The influence of sequence length on the clustering performance of MalPaCA

Johannes Hagspiel, Azqa Nadeem, Sicco Verwer

Abstract

MalPaCa is a novel, unsupervised clustering algorithm, which creates based on the network flow of a software a behavioral profile representing its actual capabilities. One of the key variables affecting is performance and usability is the sequence length or how many packets it analyzes in order to group a connection to a cluster. This article explore different sequence length amounts as well as different positions from where to extract the packets from. The findings indicate that the current default sequence length of 20 is too high, leading in many cases to no clusters being found. 8 has been determined to be the optimal length for both performance and efficiency. Additionally, it has been established that using a windowed approach whereby a connection is being sliced into multiple, smaller connections could make MalPaCa more usable in situations where connection lengths are unequally distributed. Furthermore, MalPaCas usability as a tool has been improved by automating the clustering error metric determination, thereby providing the user with a valuable, visual measure as to how well MalPaCa has fared in creating the clusters. Lastly, MalPaCas definition of behavior was compared to the "Netflow v5" behavior definition, however no substantial performance improvements could be obtained from this change.

1 Introduction

1.1 Background & Motivation

Malware Classification Problem

There has been a rapid growth in the numbers of malware variants in the past years which in part is the result of improved malware obfuscation techniques (Li, Liu, Gao, & Reiter, 2010, p.238). As an example, for the Bagle/Beagle worm, over 30,000 unique variants have been found only in the time between January and March 2007 (Commtouch, 2007)(Li et al., 2010, p.1). Correspondingly, the need to accurately identify and classify malware has only become more important. Whereas in the beginning, most of malware classification was done manually, recently more and automated approaches have been developed in order to deal with the explosion in malware variants (e.g. Bayer, Comparetti, Hlauschek, Kruegel, & Kirda, 2009; Bailey et al., 2007).

However, the current process of assigning labels has been shown to suffer from a series of shortcomings. For instance, anti-virus (AV) labels are notorious for not being consistent in that the same binary is frequently classified by one AV engine as benign and by the other as malicious (Sebastián, Rivera, Kotzias, & Caballero, 2016, p.1)(Bailey et al., 2007). Another, more fundamental problem with the labels assigned by an AV engine is that they either not accurately reflect the behavior of a sample or they only highlight one specific behavior, which is particularly troubling as multi-vector attacks are becoming more common (Bailey et al., 2007, p.179). And as the rules that these engines follow essentially are a "black-box", it is difficult to double-check the assigned labels (Nadeem, Hammerschmidt, Gañán, & Verwer, 2021, p.382).

MalPaCA

MalPaCA has been developed in response to these challenges. It is an unsupervised clustering algorithm, which creates based on the network flow of the sample a behavioral profile representing the actual capabilities of a program (Nadeem et al., 2021, p.383).

MalPaCA consists of five phases. First, the actual network flow generate by a software in the form of Pcap files is split into uni-directional connections. In step two, four features are extracted: packet size, time interval between packet arrival, source port and desination port (Nadeem et al., 2021, p.391).Then, temporal distances between connections are calculated, based on which the HDBScan clustering algorithm generates clusters. In the last step, temporal heatmaps are produced for each cluster and for each feature which are turned into a cluster membership string.

Motivation

For the actual clustering, MalPaCa uses the temporal similarity between connections in order to group them together (Nadeem et al., 2021, 382). This temporal similarity is in the base version established based on only the first 20 packets of a connection. Additionally, the behavior observed by MalPaCa is defined as the uni-directional flow of packets from a source to a destination IP address. Determining the optimal sequence length as well as the best segments in a connection from where to obtain the packets used as an input is intimately related with the usability and performance of MalPaCa. Therefore, in order to improve MalPaCas viability for malware clustering, these underlying assumptions need to be questioned and analyzed.

1.2 Research Question

The main research questions this thesis thus wants to answer are:

"RQ 1: Which packets in a network connection are required to characterize behavior?"
"RQ 2: How many packets in a network connection are required to characterize behavior?"

To determine the section of a network connection that best describes the underlying behavior, it first has to be established whether different parts of a connection relate to different behavior. Therefore, question 1 can be further broken down into:

"RQ 1.1: Do different segments of a network connection represent different behavior?"

Based on the answer to this question, it can then be discovered which segment or segments are most characteristic of the behavior of a software.

"RQ 1.2: Which network connection segment or segments are most indicative of the underlying behavior?"

Additionally, a third question is related with these previously described questions namely:

"RQ 3: On which level should behavior best be defined in order to capture this behavior in Malpaca?"

1.3 Relevant Literature

Roeling, Nadeem, and Verwer have used MalPaCa in order to find a novel way for using clustering to analyze " spatial-temporal network data" (Roeling, Nadeem, & Verwer, 2020). In this paper, they also tackled the issue of which sequencing length to chose. However, their experimentation was limited to testing four different threshold values; 5, 10, 15 and 20. They did not explore threshold values larger than 20 and they stuck to MalPaCas default mode of selecting packets from the beginning of a connection. Lastly, they also did not investigate whether a completely different definition of behavior than the unidirectional definition used by MalPaCa might not be more appropriate.

In general, a frequently selected tactic to avoid having to deal with these issues is to simply turn sequential data into aggregate values (Roeling et al., 2020, p.3)(Saad et al., 2011)(Strayer, Lapsely, Walsh, & Livadas, 2008). Other attempted solutions include using "time windows" or filtering out connections based on pre-established criteria (Roeling et al., 2020, p.3)(Gu, Zhang, & Lee, 2008)(Cai, international conference on Wireless, & undefined 2012, n.d.). However behavior is defined, be it on the host level or on the unidirectional level, some information is always lost. The question now which strategy is most appropriate for MalPaCa and leads to the best clustering performance.

MalPaCas decision to analyze behavior based on the unidirectional flow of packets from source to destination IP address is one that is not commonly observed in the existing literature. Much more frequently one encounters for instance the "Netflow" definition of a connection, i.e. (Sarhan, Layeghy, Moustafa, & Portmann, 2020)(Grill, Nikolaev, Valeros, & Rehak, 2015)(Kheir & Wolley, 2013). In line with MalPaCa, the 5th version of "netflow" is also unidirectionl, it however defines a flow based on seven criteria: 1. Source IP Address, 2. Destination IP Address, 3. Source Port, 4. Destination Port, 5 IP Type of Service, 6. IP Protocol and 5. Router interface (InterProjektWiki, 2021). "Netflow" is not only much more commonly encountered in Academia, it is also one of the most widely implemented standards for network traffic flow aggregation (Petryschuk, 2019). It is so popular in fact that it has also been turned into an official standard by the IETF in the form of the "IP Flow Information Export" (IETF, 2021).

2 Methodology

2.1 Dataset

As stated previously, the dataset used in this experiment is the publicly available "IoT-23 Dataset" by the Avast AIC laboratory (Parmisano, Garcia, & Erquiaga, 2020). The specialty of this dataset is that it has labelled the captured network flow both whether it is benign or malicious (hereafter referred to as the label of a connection) as well as what kind of behavior it is (hereafter referred to as the detailed label of a connection).

These detailed labels are not equally distributed. On the contrary, four kinds of behavior namely "benign", "okiru", "ddos" and "partofahorizontalportscan" together make up over 99,977 percent of all the captured traffic. Given that the malware observed in the "IoT-23 Dataset" are 11 different kinds of botnets and the remainder are three benign IoT home devices, this at least partially explains this distribution. The prevalence of both the "ddos" and "partofahorizontalportscan" detailed label makes sense given that for both tasks, a botnet bombards a target with a huge number of of connection requests.

The information provided by Avast online regarding the label distribution does not entirely match the information found in the actual data set. For one, 425320 less benign connections can be found in the data set. This is likely due to the fact that not for all the connections in the pcap files of the three benign scenarios can one find a corresponding label. Additionally, whereas online, it is stated that there are 60990708 "okiru" and 3 "okiru-attack" connections in the "CTU-IoT-Malware-Capture-33-1" scenario, one in fact finds 47381241 "okiru" and 13609467 connections labelled as "okiru-attack" in the actual data set.

IoT-23 Dataset

An important issue to note is that the "IoT-23 Dataset" has defined network behavior on a bi-directional flow level based on characteristics such as i.e. the host and destination IP address as well as port in addition to the protocol used and not as a uni-directional connection, solely dependent on source and destination IP address as Malpaca does.

There are significant differences as to the distribution of detailed labels when comparing these two ways of aggregation. In short, moving from the flow level to malpacas uni-directional connection level means that rarely occurring labels are becoming even rarer whereas more common behavior is even more widespread. This observation is a key insight for question 3, "On which level should behavior best be defined in order to capture this behavior in Malpaca ?", as by simply defining behavior as a uni-directional connection based only on the source and destination IP address, Malpaca has it made less likely to capture all of the behavior present in the data set given that the HDBScan algorithm used for the actual clustering is prone to placing less frequent data in a noise cluster. For the exact differences in distribution, see Figures 34 and 32 in the appendix.

Dataset Processing Pipeline

In order to obtain from the original "IoT-23 Dataset" a dataset to be used for the actual experimentation, a data processing pipeline was set up in which both data filtering as well as data enriching steps were carried out.

Two metrics were used when setting up this pipeline: realism and usability.

Usability was mainly concerned with the size of the data set. The full, original dataset was over 60GBs in size and thus it was simply too larget to be used in its entirety by the equipment available. Therefore, in a first step, only those connections were filtered out that had a minimum length of 5 packets and a maximum length of 1000. Additionally, one scenario, namely scenario "CTU-IoT-Malware-Capture-60-1" was excluded simply because it was too large with its 22 GBs. Even with this exclusion, it took three, full days to sieve out the required connections.

In the next two stages, information was then added to this filtered datatset. First, this came in the form of looking up the label as well as detailed label information of each connection from the separate Zeek files. The different aggregation levels used by Malpaca and Avast once again became a problem as the label information was again only available on a flow level. Therefore, for a few uni-directional connections, this information was only available when the source and destination IP address were switched. The decision was made to make as few assumptions as possible and instead rely on the provided information were available. For this reason, the few connections were no detailed labels were available or the information was only available with the IP addresses switched were discarded.

As the detailed labels only provide a coarse description of the actual underlying behavior of the captured network traffic, the decision was made to look for a secondary classification of the connections. Another, potentially corroborating source of information could be the application protocol used in a connection, which can be determined via deep-packet inspection. NFStream which is based on the deep-packet inspection library nDPI is an open source framework that provides exactly this kind of information. However, NFStream also uses a different aggregation level from Malpaca, which again is based on the flow level where it takes into account characteristics like the transport protocol or the involved ports. Therefore, by combining application protocol information that was intitally established on a lower level, potentially one could assign multiple distinct behaviors to the same connection. Yet, an analysis of the resulting aggregation reveals that on average, only one unique behavior is assigned to a connection. Therefore, the decision was made to include and use deep-packet inspection in order to obtain a better vision of the actual behavior under study.

The second metric, realism, guided the fourth step where from this filtered and enriched dataset, a smaller as well as more balanced dataset was constructed. As the original "IoT-23 Dataset" is based on the sandboxed behavior of actual malware, it is reasonable to assume that it is a realistic depiction of malware behavior.Therefore, in order to obtain a balanced dataset, the original detailed label ratios as seen in Figure 32 should be preserved as closely as possible. However, simply downsizing the original dataset was not possible given the fact that by filtering out all connections with a length shorter than five packets, the detailed label distribution has changed drastically, as can be seen in Figure **??** found in the appendix. The main change is that the "partofahorizontalportscan" detailed label constitutes now with 95.39% the vast majority of observable behavior. In order to address this issue, a multi-step process was developed with attempted to extract from the filtered dataset a balanced subset based on the detailed label ratios from the original dataset. An example of how this process looked like in detail can be found in Appendix 2 - "Balanced Dataset Creation Example". In the end, as can be seen in Figure 38, the biggest difference in the resulting "10000" dataset is that the "ddos" and the "okiru" behaviors are underrepresented due to the fact that not enough connections of this behavior remained in the filtered data set while the "benign" label had to be over-represented in order to make up for this shortfall.

However, this selection of connections that is predetermined in regards to the detailed label and random in regards to everything else, resulted in a data set were the average connection length for the most common detailed labels was quite short, as depicted in Figure 39 found in the appendix. As will be described in more detail in chapter 4, these connections were in fact too short for most experiments to result in valid clusters. For that reason, the decision was made to create a secondary balanced data set, "Min 20", which has roughly the same detailed label ratios as can be seen in Figure 40 in the appendix. However, instead of randomly selecting connections with the needed detailed labels, only those connections were taken into consideration which had a minimum length of 20.

In the end, by combining the deep packet inspection output from NFStream with the available labels from the "IoT-23 Dataset", 5 sources of information were available for each connection: label (whether a connection was malicious or benign), detailed label (what kind of behavior is associated with the connection), application name (what kind of protocol was used i.e. Telnet), application category name (to what kind of category the protocol belongs to i.e. RemoteAccess) as well as the name (the name of the malware).

Both of these balanced datasets were then used in the final step to create the actual input into Malpaca based on the different experimental configurations, which can be found in chapter 4. All information associated with these discussed datasets can be found in detail in Appendix 5 - "Dataset Information".

2.2 Evaluation Metrics

In order to evaluate the clusters created by Malpaca, a number of different evaluation criteria were established that fell in roughly three different categories. The actual formulas used in order to compute these metrics can be found in Appendix 1 - "Metric Calculation".

2.2.1 Validity

The first criteria is whether the sequence number results in a clustering that is valid, which comprises of a number of different sub-metrics:

- general clustering quality metrics: these include common measures such as the silhouette score and HDBScan's validity index. Broadly speaking, both of them rate how far apart individual clustering groups are and how closely together members of one cluster are to each other. A good silhouette score and validity index is one that is equal to 1.
- cluster purity: this metric measures how pure a cluster is in regards to the five sources of information. A good cluster is one, that has only i.e. one kind of behavior present. A good cluster purity is one that is close to 1.
- label cohesion: this metric assesses how concentrated all instances of i.e. one detailed label are as the goal would be to have each detailed label not spread out over multiple clusters but rather focused in one.

