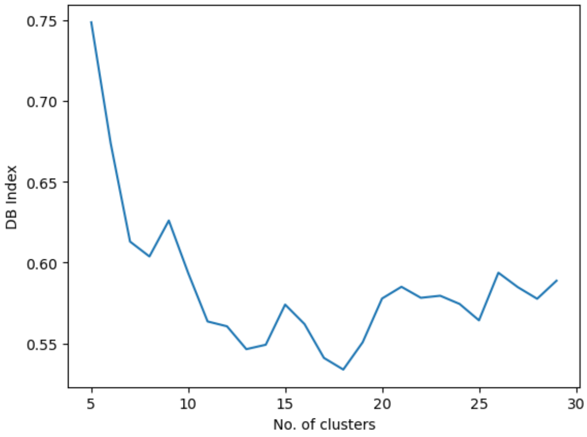
Chi-squared analysis for feature engineering

1. From the selected 17 columns, feature engineering must be done to pick the most prominent columns contributing to the clustering process

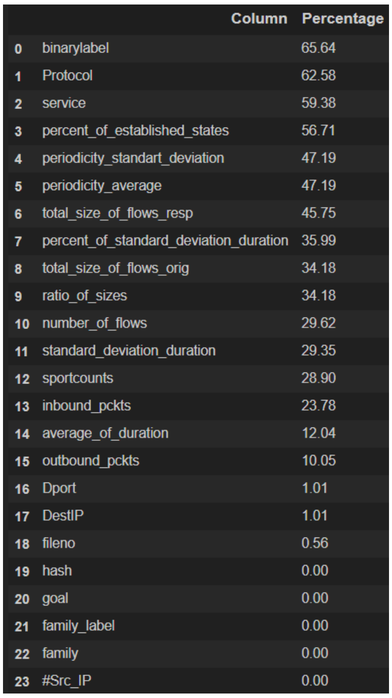
2. Selecting the 6 best features from the pool of 17 columns, we get ['DestIP', 'Dport', 'percent_of_established_states', 'inbound_pckts', 'outbound_pckts', 'fileno'] as the 6 best features

Single-linkage clustering and DB cluster validity index

1. Performing agglomerative clustering for clusters of size 5-30
2. DB score is calculated and stored in an array. This can be visualized as a plot between the DB score and the cluster size