The five purity and cohesion scored for the five sources of information were also combined into one "cohesion score" and one "purity score" for

Lastly, the "clustering error" measure as first described in the original paper is being employed once again in order for the results to be comparable with the initial findings. However, whereas determining the clustering error as intended by the original authors was a manual process, requiring inspecting per cluster all four heatmaps and then making an educated guess as to who is the rightful owner of a cluster, this paper attempts to improve upon this by automating this procedure. Like in the original paper, the algorithm still uses not the underlying data but the visual output in form of the produced heatmaps. In its new automated form, the algorithm also produces as its output what is deems to be clustered incorrectly and what clustered correctly, thereby directly providing the user with a visual overview of how the metric has come to its conclusion. An example of the new clustering error algorithm can be found in "Appendix 3 - Clustering Error Algorithm".

2.2.2 Reliability

Reliability is measured through two different metrics which are established by running the HDBScan algorithm ten times over the same data set:

- percentage cluster change: measures how frequently a particular connection is assigned to different clusters in different HDBScan iterations.
- percentage probability change: measures how frequently the probability changes with which a particular connection is assigned to a clusters in different HDBScan iterations.

Obviously the reliability of MalPaCas clustering might not depend on the input data but rather on the algorithm used. This, however, should easily be detectable as in the latter case, a systematic trend should be visible over all the different experiments.

2.2.3 Usability

The last criteria, usability, mainly centers around the amount of time needed for MalPaCa to finish creating its groupings. Only the time needed to determine the distance matrix and for the actual clustering is measured.

2.3 Netflow Behavior Definition

As established in Chapter 1, "Netflow" is one of the most common network flow aggregation standards. For that reason, it has was selected as the alternative to compare MalPaCas unidirectional definition against. Based on the seven features of "Netflow v5", the whole dataset processing pipeline was undergone once again to created the two balanced datasets.

3 Contribution

The original paper has already highlighted "performance optimizations" as one of the key areas to be focused on in future research (Nadeem et al., 2021, p.179). Already then was the dynamic-time warping algorithm singled out as a bottleneck of the problem and indeed. This indeed turned out be the case as during the initial runs, when using the original MalPaCa source code, the distance matrix was not calculated even after one hour. By switching to an algorithm that is based on the "Numba" compiler, the same distance matrix could now be computed in a manner of minutes.

As will be shown in chapter 4, the current default of taking the first 20 packets from the beginning of a connection does in some cases not lead to any valid cluster being found as not enough connections meet this requirement. Additionally, as calculating the distance matrix for the four features has a $O(N^2)$ time complexity, selecting too many connections can be prohibitively time-consuming. And depending on the computing resources available, it can also lead to memory issues causing MalPaCa to crash. Therefore, on a practical level, finding a sequence length that balances these two concerns is crucial in order to ensure that MalPaCa can actually be used for malware research purposes.

On a practical level, MalPaCa functionality as a tool has also been improved. For one, in addition to heatmaps, now a number of extra graphs are created to provide a more in-depth visual look into the composition of each cluster. As stated previously, the clustering error metric has also been improved upon as it is now automatically calculated without any manual heatmap inspections. Additionally, it also outputs both what it deems to be correct and what it deems to be incorrect clusterings, thereby providing the user with a visual means to determine the veracity of this metric. An example output of the clustering error algorithm can be found in "Appendix 3 - Clustering Error Algorithm".

4 Experimental Setups, Results and Discussion

4.1 Experimental Setups

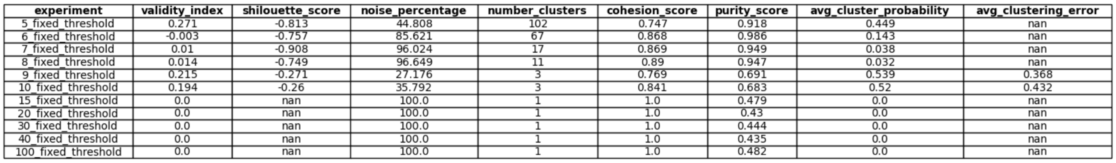
Four different experiments were conducted with both balanced data sets. The total output of these experiments are too large to include them here, they however can be found in their entirety in Appendix 4 - "Experiment Results". What follows is an excerpt of most key results of each experiment, again for the full findings refer to the Appendix.

4.2 Reliability

What can be said even before going into detail into the different experiments is that for all the different values and both data sets, neither "percentage cluster change" nor "percentage probability change" ever take on another value than 0. Therefore, one can conclude that the HDBScan algorithm used by MalPaCa is deterministic and that both the sequence length as well as the position in a connection from where the packets are being selected have no influence on the reliability of the resulting clustering.

Experiment 1 - What sequence length taken from the start of a connection leads to the best clustering results?

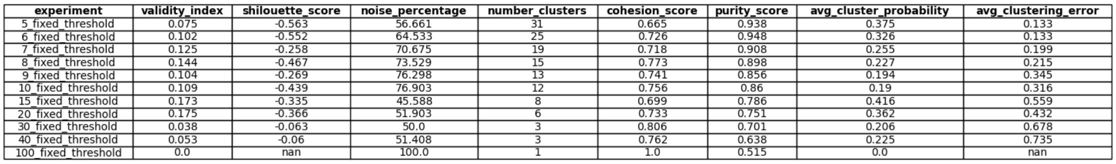
The goal of the first experiment is to emulate the current behavior of Malpaca and therefore take from the start of each connection a fixed amount of packets. What varies here is the number of packets selected. The values 5, 10, 15 and 20 have been taken from the previously discussed article by Roeling, Nadeem, and Verwer (Roeling et al., 2020). They have been extended by 30, 40 and 100 to also explore how values larger than the current default fare.



experiment	validity_index	shilouette_score	noise_percentage	number_clusters	cohesion_score	purity_score	avg_cluster_probability	avg_clustering_error
5_fixed_threshold	0.271	-0.813	44.808	102	0.747	0.918	0.449	nan
6_fixed_threshold	-0.003	-0.757	85.621	67	0.868	0.986	0.143	nan
7_fixed_threshold	0.01	-0.908	96.024	17	0.869	0.949	0.038	nan
8_fixed_threshold	0.014	-0.749	96.649	11	0.89	0.947	0.032	nan
9_fixed_threshold	0.215	-0.271	27.176	3	0.769	0.691	0.539	0.368
10_fixed_threshold	0.194	-0.26	35.792	3	0.841	0.683	0.52	0.432
15_fixed_threshold	0.0	nan	100.0	1	1.0	0.479	0.0	nan
20_fixed_threshold	0.0	nan	100.0	1	1.0	0.43	0.0	nan
30_fixed_threshold	0.0	nan	100.0	1	1.0	0.444	0.0	nan
40_fixed_threshold	0.0	nan	100.0	1	1.0	0.435	0.0	nan
100_fixed_threshold	0.0	nan	100.0	1	1.0	0.482	0.0	nan

Figure 1: 10000 Dataset - Experiment 1 - Results

The first key result observed is that for the "10000" data set, no sequence length above 10 results in a valid clustering as instead all the data is being put into the noise cluster as seen in Figure 1. Again, looking at the average connection length per detailed label of the balanced dataset as seen in Figure 39, this is no surprise as the most frequent behaviors "benign" and "partofhorizontalportscan" have an average connection length of 6.5 and 7.5 respectively. Therefore, with the current default value of 20, MalPaCa would not lead to a successful clustering.



experiment	validity_index	shilouette_score	noise_percentage	number_clusters	cohesion_score	purity_score	avg_cluster_probability	avg_clustering_error
5_fixed_threshold	0.075	-0.563	56.661	31	0.665	0.938	0.375	0.133
6_fixed_threshold	0.102	-0.552	64.533	25	0.726	0.948	0.326	0.133
7_fixed_threshold	0.125	-0.258	70.675	19	0.718	0.908	0.255	0.199
8_fixed_threshold	0.144	-0.467	73.529	15	0.773	0.898	0.227	0.215
9_fixed_threshold	0.104	-0.269	76.298	13	0.741	0.856	0.194	0.345
10_fixed_threshold	0.109	-0.439	76.903	12	0.756	0.86	0.19	0.316
15_fixed_threshold	0.173	-0.335	45.588	8	0.699	0.786	0.416	0.559
20_fixed_threshold	0.175	-0.366	51.903	6	0.733	0.751	0.362	0.432
30_fixed_threshold	0.038	-0.063	50.0	3	0.806	0.701	0.206	0.678
40_fixed_threshold	0.053	-0.06	51.408	3	0.762	0.638	0.225	0.735
100_fixed_threshold	0.0	nan	100.0	1	1.0	0.515	0.0	nan

Figure 2: Min 20 Dataset - Experiment 1 - Results

The "Min 20" data set does not suffer from the same problem as seen in Figure 2, as for all but the 100 packet sequence length, HDBScan resulted in valid, albeit oftentimes small clustering groups. Again, by including only those connections in the "Min 20" balanced dataset that are longer than 20, obviously this leads to better clustering results at higher threshold values.

However, the best result was achieved with a sequence length of eight. This is in line with the results of the previously mentioned paper by Roeling, Nadeem, and Verwer as they in the end identified 10 to be the optimal value (Roeling et al., 2020). As can be seen in Figure 3, at that threshold, the underlying behavior as represented by the detailed labels is nicely distributed over the clusters.

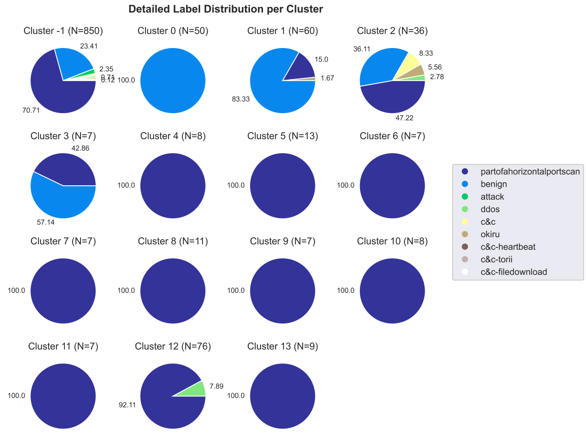


Figure 3: Min 20 Dataset - 8 Threshold - Detailed Labels per Cluster

Additionally, this threshold also does a great job of separating malicious from benign behavior as can be seen in Figure 4.

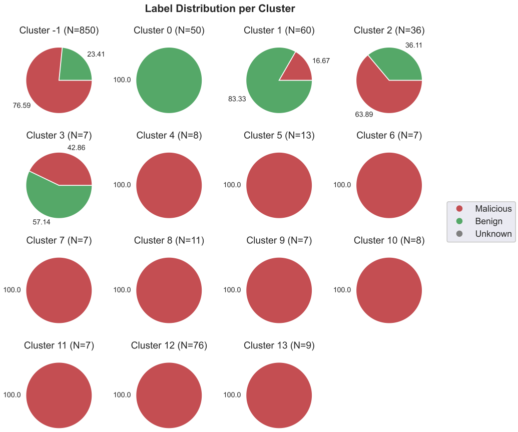


Figure 4: Min 20 Dataset - 8 Threshold - Labels per Cluster

Benign behavior is concentrated in clusters 0, 1, 2 and 3, leading to a high overall purity score of 0.947 and a cohesion score of 0.89. Unfortunately, the algorithm used to determine the clustering error could not handle this particular dataset, thereby no corresponding value can be reported.

Neither dataset performed particularly well in the general clustering quality metrics categories. Particularly on the silhouette score perform both datasets poorly as neither achieves even a positive value for one experiment. This indicates that on conventional clustering quality metrics, these resulting clusters do not perform well. For the "clustering error" metrics, a clear trend can be determined in that the smaller the threshold is, the smaller the clustering error becomes. This could be due to the fact that "partofhorizontalportscan" makes up a vast majority of the connections present in both datasets. This behavior leads to very short connections, and what can be observed is that as the threshold decreases more and more clusters pop up that only contain "partofhorizontalportscan". Therefore, the clustering error decreases with decreasing threshold values due to the fact that in this scenario, there are now many clusters with only one behavior present, namely "partofhorizontalportscan".

Experiment 2 - Is there a difference in the clustering results depending on which part of a connection is being selected?

The second experiment is carried out in order to determine whether changing the position from where to take the packets has an influence on the resulting grouping. To that end, varying amounts of packets are being selected from three positions, once from the start, middle and end. Each of these three sequences are used as a distinct input into Malpaca and the resulting clusters are than compared to see if there are substantial differences.

When it comes to the question whether sequencing different segments of a connection captures different behavior, there is enough evidence to suggest that this indeed is the case. As seen on Figure 5, a custom made transition graph, in this representative example from the "10000" dataset using a fixed threshold of five, the same connection is being grouped into different clusters when different segments of the same connection are being analyzed.

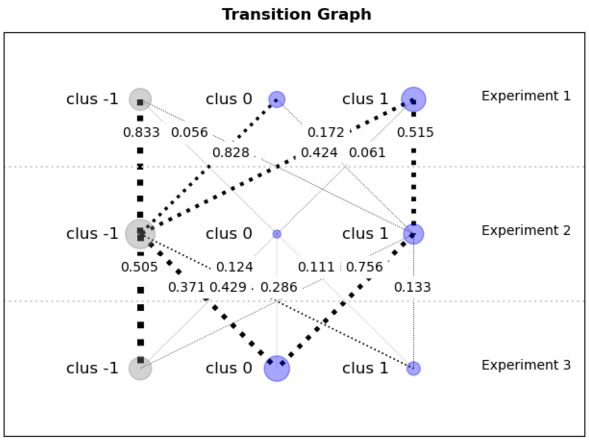


Figure 5: 10000 Dataset - Transition Graph

The focus of the interpretation of the transition graph should not be on how large a particular cluster is in different experiments but rather on the name-sake transition of how connection assignment to a cluster shifts between experiment. In Figure 5, for instance between Experiment 2 and 3, over 3/4s of the connections in cluster 1 are being assigned to cluster 0 while only roughly 13% stay in cluster 0.

However, not only the cluster assignment differs between experiments, the cluster makeup varies as well.

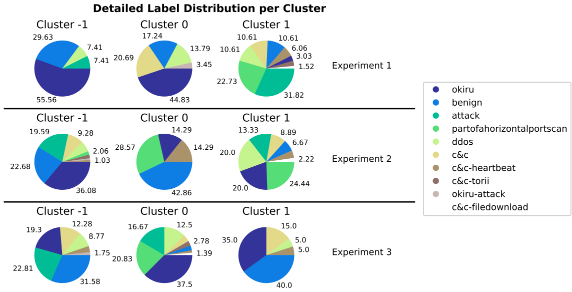


Figure 6: 10000 Dataset - 5 Threshold - Detailed Labels per Cluster per Experiment

As can be seen in Figure 6, there are substantial differences in the make-up of each cluster in the different experiments. Cluster 0 in experiment 1 for instance is dominated by the "okiru" behavior while in experiment 2, the "benign" behavior is most wide spread.

Experiment 3 - What is the effect of taking packets from the end of a connection and of skipping some packets ?

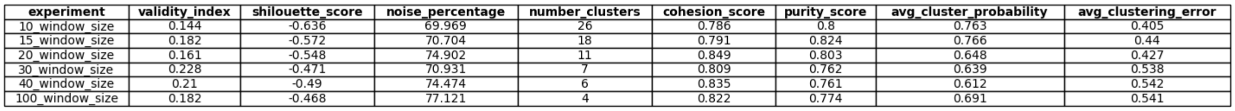
In contrast to the first study, in experiment three both the sequence length and the position from which packets are selected are altered. In particular, two new setups are being tested: taking packets from the end of a connection as well as skipping a certain amount of and then taking a fixed amount of packets. The intuition behind these experiment layouts is that perhaps the behavior of a connection can best be established not from the first but from the last packets send. The threshold values chosen for experiment 4 were the same as for experiment 1 in order to facilitate easy comparison. The skip amounts chosen are 5 and 10 as the previous experiments have indicated that already many connections are shorter than that.

Due to the combination of all possible threshold values with all possible skip values, too many experiments have been conducted in order to depict the results here. However, in Appendix 4 - "Experiment Results", one can find all outcomes. In short, no experiment has performed better than the fixed threshold value 8 found in experiment 1. Additionally, even after closer inspection, the benefits from analyzing a connection from the end or skipping certain packets are not clear.

Experiment 4 - What is the effect of breaking up one connection into multiple smaller connections of equal length ?

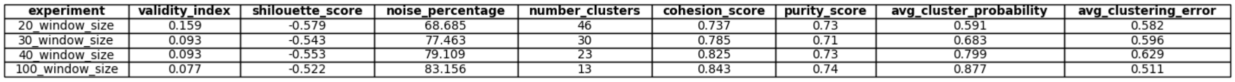
One possible solution to the issue of some data sets not containing enough connections for MalPaCa to create valid clusters could be to instead split one connection into a number of sub-connections of equal length. This also has the added benefit of covering the behavior of a connection throughout its entire lifetime and not just at one particular moment in time. The threshold values chosen for experiment 4 were the same as for experiment 1 in order to facilitate easy comparison.

The first thing to note for these experiments is that for many of the selected values, the MalPaCa Algorithm could not actually create a successful clustering as there simply were too many connections to be compared for the computing hardware that was available. For



experiment	validity_index	shilouette_score	noise_percentage	number_clusters	cohesion_score	purity_score	avg_cluster_probability	avg_clustering_error
10_window_size	0.144	-0.636	69.969	26	0.786	0.8	0.763	0.405
15_window_size	0.182	-0.572	70.704	18	0.791	0.824	0.766	0.44
20_window_size	0.161	-0.548	74.902	11	0.849	0.803	0.648	0.427
30_window_size	0.228	-0.471	70.931	7	0.809	0.762	0.639	0.538
40_window_size	0.21	-0.49	74.474	6	0.835	0.761	0.612	0.542
100_window_size	0.182	-0.468	77.121	4	0.822	0.774	0.691	0.541

Figure 7: 10000 Dataset - Experiment 4 - Results



experiment	validity_index	shilouette_score	noise_percentage	number_clusters	cohesion_score	purity_score	avg_cluster_probability	avg_clustering_error
20_window_size	0.159	-0.579	68.685	46	0.737	0.73	0.591	0.582
30_window_size	0.093	-0.543	77.463	30	0.785	0.71	0.683	0.596
40_window_size	0.093	-0.553	79.109	23	0.825	0.73	0.799	0.629
100_window_size	0.077	-0.522	83.156	13	0.843	0.74	0.877	0.511

Figure 8: Min 20 Dataset - Experiment 4 - Results

the values that could be computed, the overall performance was worse compared to that of the previously identified best threshold of 8. However, for neither the "10000" nor the "Min 20" dataset could experiment 4 actually be conducted for this value. Thus, it is entirely possible that with this value, the windowed approach would actually fare best.

Experiment 5 - What is the effect of defining behavior according to Netflow 5?

All of experiments 1 to 4 were also conducted once again on both the "10000" and the "Min 20" dataset, however this time using the "Netflow v5" definition of a connection.

The precise results can be found once again in Appendix 4 - "Experiment Results". In short, no systematic differences in terms of clustering results can be observed when behavior is defined along the lines of "Netflow v5". This is likely due to the fact that both the current MalPaCa definition and "Netflow v5s" definition are quite similar. Both are unidirectional and both take into account the source and destination IP address. As "Netflow v5" posses additional requirements based on which a connection is defined, there are more and shorter "Netflow v5" connections. This, however, seemingly has no effects on the clustering outcomes of MalPaCa. Therefore, no immediate performance boost would be obtained by changing the behavior definition to that of "Netflow v5". However, MalPaCa would then follow a more universally accepted definition of behavior, thereby making its results more comparable.

5 Responsible Research

As stated previously, both the "IoT-23 Dataset" [1] and the MalPaCa source code with all the modifications carried out in the context of this scientific article [2] are publicly available. The conditions needed in order to obtain the filtered dataset as well as the precise ratios used to create the balanced datasets are described in chapter 2 as well as in the appendix. If desired, the NFStream library [3] used for deep-packet inspection is also freely available. Therefore, the results of the previously discussed experiments can easily be recreated. And as the HDBScan clustering algorithm used by MalPaCa was shown to be deterministic, the groupings should be the same given that the identical data set is used.

Additionally, wherever possible, attempts where made to make the findings comparable to that of the original paper. For instance, by re-using the "clustering error" metric, the results can be put into context of the initial findings. Some of the tested threshold values were being taken from related research that was carried out using MalPaCa so that the findings could be compared (Roeling et al., 2020).

A great deal of effort was being put into ensuring that the datasets used do not suffer from data manipulation in any form. For this reason, the experiments were carried out on both the "10000" and the "Min 20" balanced dataset so that a clear view is being obtained as to how the results look like in more favorable and in more realistic conditions.

As the "IoT-23 Dataset" is based on traffic simulated by the Stratosphere Laboratory of the czech CTU University, no sensitive personal information is being transmitted via the captured network traffic. Therefore, no need exists to anonymize or otherwise obscure information like the IP addresses.

6 Conclusions and Future Work

6.1 Limitations

The files of the "IoT-23 Dataset" containing the label type of a connection did not include any information regarding two of the features based on which a netflow connection can be determined namely "IP Type of Service" and "Router or switch interface" (InterProjektWiki, 2021). Therefore, the label attribution had to be done solely on the basis of the available features. It could be now, that without this information, the detailed label association is not correct. However, a cursory examination of the "IP Type of Service" values show that there is not a lot of variation among them, indicating that this is unlikely the case. Nonetheless, recreating the experiment of this article would further strengthen its findings.

As stated previously, some of the planned experiments, particularly for the smaller threshold values and for the windowed approach could not be executed simply due to the fact that the computing hardware available could not handle the memory requirements. It is now possible that in those instances, the clustering algorithm would perform much better.

6.2 Future Work

On a practical level, the process used to calculate the "clustering error" should further be improved upon. In particular, a thorough investigation is necessary to understand what

[1] https://www.stratosphereips.org/datasets-iot23

[2] https://github.com/mrjojo11/malpaca-pub

[3] https://github.com/nfstream/nfstream

exactly causes the algorithm to fail when facing larger data sets. Additionally, instead of requiring an exact match in terms of feature value in order to be classified as clustered correctly, correctness should be based on a range of acceptable values.

More experimentation with splitting up a connection into a number of windows is necessary. As it stands, each window is treated as an individual behavior and no upper limit is set on how many windows can be created from one connection. As stated previously, this decision was made to capture the entire behavior of a connection and to prevent malware obfuscation techniques such as adding a random delay or sending unnecessary packets to mask the actual behavior. However, as seen, this approach can be extremely resource intensive to a point where multiple experiments could not be completed due to lack of memory. One obvious change would be to put a limit on how many windows can be extracted from one connection. However, then again one runs into the problem of potentially not capturing behavior that occurs towards the end of a connection. A preprocessing step thus could be explored, where longer connections are examined one the basis of the same features that MalPaCa uses in order to identify segments of the connection that are most dissimilar. These segments then could be then be used for the windows.

Lastly, the network traffic captured in the "IoT-23 Dataset" is quite bifurcated in that a connection is either very large or very small as Figure 33 found in the appendix reveals. Additionally, the behavior distribution is skewed in that four of the 12 types of behavior account for over 90% of all observed behavior. Testing MalPaCa on another, more balanced dataset would be an important step towards fully understanding its clustering behavior.

6.3 Conclusion

The current default sequencing length value of 20 is definitely too high and should in any case be reduced to 10 or 8. Threshold values larger than 20 did not result in valid clusterings which is largely due to the fact that the vast majority of behavior in the "IoT-23 Dataset" results in connections that are much shorter than 20. The window approach where one connection is repeatedly sliced into shorter, equally sliced windows has shown great promise for dealing with heterogeneous connections. As the "IoT-23 Dataset" has shown, frequently behavior can lead to connections that are either very long or very short. A static threshold of any value can no do this justice as it runs into the danger of not picking up important behavior or cutting short connections representing longer chains of action. Using a windowed approach can counteract both issues at least partially. By setting the window size small, shorter behavior can be picked up and by repeatedly extracting this small window, even behavior spread out over long connections can be included. Hardware limitations however have prevented many of the experiments using this approach from being completed, particularly for small threshold values. Therefore, this indicates that one potential downside of this approach could be its resource intensive nature. For research is necessary to investigate both how MalPaCa performs in these circumstances as well as whether these processing problems could not be mitigated. Lastly, changing MalPaCas definition of behavior to the more common "Netflow v5" does not lead to performance improvements, however MalPaCas results would become more comparable.

References

Bailey, M., Oberheide, J., Andersen, J., Mao, Z. M., Jahanian, F., & Nazario, J. (2007). Automated classification and analysis of Internet malware. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 4637 LNCS, pp. 178–197). Springer Verlag. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-540-74320-0_10` doi: 10.1007/978-3-540-74320-0_10

Bayer, U., Comparetti, P., Hlauschek, C., Kruegel, C., & Kirda, E. (2009). Scalable, behavior-based malware clustering. In *Network and distributed system security symposium (ndss).* Retrieved from `https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.7690http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.7690&rep=rep1&type=pdf`

Cai, T., international conference on Wireless, F. Z. . t., & undefined 2012. (n.d.). Detecting HTTP botnet with clustering network traffic. *ieeexplore.ieee.org*. Retrieved from `https://ieeexplore.ieee.org/abstract/document/6478491/?casa_token=FzbaQWbfU7kAAAAA:l6wqhy4gti1CiP7zv9SHAbV0mEdqDEcWCSmnNuf_pFI6eGhF0CivgbFVbhjPb8XFIuAwG5mPVg`

Commtouch. (2007). *Inc. Malware outbreak trend report: Bagle/beagle.* Retrieved from `http://www.commtouch.com/documents/Bagle-Worm_MOTR.pdf%0A`

Grill, M., Nikolaev, I., Valeros, V., & Rehak, M. (2015, jun). Detecting DGA malware using NetFlow. In *Proceedings of the 2015 ifip/ieee international symposium on integrated network management, im 2015* (pp. 1304–1309). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/INM.2015.7140486

Gu, G., Zhang, J., & Lee, W. (2008, feb). BotSniffer : Detecting Botnet Command and Control Channels in Network Traffic. *Proceedings of the 15th Annual Network and Distributed System Security Symposium.*, *53*(1), 1–13. Retrieved from `https://corescholar.libraries.wright.edu/cse/7http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.8092&rep=rep1&type=pdf`

IETF. (2021). *IP flow information export (ipfix).* Retrieved 2021-06-27, from `https://tools.ietf.org/html/rfc7011`

InterProjektWiki. (2021). *NetFlow.* Retrieved 2021-06-27, from `https://pliki.ip-sa.pl/wiki/Wiki.jsp?page=NetFlow`

Kheir, N., & Wolley, C. (2013, nov). BotSuer: Suing stealthy P2P bots in network traffic through netflow analysis. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 8257 LNCS, pp. 162–178). Springer Verlag. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-319-02937-5_9` doi: 10.1007/978-3-319-02937-5_9

Li, P., Liu, L., Gao, D., & Reiter, M. K. (2010). On challenges in evaluating malware clustering. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 6307 LNCS, pp. 238–255). Springer Verlag. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-642-15512-3_13` doi: 10.1007/978-3-642-15512-3_13