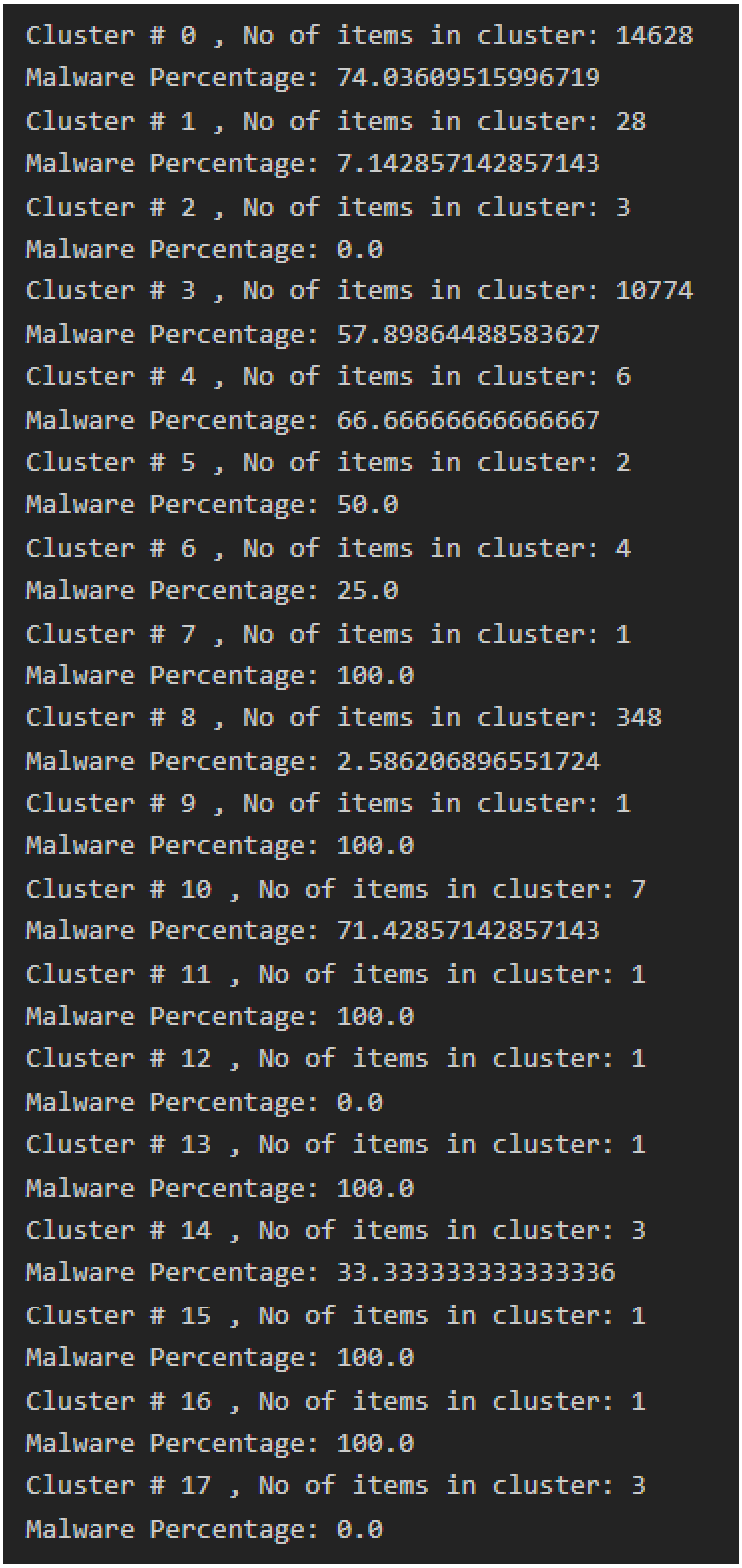


3. From the plot, it is clear that at around **18 cluster size**, the DB index is minimized hence that is the optimal cluster size



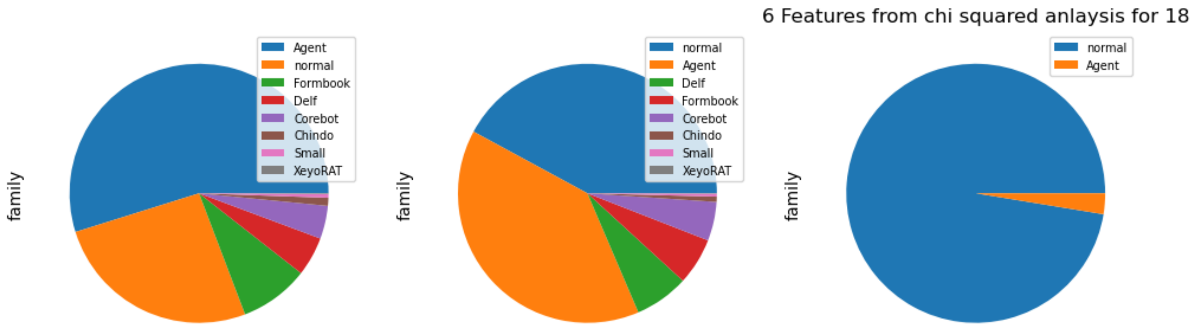
	Column	Percentage
0	binarylabel	65.64
1	Protocol	62.58
2	service	59.38
3	percent_of_established_states	56.71
4	periodicity_standart_deviation	47.19
5	periodicity_average	47.19
6	total_size_of_flows_resp	45.75
7	percent_of_standard_deviation_duration	35.99
8	total_size_of_flows_orig	34.18
9	ratio_of_sizes	34.18
10	number_of_flows	29.62
11	standard_deviation_duration	29.35
12	sportcounts	28.90
13	inbound_pckts	23.78
14	average_of_duration	12.04
15	outbound_pckts	10.05
16	Dport	1.01
17	DestIP	1.01
18	fileno	0.56
19	hash	0.00
20	goal	0.00
21	family_label	0.00
22	family	0.00
23	#Src_IP	0.00