Nadeem, A., Hammerschmidt, C., Gañán, C. H., & Verwer, S. (2021). Beyond Labeling: Using Clustering to Build Network Behavioral Profiles of Malware Families. In *Malware analysis using artificial intelligence and deep learning* (pp. 381–409). Springer International Publishing. Retrieved from `https://link-springer-com.eur.idm.oclc.org/chapter/10.1007/978-3-030-62582-5_15` doi: 10.1007/978-3-030-62582-5_15

Parmisano, A., Garcia, S., & Erquiaga, M. J. (2020, jan). *Stratosphere Laboratory. A labeled dataset with malicious and benign IoT network traffic.* Retrieved 2021-06-16, from `https://zenodo.org/record/4743746https://www.stratosphereips.org/datasets-iot23` doi: 10.5281/ZENODO.4743746

Petryschuk, S. (2019). *What Is NetFlow? How It Works, Why to Use It, and Examples.* Retrieved 2021-06-27, from `https://www.auvik.com/franklyit/blog/netflow-basics/`

Roeling, M. P., Nadeem, A., & Verwer, S. (2020, sep). Hybrid Connection and Host Clustering for Community Detection in Spatial-Temporal Network Data. In *Communications in computer and information science* (Vol. 1323, pp. 178–204). Springer Science and Business Media Deutschland GmbH. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-030-65965-3_12` doi: 10.1007/978-3-030-65965-3_12

Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., ... Hakimian, P. (2011). Detecting P2P botnets through network behavior analysis and machine learning. In *2011 9th annual international conference on privacy, security and trust, pst 2011* (pp. 174–180). doi: 10.1109/PST.2011.5971980

Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2020, nov). NetFlow Datasets for Machine Learning-based Network Intrusion Detection Systems. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *371 LNICST*, 117–135. Retrieved from `http://arxiv.org/abs/2011.09144http://dx.doi.org/10.1007/978-3-030-72802-1_9` doi: 10.1007/978-3-030-72802-1_9

Sebastián, M., Rivera, R., Kotzias, P., & Caballero, J. (2016). Avclass: A tool for massive malware labeling. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9854 LNCS, pp. 230–253). Springer Verlag. doi: 10.1007/978-3-319-45719-2_11

Strayer, W. T., Lapsely, D., Walsh, R., & Livadas, C. (2008). Botnet detection based on network behavior. *Advances in Information Security*, *36*, 1–24. Retrieved from `https://link.springer.com/chapter/10.1007/978-0-387-68768-1_1` doi: 10.1007/978-0-387-68768-1_1

7 Appendices

7.1 Metric Calculation

Variables:
mfl = most frequent label in cluster
mfdl = most frequent detailed label in cluster
mfan = most frequent application name in cluster
mfacn = most frequent application category name in cluster
mfn = most frequent name in cluster

lalc = largest amount of one label present in one cluster
tsl = total amount one label present
nl = number of distinct labels
ladlc = largest amount of one detailed label present in one cluster
tsdl = total amount one detailed label present
ndl = number of distinct detailed labels
laanc = largest amount of one application name present in one cluster
tsan = total amount one application name present
nan = number of distinct application names
laacnc = largest amount of one application category name present in one cluster
tsacn = total amount one application category name present
nacn = number of distinct application category names
lanc = largest amount of one name present in one cluster
tsn = total amount one name present
nn = number of distinct names

tcc = total number of connections in cluster
tc = total number of connections
c = number of clusters

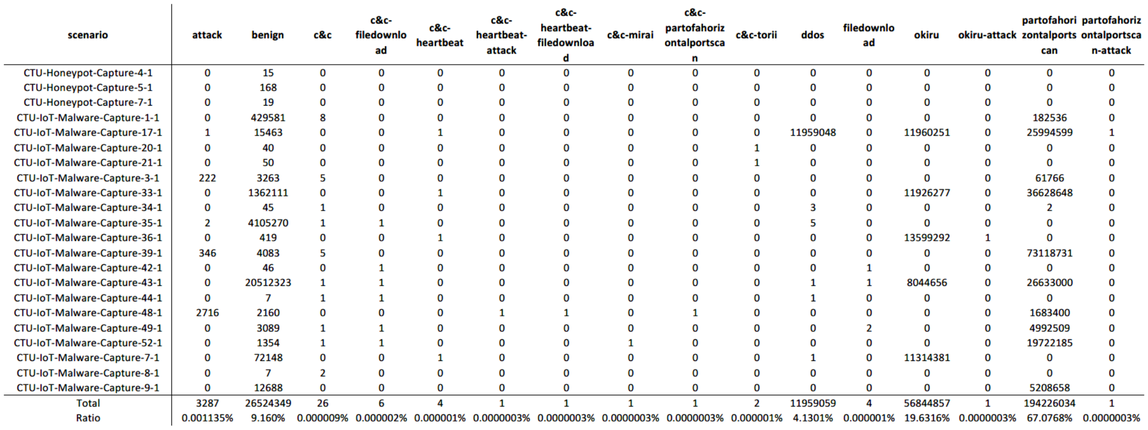
Formulas:
avg_label_purity = $\frac{\sum \frac{mfl}{tcc}}{c}$
avg_detailed_label_purity = $\frac{\sum \frac{mfdl}{tcc}}{c}$
avg_application_name_purity = $\frac{\sum \frac{mfan}{tcc}}{c}$
avg_application_category_name_purity = $\frac{\sum \frac{mfacn}{tcc}}{c}$
avg_name_purity = $\frac{\sum \frac{mfn}{tcc}}{c}$

avg_label_cohesion = $\frac{\sum_{n=1}^{\#labels} \frac{lalc}{tsl}}{nl}$
avg_detailed_label_cohesion = $\frac{\sum_{n=1}^{\#detailed\_labels} \frac{ladlc}{tsdl}}{ndl}$
avg_application_name_cohesion = $\frac{\sum_{n=1}^{\#application\_name} \frac{laanc}{tsan}}{nan}$
avg_application_category_name_cohesion = $\frac{\sum_{n=1}^{\#application\_category\_names} \frac{laacnc}{tsacn}}{nacn}$
avg_name_cohesion = $\frac{\sum_{n=1}^{\#name} \frac{lanc}{tsn}}{nn}$

$\text{purity\_score} = 0.35 * avg\_overall\_label\_purity + 0.45 * avg\_overall\_detailed\_label\_purity + 0.05 * avg\_overall\_application\_name\_purity + 0.05 * avg\_overall\_application\_category\_name\_purity + 0.1 * avg\_overall\_name\_purity$
$\text{cohesion\_score} = 0.35 * avg\_label\_cohesion + 0.45 * avg\_detailed\_label\_cohesion + 0.05 * avg\_application\_name\_cohesion + 0.05 * avg\_application\_category\_name\_cohesion + 0.1 * avg\_name\_cohesion$

7.2 Balanced Dataset Creation Example

This example recaptures how for the unidirectional behavior, the "10000" balanced dataset was created.
Step 1: Calculate the detailed label distribution of the original dataset based on the desired definition of behavior

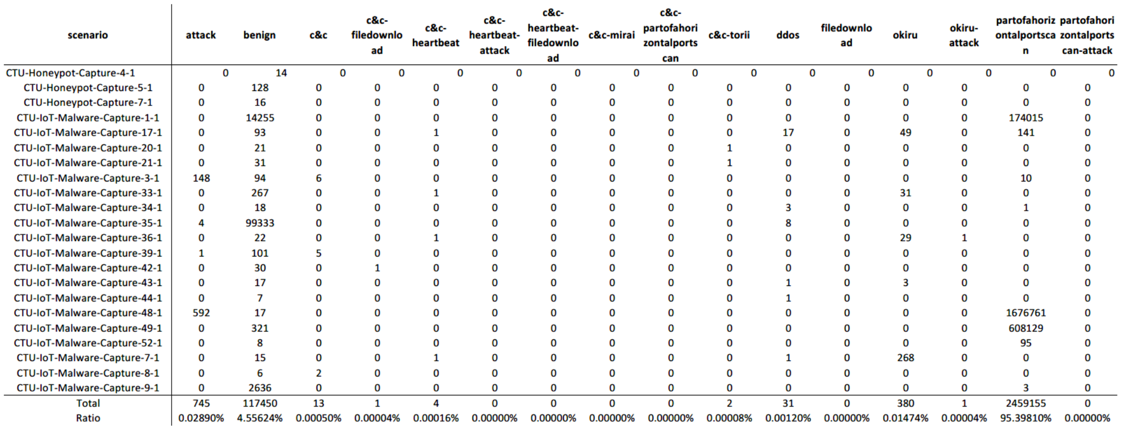


scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	168	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	429581	8	0	0	0	0	0	0	0	0	0	0	0	182536	0
CTU-IoT-Malware-Capture-17-1	1	15463	0	0	1	0	0	0	0	0	11959048	0	11960251	0	25994599	1
CTU-IoT-Malware-Capture-20-1	0	40	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	50	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	222	3263	5	0	0	0	0	0	0	0	0	0	0	0	61766	0
CTU-IoT-Malware-Capture-33-1	0	1362111	0	0	1	0	0	0	0	0	0	0	11926277	0	36628648	0
CTU-IoT-Malware-Capture-34-1	0	45	1	0	0	0	0	0	0	0	3	0	0	0	2	0
CTU-IoT-Malware-Capture-35-1	2	4105270	1	1	0	0	0	0	0	0	5	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	419	0	0	1	0	0	0	0	0	0	0	13599292	1	0	0
CTU-IoT-Malware-Capture-39-1	346	4083	5	0	0	0	0	0	0	0	0	0	0	0	73118731	0
CTU-IoT-Malware-Capture-42-1	0	46	0	1	0	0	0	0	0	0	0	1	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	20512323	1	1	0	0	0	0	0	0	1	1	8044656	0	26633000	0
CTU-IoT-Malware-Capture-44-1	0	7	1	1	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	2716	2160	0	0	0	1	1	0	1	0	0	0	0	0	1683400	0
CTU-IoT-Malware-Capture-49-1	0	3089	1	1	0	0	0	0	0	0	0	2	0	0	4992509	0
CTU-IoT-Malware-Capture-52-1	0	1354	1	1	0	0	0	1	0	0	0	0	0	0	19722185	0
CTU-IoT-Malware-Capture-7-1	0	72148	0	0	1	0	0	0	0	0	1	0	11314381	0	0	0
CTU-IoT-Malware-Capture-8-1	0	7	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	12688	0	0	0	0	0	0	0	0	0	0	0	0	5208658	0
Total	3287	26524349	26	6	4	1	1	1	1	2	11959059	4	56844857	1	194226034	1
Ratio	0.001135%	9.160%	0.000009%	0.000002%	0.000001%	0.0000003%	0.0000003%	0.0000003%	0.0000003%	0.000001%	4.1301%	0.000001%	19.6316%	0.0000003%	67.0768%	0.0000003%

Figure 9: Detailed Label Distribution - Unidirectional Behavior

In this example, the used definition of behavior is the original MalPaCa definition of the "Unidirectional Behavior". Based on this level of aggregation, it is assessed per scenario how many connections of each detailed label are present. Then, the percentage of each detailed label being found in the overall "IoT 23" dataset is calculated.

Step 2: Calculate the detailed label distribution of the filtered dataset

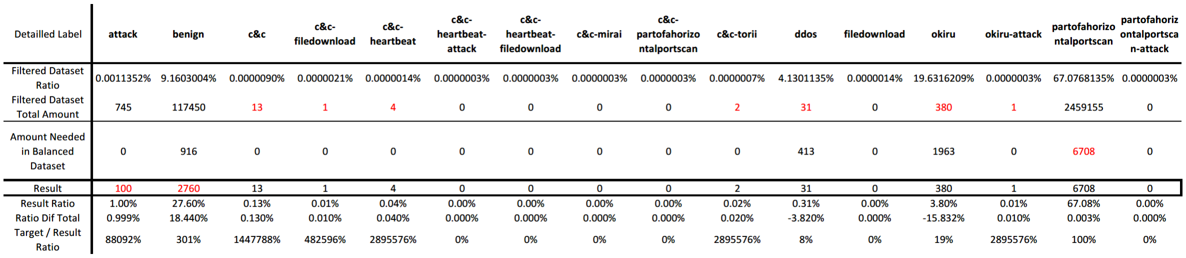


scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	14255	0	0	0	0	0	0	0	0	0	0	0	0	174015	0
CTU-IoT-Malware-Capture-17-1	0	93	0	0	1	0	0	0	0	0	17	0	49	0	141	0
CTU-IoT-Malware-Capture-20-1	0	21	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	31	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	148	94	6	0	0	0	0	0	0	0	0	0	0	0	10	0
CTU-IoT-Malware-Capture-33-1	0	267	0	0	1	0	0	0	0	0	0	0	31	0	0	0
CTU-IoT-Malware-Capture-34-1	0	18	0	0	0	0	0	0	0	0	3	0	0	0	1	0
CTU-IoT-Malware-Capture-35-1	4	99333	0	0	0	0	0	0	0	0	8	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	22	0	0	1	0	0	0	0	0	0	0	29	1	0	0
CTU-IoT-Malware-Capture-39-1	1	101	5	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	30	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	17	0	0	0	0	0	0	0	0	1	0	3	0	0	0
CTU-IoT-Malware-Capture-44-1	0	7	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	592	17	0	0	0	0	0	0	0	0	0	0	0	0	1676761	0
CTU-IoT-Malware-Capture-49-1	0	321	0	0	0	0	0	0	0	0	0	0	0	0	608129	0
CTU-IoT-Malware-Capture-52-1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	95	0
CTU-IoT-Malware-Capture-7-1	0	15	0	0	1	0	0	0	0	0	1	0	268	0	0	0
CTU-IoT-Malware-Capture-8-1	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	2636	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Total	745	117450	13	1	4	0	0	0	0	2	31	0	380	1	2459155	0
Ratio	0.02890%	4.55624%	0.00050%	0.00004%	0.00016%	0.00000%	0.00000%	0.00000%	0.00000%	0.00008%	0.00120%	0.00000%	0.01474%	0.00004%	95.39810%	0.00000%

Figure 10: Detailed Label Distribution - Filtered Dataset - Unidirectional Behavior

Step 1 now is repeated, but this time on the filtered dataset where all connections shorter than 5 are being excluded and connections longer than 1000 packets are being cut off. This step is necessary to make the data set usable for the computation equipment available.

Step 3: Match up the filtered dataset with the original ratios

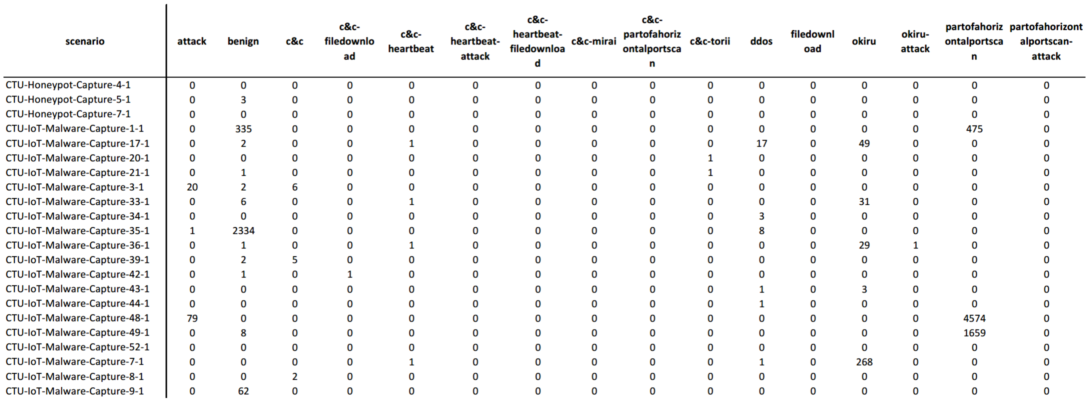


Detailled Label	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
Filtered Dataset Ratio	0.0011352%	9.1603004%	0.0000090%	0.0000021%	0.0000014%	0.0000003%	0.0000003%	0.0000003%	0.0000003%	0.0000007%	4.1301135%	0.0000014%	19.6316209%	0.0000003%	67.0768135%	0.0000003%
Filtered Dataset Total Amount	745	117450	13	1	4	0	0	0	0	2	31	0	380	1	2459155	0
Amount Needed in Balanced Dataset	0	916	0	0	0	0	0	0	0	0	413	0	1963	0	6708	0
Result	100	2760	13	1	4	0	0	0	0	2	31	0	380	1	6708	0
Result Ratio	1.00%	27.60%	0.13%	0.01%	0.04%	0.00%	0.00%	0.00%	0.00%	0.02%	0.31%	0.00%	3.80%	0.01%	67.08%	0.00%
Ratio Dif Total	0.999%	18.440%	0.130%	0.010%	0.040%	0.000%	0.000%	0.000%	0.000%	0.020%	-3.820%	0.000%	-15.832%	0.010%	0.003%	0.000%
Target / Result Ratio	88092%	301%	1447788%	482596%	2895576%	0%	0%	0%	0%	2895576%	8%	0%	19%	2895576%	100%	0%

Figure 11: Matching Filtered Dataset with Original Dataset

In this step, it is determined how many of each detailed labels should be present in the balanced dataset. To that end, the first step is to attempt to create a subset of filtered dataset based on the original ratios with which each detailed label was present in the "IoT 23" dataset. Firstly, a targeted data size is established, which in this example was 10000, based on which then the needed amount of each detailed label is determined by calculating the original ratio with the selected total data size. In many cases now, it is not possible to actually select as connections of a detailed label as required as the previously conducted filtering step fundamentally changed the detailed label distribution. To solve this, first for those detailed labels where less connections are present in the filtered data set than needed, such as in the case of the "ddos" label, simply as many connections as exist are selected. Then, in order to still arrive at the desired data set size, first all of those detailed labels are picked out, which are so rare that they would not show up in the balanced dataset such as the "okiru-attack" or the "c&c" detailed labels. Lastly, from those detailed labels where more exist than needed, such as "attack" or "benign", as many connections are selected as needed to reach the desired dataset size. The "benign" detailed label was favored as this behavior is most likely to be highly represented in real-world traffic. The end result of this matching process can be observed in Figure 11

Step 4: Extract from the filtered dataset connections based on the matched detailed label distribution



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	335	0	0	0	0	0	0	0	0	0	0	0	0	475	0
CTU-IoT-Malware-Capture-17-1	0	2	0	0	1	0	0	0	0	0	17	0	49	0	0	0
CTU-IoT-Malware-Capture-20-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	20	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-33-1	0	6	0	0	1	0	0	0	0	0	0	0	31	0	0	0
CTU-IoT-Malware-Capture-34-1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
CTU-IoT-Malware-Capture-35-1	1	2334	0	0	0	0	0	0	0	0	8	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	1	0	0	1	0	0	0	0	0	0	0	29	1	0	0
CTU-IoT-Malware-Capture-39-1	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0
CTU-IoT-Malware-Capture-44-1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	79	0	0	0	0	0	0	0	0	0	0	0	0	0	4574	0
CTU-IoT-Malware-Capture-49-1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	1659	0
CTU-IoT-Malware-Capture-52-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-7-1	0	0	0	0	1	0	0	0	0	0	1	0	268	0	0	0
CTU-IoT-Malware-Capture-8-1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 12: Extracting Table - 10000 Dataset

The detailed label distribution created in step 3 is then matched up with the detailed label distribution from the filtered dataset, seen in Figure 10. To do this, the ratio from the filtered dataset with which each detailed label is distributed over the different scenarios is used in order to make sure that the required amount of connections can actually be extracted from one scenario. As seen in Figure 12, this means that for instance from scenario "CTU-IoT-Malware-Capture-1-1", 335 connections with the "benign" detailed label have to be selected for the balanced dataset. These 335 connections are randomly chosen from the 14625 "benign" connections in "CTU-IoT-Malware-Capture-1-1" in order to ensure a non-biased selection.

Creating other types of balanced datasets

In order to create the "Min 20" balanced dataset, all that is necessary to change is in step 2, only those connections should be taken into account when determining the detailed label distribution of the filtered dataset that are longer than at least 20 packets. Step 4 also needs to be changed, so that instead of randomly selecting a connection from the total pool of connections of a particular detailed label in a particular scenario, now the connection should be randomly selected from the total pool of connections of a particular detailed label in a particular scenario that are longer than 20.

In order to create a balanced dataset for the "netflow" behavior definition, instead of using the "unidirectional" definition of a connection when determining the detailed label distribution, use the "netflow" definition of a connection. All other tasks can remain the same.

7.3 Clustering Error Algorithm

Example Output

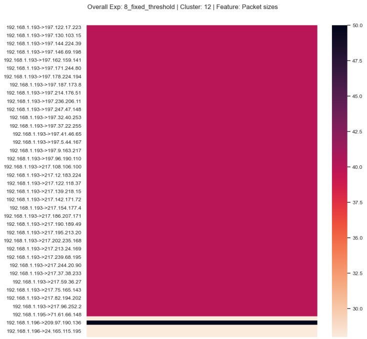


Figure 13: Overall heatmap for the bytes feature

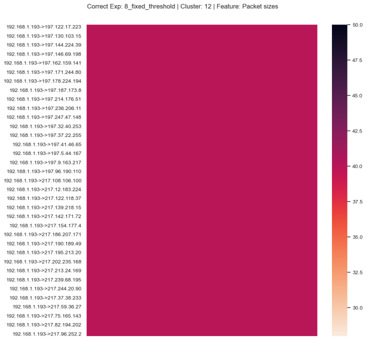


Figure 14: Heatmap representing what the algorithm has deemed as correctly clustered

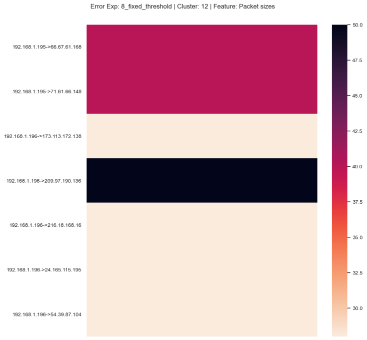
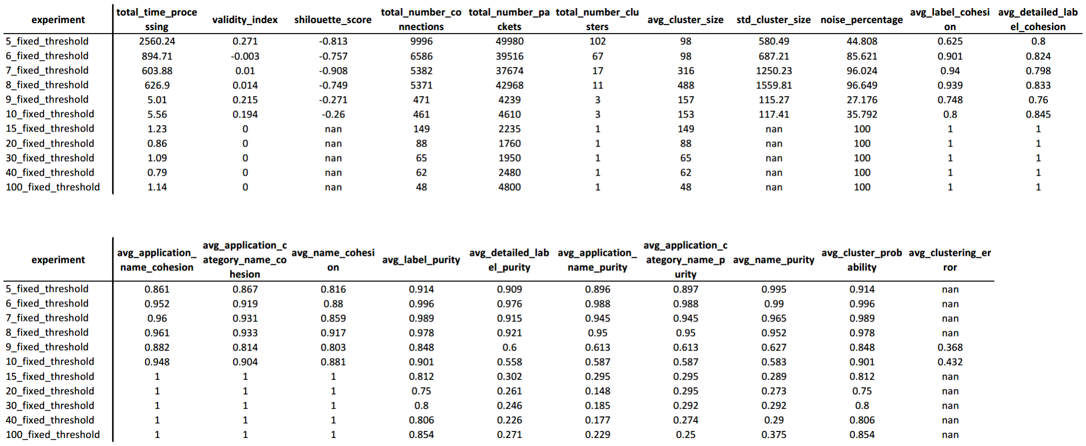


Figure 15: Heatmap representing what the algorithm has deemed as incorrectly clustered

7.4 Experiment Results

Unidirectional Behavior

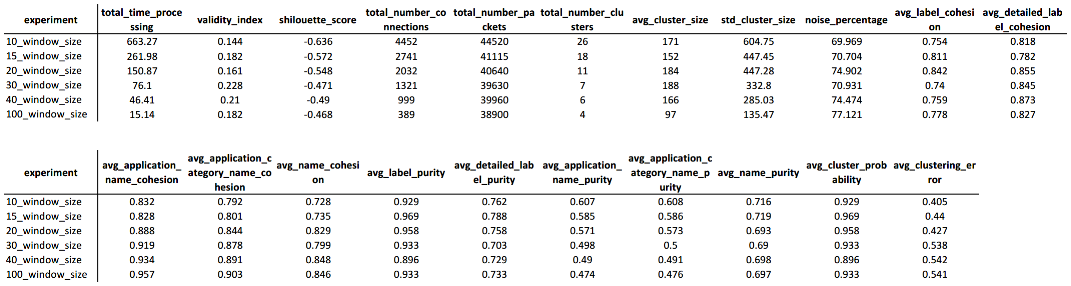
10000 Dataset



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
5_fixed_threshold	2560.24	0.271	-0.813	9996	49980	102	98	580.49	44.808	0.625	0.8
6_fixed_threshold	894.71	-0.003	-0.757	6586	39516	67	98	687.21	85.621	0.901	0.824
7_fixed_threshold	603.88	0.01	-0.908	5382	37674	17	316	1250.23	96.024	0.94	0.798
8_fixed_threshold	626.9	0.014	-0.749	5371	42968	11	488	1559.81	96.649	0.939	0.833
9_fixed_threshold	5.01	0.215	-0.271	471	4239	3	157	115.27	27.176	0.748	0.76
10_fixed_threshold	5.56	0.194	-0.26	461	4610	3	153	117.41	35.792	0.8	0.845
15_fixed_threshold	1.23	0	nan	149	2235	1	149	nan	100	1	1
20_fixed_threshold	0.86	0	nan	88	1760	1	88	nan	100	1	1
30_fixed_threshold	1.09	0	nan	65	1950	1	65	nan	100	1	1
40_fixed_threshold	0.79	0	nan	62	2480	1	62	nan	100	1	1
100_fixed_threshold	1.14	0	nan	48	4800	1	48	nan	100	1	1

experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold	0.861	0.867	0.816	0.914	0.909	0.896	0.897	0.995	0.914	nan
6_fixed_threshold	0.952	0.919	0.88	0.996	0.976	0.988	0.988	0.99	0.996	nan
7_fixed_threshold	0.96	0.931	0.859	0.989	0.915	0.945	0.945	0.965	0.989	nan
8_fixed_threshold	0.961	0.933	0.917	0.978	0.921	0.95	0.95	0.952	0.978	nan
9_fixed_threshold	0.882	0.814	0.803	0.848	0.6	0.613	0.613	0.627	0.848	0.368
10_fixed_threshold	0.948	0.904	0.881	0.901	0.558	0.587	0.587	0.583	0.901	0.432
15_fixed_threshold	1	1	1	0.812	0.302	0.295	0.295	0.289	0.812	nan
20_fixed_threshold	1	1	1	0.75	0.261	0.148	0.295	0.273	0.75	nan
30_fixed_threshold	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
40_fixed_threshold	1	1	1	0.806	0.226	0.177	0.274	0.29	0.806	nan
100_fixed_threshold	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan