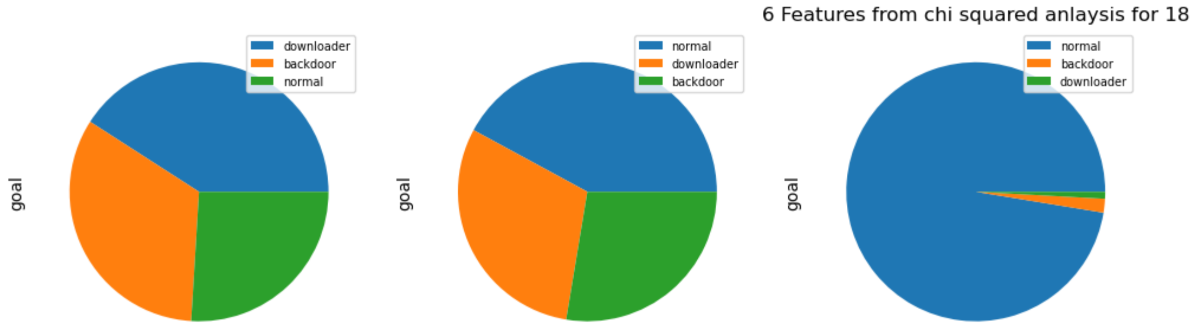
4. Inside the 18 clusters, we have to do two things. First, get the accuracy of clustering between malware and benign. Secondly, we have to get the distribution of malware inside the clusters.



```
Cluster # 0 , No of items in cluster: 14628
Malware Percentage: 74.03609515996719
Cluster # 1 , No of items in cluster: 28
Malware Percentage: 7.142857142857143
Cluster # 2 , No of items in cluster: 3
Malware Percentage: 0.0
Cluster # 3 , No of items in cluster: 10774
Malware Percentage: 57.89864488583627
Cluster # 4 , No of items in cluster: 6
Malware Percentage: 66.66666666666667
Cluster # 5 , No of items in cluster: 2
Malware Percentage: 50.0
Cluster # 6 , No of items in cluster: 4
Malware Percentage: 25.0
Cluster # 7 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 8 , No of items in cluster: 348
Malware Percentage: 2.586206896551724
Cluster # 9 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 10 , No of items in cluster: 7
Malware Percentage: 71.42857142857143
Cluster # 11 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 12 , No of items in cluster: 1
Malware Percentage: 0.0
Cluster # 13 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 14 , No of items in cluster: 3
Malware Percentage: 33.333333333333336
Cluster # 15 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 16 , No of items in cluster: 1
Malware Percentage: 100.0
Cluster # 17 , No of items in cluster: 3
Malware Percentage: 0.0
```

5. It is clear that except clusters 0, 3 and 8, all clusters have very little items in them. Hence, we are going to see the visualization of these three clusters alone, family wise and goal wise