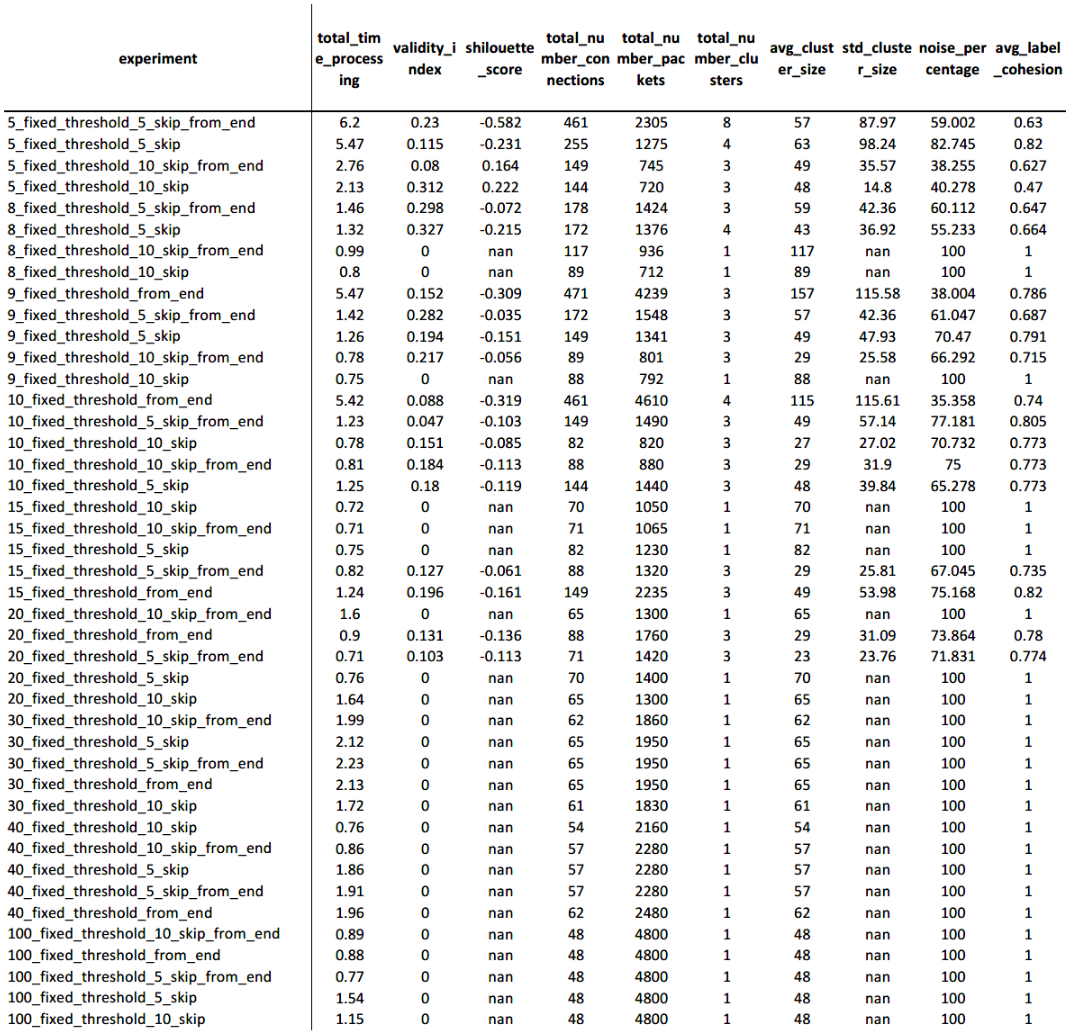
Figure 16: 10000 Dataset - Experiment 1 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
10_window_size	663.27	0.144	-0.636	4452	44520	26	171	604.75	69.969	0.754	0.818
15_window_size	261.98	0.182	-0.572	2741	41115	18	152	447.45	70.704	0.811	0.782
20_window_size	150.87	0.161	-0.548	2032	40640	11	184	447.28	74.902	0.842	0.855
30_window_size	76.1	0.228	-0.471	1321	39630	7	188	332.8	70.931	0.74	0.845
40_window_size	46.41	0.21	-0.49	999	39960	6	166	285.03	74.474	0.759	0.873
100_window_size	15.14	0.182	-0.468	389	38900	4	97	135.47	77.121	0.778	0.827

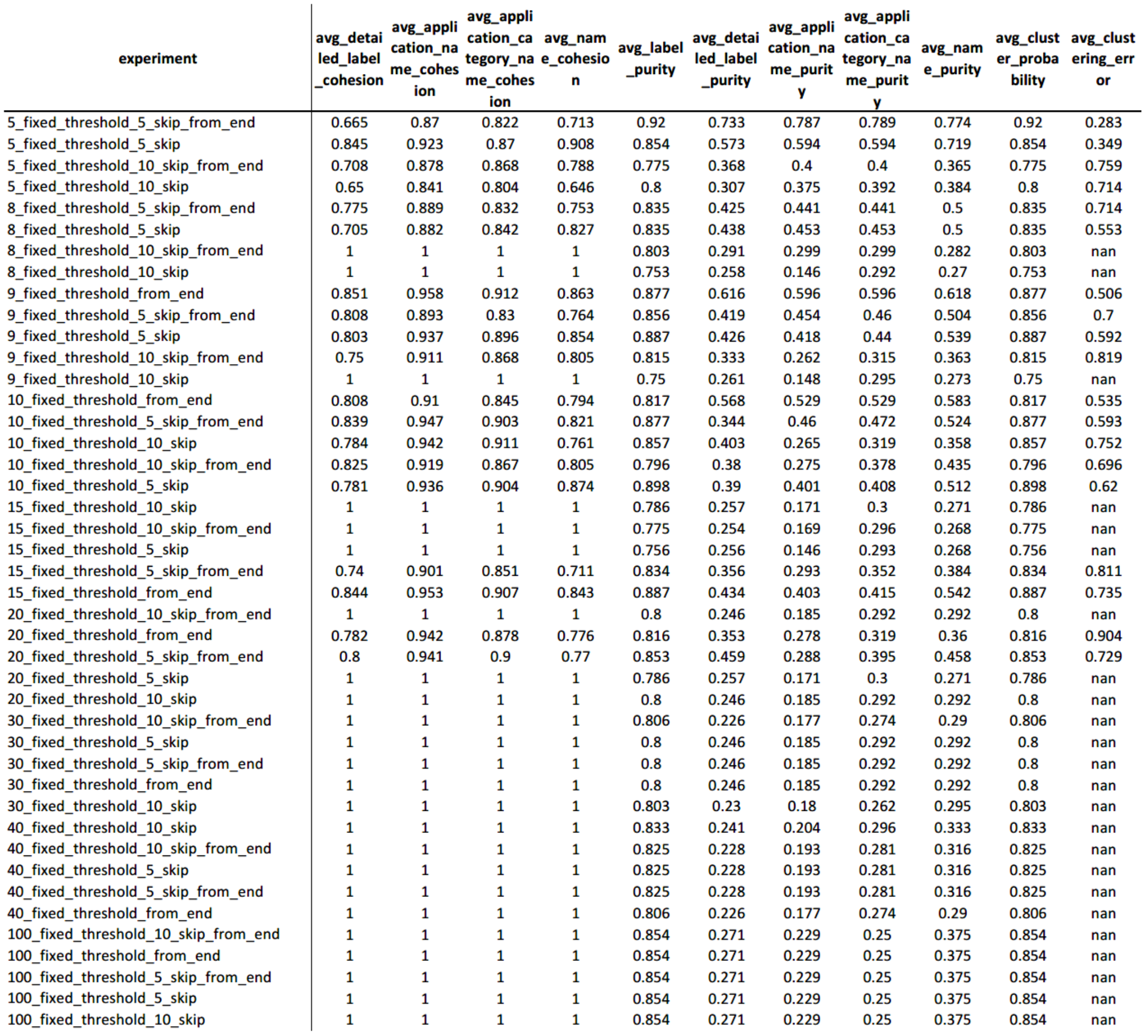
experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
10_window_size	0.832	0.792	0.728	0.929	0.762	0.607	0.608	0.716	0.929	0.405
15_window_size	0.828	0.801	0.735	0.969	0.788	0.585	0.586	0.719	0.969	0.44
20_window_size	0.888	0.844	0.829	0.958	0.758	0.571	0.573	0.693	0.958	0.427
30_window_size	0.919	0.878	0.799	0.933	0.703	0.498	0.5	0.69	0.933	0.538
40_window_size	0.934	0.891	0.848	0.896	0.729	0.49	0.491	0.698	0.896	0.542
100_window_size	0.957	0.903	0.846	0.933	0.733	0.474	0.476	0.697	0.933	0.541

Figure 17: 10000 Dataset - Experiment 4 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion
5_fixed_threshold_5_skip_from_end	6.2	0.23	-0.582	461	2305	8	57	87.97	59.002	0.63
5_fixed_threshold_5_skip	5.47	0.115	-0.231	255	1275	4	63	98.24	82.745	0.82
5_fixed_threshold_10_skip_from_end	2.76	0.08	0.164	149	745	3	49	35.57	38.255	0.627
5_fixed_threshold_10_skip	2.13	0.312	0.222	144	720	3	48	14.8	40.278	0.47
8_fixed_threshold_5_skip_from_end	1.46	0.298	-0.072	178	1424	3	59	42.36	60.112	0.647
8_fixed_threshold_5_skip	1.32	0.327	-0.215	172	1376	4	43	36.92	55.233	0.664
8_fixed_threshold_10_skip_from_end	0.99	0	nan	117	936	1	117	nan	100	1
8_fixed_threshold_10_skip	0.8	0	nan	89	712	1	89	nan	100	1
9_fixed_threshold_from_end	5.47	0.152	-0.309	471	4239	3	157	115.58	38.004	0.786
9_fixed_threshold_5_skip_from_end	1.42	0.282	-0.035	172	1548	3	57	42.36	61.047	0.687
9_fixed_threshold_5_skip	1.26	0.194	-0.151	149	1341	3	49	47.93	70.47	0.791
9_fixed_threshold_10_skip_from_end	0.78	0.217	-0.056	89	801	3	29	25.58	66.292	0.715
9_fixed_threshold_10_skip	0.75	0	nan	88	792	1	88	nan	100	1
10_fixed_threshold_from_end	5.42	0.088	-0.319	461	4610	4	115	115.61	35.358	0.74
10_fixed_threshold_5_skip_from_end	1.23	0.047	-0.103	149	1490	3	49	57.14	77.181	0.805
10_fixed_threshold_10_skip	0.78	0.151	-0.085	82	820	3	27	27.02	70.732	0.773
10_fixed_threshold_10_skip_from_end	0.81	0.184	-0.113	88	880	3	29	31.9	75	0.773
10_fixed_threshold_5_skip	1.25	0.18	-0.119	144	1440	3	48	39.84	65.278	0.773
15_fixed_threshold_10_skip	0.72	0	nan	70	1050	1	70	nan	100	1
15_fixed_threshold_10_skip_from_end	0.71	0	nan	71	1065	1	71	nan	100	1
15_fixed_threshold_5_skip	0.75	0	nan	82	1230	1	82	nan	100	1
15_fixed_threshold_5_skip_from_end	0.82	0.127	-0.061	88	1320	3	29	25.81	67.045	0.735
15_fixed_threshold_from_end	1.24	0.196	-0.161	149	2235	3	49	53.98	75.168	0.82
20_fixed_threshold_10_skip_from_end	1.6	0	nan	65	1300	1	65	nan	100	1
20_fixed_threshold_from_end	0.9	0.131	-0.136	88	1760	3	29	31.09	73.864	0.78
20_fixed_threshold_5_skip_from_end	0.71	0.103	-0.113	71	1420	3	23	23.76	71.831	0.774
20_fixed_threshold_5_skip	0.76	0	nan	70	1400	1	70	nan	100	1
20_fixed_threshold_10_skip	1.64	0	nan	65	1300	1	65	nan	100	1
30_fixed_threshold_10_skip_from_end	1.99	0	nan	62	1860	1	62	nan	100	1
30_fixed_threshold_5_skip	2.12	0	nan	65	1950	1	65	nan	100	1
30_fixed_threshold_5_skip_from_end	2.23	0	nan	65	1950	1	65	nan	100	1
30_fixed_threshold_from_end	2.13	0	nan	65	1950	1	65	nan	100	1
30_fixed_threshold_10_skip	1.72	0	nan	61	1830	1	61	nan	100	1
40_fixed_threshold_10_skip	0.76	0	nan	54	2160	1	54	nan	100	1
40_fixed_threshold_10_skip_from_end	0.86	0	nan	57	2280	1	57	nan	100	1
40_fixed_threshold_5_skip	1.86	0	nan	57	2280	1	57	nan	100	1
40_fixed_threshold_5_skip_from_end	1.91	0	nan	57	2280	1	57	nan	100	1
40_fixed_threshold_from_end	1.96	0	nan	62	2480	1	62	nan	100	1
100_fixed_threshold_10_skip_from_end	0.89	0	nan	48	4800	1	48	nan	100	1
100_fixed_threshold_from_end	0.88	0	nan	48	4800	1	48	nan	100	1
100_fixed_threshold_5_skip_from_end	0.77	0	nan	48	4800	1	48	nan	100	1
100_fixed_threshold_5_skip	1.54	0	nan	48	4800	1	48	nan	100	1
100_fixed_threshold_10_skip	1.15	0	nan	48	4800	1	48	nan	100	1

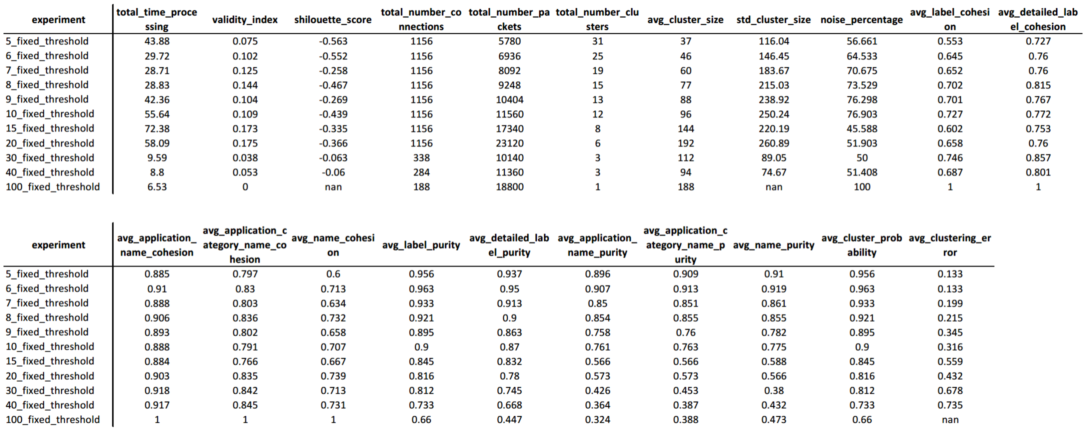
Figure 18: 10000 Dataset - Experiment 3 Results - 1



experiment	avg_detailed_label_cohesion	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold_5_skip_from_end	0.665	0.87	0.822	0.713	0.92	0.733	0.787	0.789	0.774	0.92	0.283
5_fixed_threshold_5_skip	0.845	0.923	0.87	0.908	0.854	0.573	0.594	0.594	0.719	0.854	0.349
5_fixed_threshold_10_skip_from_end	0.708	0.878	0.868	0.788	0.775	0.368	0.4	0.4	0.365	0.775	0.759
5_fixed_threshold_10_skip	0.65	0.841	0.804	0.646	0.8	0.307	0.375	0.392	0.384	0.8	0.714
8_fixed_threshold_5_skip_from_end	0.775	0.889	0.832	0.753	0.835	0.425	0.441	0.441	0.5	0.835	0.714
8_fixed_threshold_5_skip	0.705	0.882	0.842	0.827	0.835	0.438	0.453	0.453	0.5	0.835	0.553
8_fixed_threshold_10_skip_from_end	1	1	1	1	0.803	0.291	0.299	0.299	0.282	0.803	nan
8_fixed_threshold_10_skip	1	1	1	1	0.753	0.258	0.146	0.292	0.27	0.753	nan
9_fixed_threshold_from_end	0.851	0.958	0.912	0.863	0.877	0.616	0.596	0.596	0.618	0.877	0.506
9_fixed_threshold_5_skip_from_end	0.808	0.893	0.83	0.764	0.856	0.419	0.454	0.46	0.504	0.856	0.7
9_fixed_threshold_5_skip	0.803	0.937	0.896	0.854	0.887	0.426	0.418	0.44	0.539	0.887	0.592
9_fixed_threshold_10_skip_from_end	0.75	0.911	0.868	0.805	0.815	0.333	0.262	0.315	0.363	0.815	0.819
9_fixed_threshold_10_skip	1	1	1	1	0.75	0.261	0.148	0.295	0.273	0.75	nan
10_fixed_threshold_from_end	0.808	0.91	0.845	0.794	0.817	0.568	0.529	0.529	0.583	0.817	0.535
10_fixed_threshold_5_skip_from_end	0.839	0.947	0.903	0.821	0.877	0.344	0.46	0.472	0.524	0.877	0.593
10_fixed_threshold_10_skip	0.784	0.942	0.911	0.761	0.857	0.403	0.265	0.319	0.358	0.857	0.752
10_fixed_threshold_10_skip_from_end	0.825	0.919	0.867	0.805	0.796	0.38	0.275	0.378	0.435	0.796	0.696
10_fixed_threshold_5_skip	0.781	0.936	0.904	0.874	0.898	0.39	0.401	0.408	0.512	0.898	0.62
15_fixed_threshold_10_skip	1	1	1	1	0.786	0.257	0.171	0.3	0.271	0.786	nan
15_fixed_threshold_10_skip_from_end	1	1	1	1	0.775	0.254	0.169	0.296	0.268	0.775	nan
15_fixed_threshold_5_skip	1	1	1	1	0.756	0.256	0.146	0.293	0.268	0.756	nan
15_fixed_threshold_5_skip_from_end	0.74	0.901	0.851	0.711	0.834	0.356	0.293	0.352	0.384	0.834	0.811
15_fixed_threshold_from_end	0.844	0.953	0.907	0.843	0.887	0.434	0.403	0.415	0.542	0.887	0.735
20_fixed_threshold_10_skip_from_end	1	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
20_fixed_threshold_from_end	0.782	0.942	0.878	0.776	0.816	0.353	0.278	0.319	0.36	0.816	0.904
20_fixed_threshold_5_skip_from_end	0.8	0.941	0.9	0.77	0.853	0.459	0.288	0.395	0.458	0.853	0.729
20_fixed_threshold_5_skip	1	1	1	1	0.786	0.257	0.171	0.3	0.271	0.786	nan
20_fixed_threshold_10_skip	1	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
30_fixed_threshold_10_skip_from_end	1	1	1	1	0.806	0.226	0.177	0.274	0.29	0.806	nan
30_fixed_threshold_5_skip	1	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
30_fixed_threshold_5_skip_from_end	1	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
30_fixed_threshold_from_end	1	1	1	1	0.8	0.246	0.185	0.292	0.292	0.8	nan
30_fixed_threshold_10_skip	1	1	1	1	0.803	0.23	0.18	0.262	0.295	0.803	nan
40_fixed_threshold_10_skip	1	1	1	1	0.833	0.241	0.204	0.296	0.333	0.833	nan
40_fixed_threshold_10_skip_from_end	1	1	1	1	0.825	0.228	0.193	0.281	0.316	0.825	nan
40_fixed_threshold_5_skip	1	1	1	1	0.825	0.228	0.193	0.281	0.316	0.825	nan
40_fixed_threshold_5_skip_from_end	1	1	1	1	0.825	0.228	0.193	0.281	0.316	0.825	nan
40_fixed_threshold_from_end	1	1	1	1	0.806	0.226	0.177	0.274	0.29	0.806	nan
100_fixed_threshold_10_skip_from_end	1	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan
100_fixed_threshold_from_end	1	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan
100_fixed_threshold_5_skip_from_end	1	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan
100_fixed_threshold_5_skip	1	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan
100_fixed_threshold_10_skip	1	1	1	1	0.854	0.271	0.229	0.25	0.375	0.854	nan

Figure 19: 10000 Dataset - Experiment 3 Results - 2

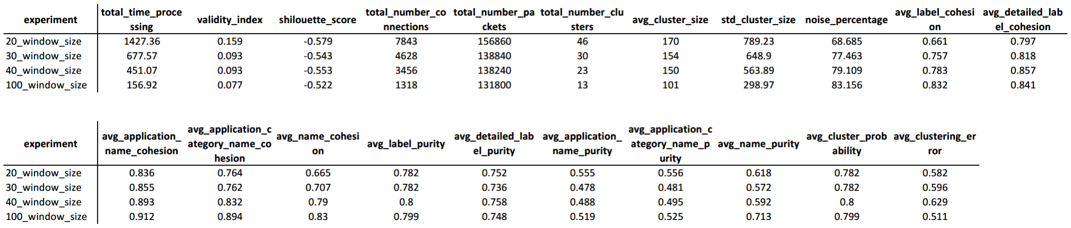
Min 20 Dataset



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
5_fixed_threshold	43.88	0.075	-0.563	1156	5780	31	37	116.04	56.661	0.553	0.727
6_fixed_threshold	29.72	0.102	-0.552	1156	6936	25	46	146.45	64.533	0.645	0.76
7_fixed_threshold	28.71	0.125	-0.258	1156	8092	19	60	183.67	70.675	0.652	0.76
8_fixed_threshold	28.83	0.144	-0.467	1156	9248	15	77	215.03	73.529	0.702	0.815
9_fixed_threshold	42.36	0.104	-0.269	1156	10404	13	88	238.92	76.298	0.701	0.767
10_fixed_threshold	55.64	0.109	-0.439	1156	11560	12	96	250.24	76.903	0.727	0.772
15_fixed_threshold	72.38	0.173	-0.335	1156	17340	8	144	220.19	45.588	0.602	0.753
20_fixed_threshold	58.09	0.175	-0.366	1156	23120	6	192	260.89	51.903	0.658	0.76
30_fixed_threshold	9.59	0.038	-0.063	338	10140	3	112	89.05	50	0.746	0.857
40_fixed_threshold	8.8	0.053	-0.06	284	11360	3	94	74.67	51.408	0.687	0.801
100_fixed_threshold	6.53	0	nan	188	18800	1	188	nan	100	1	1

experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold	0.885	0.797	0.6	0.956	0.937	0.896	0.909	0.91	0.956	0.133
6_fixed_threshold	0.91	0.83	0.713	0.963	0.95	0.907	0.913	0.919	0.963	0.133
7_fixed_threshold	0.888	0.803	0.634	0.933	0.913	0.85	0.851	0.861	0.933	0.199
8_fixed_threshold	0.906	0.836	0.732	0.921	0.9	0.854	0.855	0.855	0.921	0.215
9_fixed_threshold	0.893	0.802	0.658	0.895	0.863	0.758	0.76	0.782	0.895	0.345
10_fixed_threshold	0.888	0.791	0.707	0.9	0.87	0.761	0.763	0.775	0.9	0.316
15_fixed_threshold	0.884	0.766	0.667	0.845	0.832	0.566	0.566	0.588	0.845	0.559
20_fixed_threshold	0.903	0.835	0.739	0.816	0.78	0.573	0.573	0.566	0.816	0.432
30_fixed_threshold	0.918	0.842	0.713	0.812	0.745	0.426	0.453	0.38	0.812	0.678
40_fixed_threshold	0.917	0.845	0.731	0.733	0.668	0.364	0.387	0.432	0.733	0.735
100_fixed_threshold	1	1	1	0.66	0.447	0.324	0.388	0.473	0.66	nan

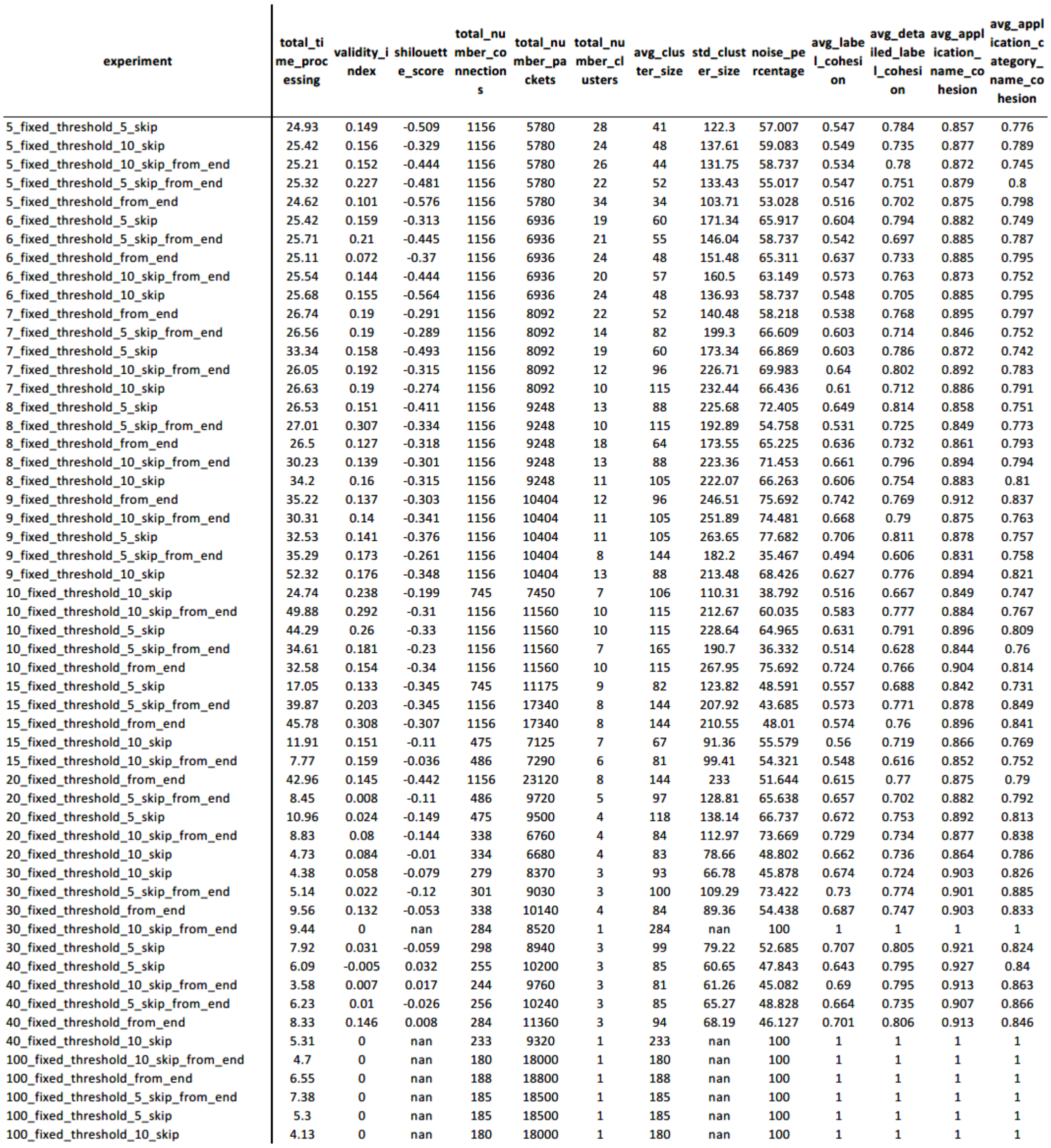
Figure 20: Min_20 Dataset - Experiment 1 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
20_window_size	1427.36	0.159	-0.579	7843	156860	46	170	789.23	68.685	0.661	0.797
30_window_size	677.57	0.093	-0.543	4628	138840	30	154	648.9	77.463	0.757	0.818
40_window_size	451.07	0.093	-0.553	3456	138240	23	150	563.89	79.109	0.783	0.857
100_window_size	156.92	0.077	-0.522	1318	131800	13	101	298.97	83.156	0.832	0.841