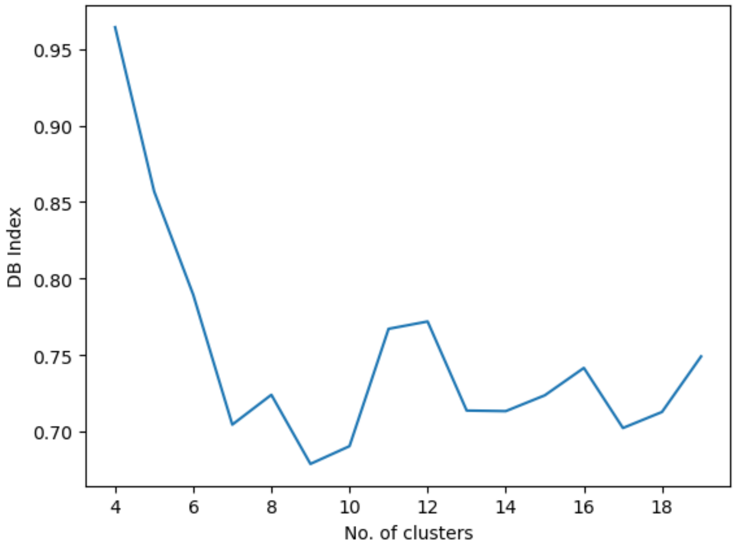




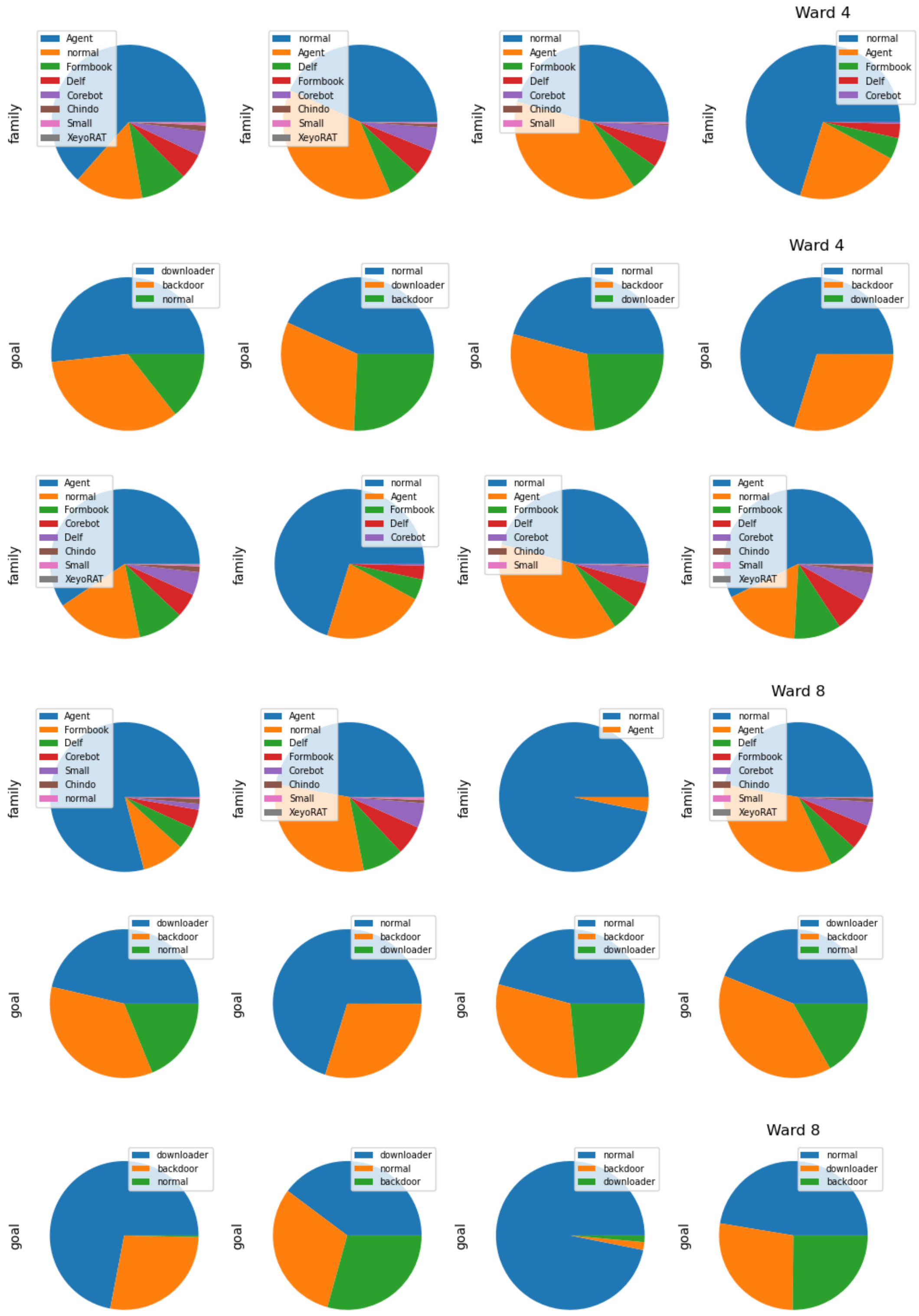
6. The Accuracy of the clustering process is: **67.64%**

Ward Linkage Hierarchical clustering

1. Using the same Agglmerative clustering but this time with linkage == ‘ward’
2. Check the cluster sizes 4-20



3. The curve shows an optimal number of clusters as **9**
4. Turns out, cluster number **4 and 8** have the most number of elements and their visualization are as follows



5. The accuracy of the clustering process is: **72.33%**

K-Means clustering

1. Applying the K-Means clustering process after dropping irrelevant columns and fitting an SVM and an ANN for 100 epochs

2. This analysis is done to check if the traditional ML process is enough for this task but the accuracy proves the importance of the clustering process
3. SVM accuracy: **57.83%**
4. ANN accuracy (100 epochs): **56.61%**

Inferences

1. In the paper, Single-linkage hierarchical clustering gave the best results while in our case **ward linkage** gave the best results
2. Traditional ML processes cannot be used directly and thus this validates the procedure of clustering present in the paper. Traditional processes give inferior accuracy compared to the prescribed techniques

Experiments on each class of malware

The same set of 3 clustering process was carried out for each of the classes of malware found in the RaDaR dataset.

The classes of malware are: backdoor, cryptominer, deceptor, dropper, pua, ransomware and spyware

1. Backdoor and Normal

Single Linkage optimal number of clusters	26
Single Linkage Clustering accuracy	58.78%
Ward Linkage optimal number of clusters	23
Wark Linkage Hierarchical Clustering accuracy	**63.39%**
SVM classifier accuracy	49.95%
ANN classifier accuracy	51.40%

2. Banker and Normal

Single Linkage optimal number of clusters	18
Single Linkage Clustering accuracy	61.77%
Ward Linkage optimal number of clusters	27
Wark Linkage Hierarchical Clustering accuracy	64.98%
SVM classifier accuracy	**65.79%**
ANN classifier accuracy	63.46%

3. Cryptominer and Normal

Single Linkage optimal number of clusters	13
Single Linkage Clustering accuracy	63.74%
Ward Linkage optimal number of clusters	40
Wark Linkage Hierarchical Clustering accuracy	63.74%
SVM classifier accuracy	**68.73%**
ANN classifier accuracy	67.69%

4. Deceptor and Normal

Single Linkage optimal number of clusters	24
Single Linkage Clustering accuracy	63.51%
Ward Linkage optimal number of clusters	45
Wark Linkage Hierarchical Clustering accuracy	63.99%
SVM classifier accuracy	**65.00%**
ANN classifier accuracy	64.17%

5. Downloader and Normal

Single Linkage optimal number of clusters	25
Single Linkage Clustering accuracy	66.05%
Ward Linkage optimal number of clusters	24
Wark Linkage Hierarchical Clustering accuracy	**78.49%**
SVM classifier accuracy	71.16%
ANN classifier accuracy	70.37%

6. Dropper and Normal

Single Linkage optimal number of clusters	24
Single Linkage Clustering accuracy	63.51%
Ward Linkage optimal number of clusters	26
Wark Linkage Hierarchical Clustering accuracy	63.51%
SVM classifier accuracy	**65.00%**
ANN classifier accuracy	64.17%

7. PUA and Normal

Single Linkage optimal number of clusters	13
Single Linkage Clustering accuracy	60.69%
Ward Linkage optimal number of clusters	33
Wark Linkage Hierarchical Clustering accuracy	**63.93%**
SVM classifier accuracy	61.89%
ANN classifier accuracy	62.81%

8. Ransomware and Normal

**** There are no ransomware traces found in the dataset**

9. Spyware and Normal

Single Linkage optimal number of clusters	25
Single Linkage Clustering accuracy	65.69%
Ward Linkage optimal number of clusters	37
Wark Linkage Hierarchical Clustering	**70.76%**

accuracy	
SVM classifier accuracy	46.73%
ANN classifier accuracy	46.63%

Inferences

1. For Backdoor, PUA and Downloader => Best classifier model => Ward Linkage Clustering
2. For Banker, Cryptominer, Deceptor and Dropper => Best classifier model => SVM classifier