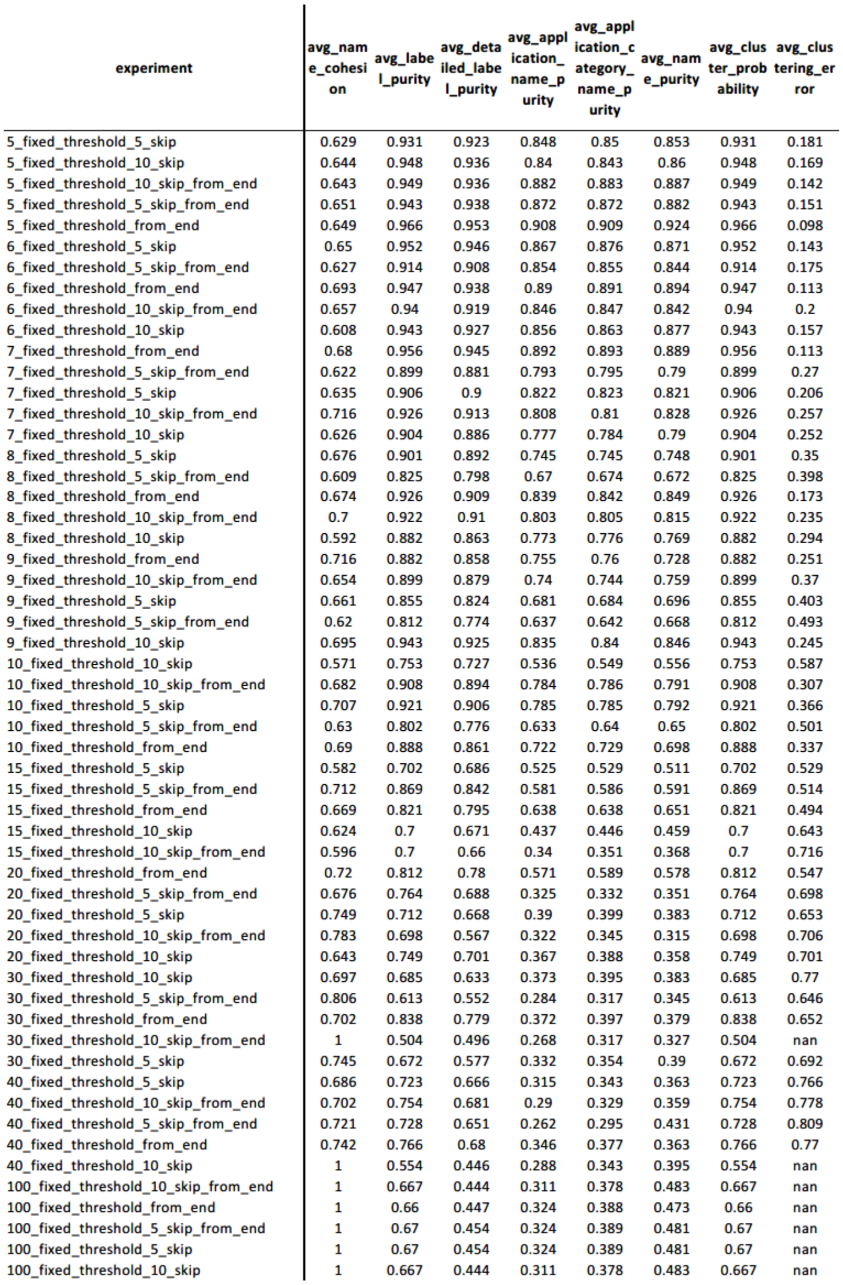
experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
20_window_size	0.836	0.764	0.665	0.782	0.752	0.555	0.556	0.618	0.782	0.582
30_window_size	0.855	0.762	0.707	0.782	0.736	0.478	0.481	0.572	0.782	0.596
40_window_size	0.893	0.832	0.79	0.8	0.758	0.488	0.495	0.592	0.8	0.629
100_window_size	0.912	0.894	0.83	0.799	0.748	0.519	0.525	0.713	0.799	0.511

Figure 21: Min_20 Dataset - Experiment 4 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion	avg_application_name_cohesion	avg_application_category_name_cohesion
5_fixed_threshold_5_skip	24.93	0.149	-0.509	1156	5780	28	41	122.3	57.007	0.547	0.784	0.857	0.776
5_fixed_threshold_10_skip	25.42	0.156	-0.329	1156	5780	24	48	137.61	59.083	0.549	0.735	0.877	0.789
5_fixed_threshold_10_skip_from_end	25.21	0.152	-0.444	1156	5780	26	44	131.75	58.737	0.534	0.78	0.872	0.745
5_fixed_threshold_5_skip_from_end	25.32	0.227	-0.481	1156	5780	22	52	133.43	55.017	0.547	0.751	0.879	0.8
5_fixed_threshold_from_end	24.62	0.101	-0.576	1156	5780	34	34	103.71	53.028	0.516	0.702	0.875	0.798
6_fixed_threshold_5_skip	25.42	0.159	-0.313	1156	6936	19	60	171.34	65.917	0.604	0.794	0.882	0.749
6_fixed_threshold_5_skip_from_end	25.71	0.21	-0.445	1156	6936	21	55	146.04	58.737	0.542	0.697	0.885	0.787
6_fixed_threshold_from_end	25.11	0.072	-0.37	1156	6936	24	48	151.48	65.311	0.637	0.733	0.885	0.795
6_fixed_threshold_10_skip_from_end	25.54	0.144	-0.444	1156	6936	20	57	160.5	63.149	0.573	0.763	0.873	0.752
6_fixed_threshold_10_skip	25.68	0.155	-0.564	1156	6936	24	48	136.93	58.737	0.548	0.705	0.885	0.795
7_fixed_threshold_from_end	26.74	0.19	-0.291	1156	8092	22	52	140.48	58.218	0.538	0.768	0.895	0.797
7_fixed_threshold_5_skip_from_end	26.56	0.19	-0.289	1156	8092	14	82	199.3	66.609	0.603	0.714	0.846	0.752
7_fixed_threshold_5_skip	33.34	0.158	-0.493	1156	8092	19	60	173.34	66.869	0.603	0.786	0.872	0.742
7_fixed_threshold_10_skip_from_end	26.05	0.192	-0.315	1156	8092	12	96	226.71	69.983	0.64	0.802	0.892	0.783
7_fixed_threshold_10_skip	26.63	0.19	-0.274	1156	8092	10	115	232.44	66.436	0.61	0.712	0.886	0.791
8_fixed_threshold_5_skip	26.53	0.151	-0.411	1156	9248	13	88	225.68	72.405	0.649	0.814	0.858	0.751
8_fixed_threshold_5_skip_from_end	27.01	0.307	-0.334	1156	9248	10	115	192.89	54.758	0.531	0.725	0.849	0.773
8_fixed_threshold_from_end	26.5	0.127	-0.318	1156	9248	18	64	173.55	65.225	0.636	0.732	0.861	0.793
8_fixed_threshold_10_skip_from_end	30.23	0.139	-0.301	1156	9248	13	88	223.36	71.453	0.661	0.796	0.894	0.794
8_fixed_threshold_10_skip	34.2	0.16	-0.315	1156	9248	11	105	222.07	66.263	0.606	0.754	0.883	0.81
9_fixed_threshold_from_end	35.22	0.137	-0.303	1156	10404	12	96	246.51	75.692	0.742	0.769	0.912	0.837
9_fixed_threshold_10_skip_from_end	30.31	0.14	-0.341	1156	10404	11	105	251.89	74.481	0.668	0.79	0.875	0.763
9_fixed_threshold_5_skip	32.53	0.141	-0.376	1156	10404	11	105	263.65	77.682	0.706	0.811	0.878	0.757
9_fixed_threshold_5_skip_from_end	35.29	0.173	-0.261	1156	10404	8	144	182.2	35.467	0.494	0.606	0.831	0.758
9_fixed_threshold_10_skip	52.32	0.176	-0.348	1156	10404	13	88	213.48	68.426	0.627	0.776	0.894	0.821
10_fixed_threshold_10_skip	24.74	0.238	-0.199	745	7450	7	106	110.31	38.792	0.516	0.667	0.849	0.747
10_fixed_threshold_10_skip_from_end	49.88	0.292	-0.31	1156	11560	10	115	212.67	60.035	0.583	0.777	0.884	0.767
10_fixed_threshold_5_skip	44.29	0.26	-0.33	1156	11560	10	115	228.64	64.965	0.631	0.791	0.896	0.809
10_fixed_threshold_5_skip_from_end	34.61	0.181	-0.23	1156	11560	7	165	190.7	36.332	0.514	0.628	0.844	0.76
10_fixed_threshold_from_end	32.58	0.154	-0.34	1156	11560	10	115	267.95	75.692	0.724	0.766	0.904	0.814
15_fixed_threshold_5_skip	17.05	0.133	-0.345	745	11175	9	82	123.82	48.591	0.557	0.688	0.842	0.731
15_fixed_threshold_5_skip_from_end	39.87	0.203	-0.345	1156	17340	8	144	207.92	43.685	0.573	0.771	0.878	0.849
15_fixed_threshold_from_end	45.78	0.308	-0.307	1156	17340	8	144	210.55	48.01	0.574	0.76	0.896	0.841
15_fixed_threshold_10_skip	11.91	0.151	-0.11	475	7125	7	67	91.36	55.579	0.56	0.719	0.866	0.769
15_fixed_threshold_10_skip_from_end	7.77	0.159	-0.036	486	7290	6	81	99.41	54.321	0.548	0.616	0.852	0.752
20_fixed_threshold_from_end	42.96	0.145	-0.442	1156	23120	8	144	233	51.644	0.615	0.77	0.875	0.79
20_fixed_threshold_5_skip_from_end	8.45	0.008	-0.11	486	9720	5	97	128.81	65.638	0.657	0.702	0.882	0.792
20_fixed_threshold_5_skip	10.96	0.024	-0.149	475	9500	4	118	138.14	66.737	0.672	0.753	0.892	0.813
20_fixed_threshold_10_skip_from_end	8.83	0.08	-0.144	338	6760	4	84	112.97	73.669	0.729	0.734	0.877	0.838
20_fixed_threshold_10_skip	4.73	0.084	-0.01	334	6680	4	83	78.66	48.802	0.662	0.736	0.864	0.786
30_fixed_threshold_10_skip	4.38	0.058	-0.079	279	8370	3	93	66.78	45.878	0.674	0.724	0.903	0.826
30_fixed_threshold_5_skip_from_end	5.14	0.022	-0.12	301	9030	3	100	109.29	73.422	0.73	0.774	0.901	0.885
30_fixed_threshold_from_end	9.56	0.132	-0.053	338	10140	4	84	89.36	54.438	0.687	0.747	0.903	0.833
30_fixed_threshold_10_skip_from_end	9.44	0	nan	284	8520	1	284	nan	100	1	1	1	1
30_fixed_threshold_5_skip	7.92	0.031	-0.059	298	8940	3	99	79.22	52.685	0.707	0.805	0.921	0.824
40_fixed_threshold_5_skip	6.09	-0.005	0.032	255	10200	3	85	60.65	47.843	0.643	0.795	0.927	0.84
40_fixed_threshold_10_skip_from_end	3.58	0.007	0.017	244	9760	3	81	61.26	45.082	0.69	0.795	0.913	0.863
40_fixed_threshold_5_skip_from_end	6.23	0.01	-0.026	256	10240	3	85	65.27	48.828	0.664	0.735	0.907	0.866
40_fixed_threshold_from_end	8.33	0.146	0.008	284	11360	3	94	68.19	46.127	0.701	0.806	0.913	0.846
40_fixed_threshold_10_skip	5.31	0	nan	233	9320	1	233	nan	100	1	1	1	1
100_fixed_threshold_10_skip_from_end	4.7	0	nan	180	18000	1	180	nan	100	1	1	1	1
100_fixed_threshold_from_end	6.55	0	nan	188	18800	1	188	nan	100	1	1	1	1
100_fixed_threshold_5_skip_from_end	7.38	0	nan	185	18500	1	185	nan	100	1	1	1	1
100_fixed_threshold_5_skip	5.3	0	nan	185	18500	1	185	nan	100	1	1	1	1
100_fixed_threshold_10_skip	4.13	0	nan	180	18000	1	180	nan	100	1	1	1	1

Figure 22: Min_20 Dataset - Experiment 3 Results - 1

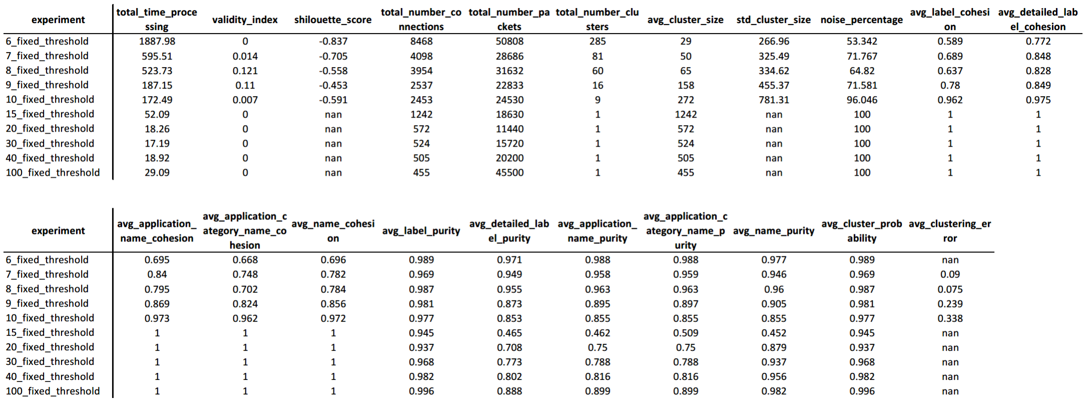


experiment	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold_5_skip	0.629	0.931	0.923	0.848	0.85	0.853	0.931	0.181
5_fixed_threshold_10_skip	0.644	0.948	0.936	0.84	0.843	0.86	0.948	0.169
5_fixed_threshold_10_skip_from_end	0.643	0.949	0.936	0.882	0.883	0.887	0.949	0.142
5_fixed_threshold_5_skip_from_end	0.651	0.943	0.938	0.872	0.872	0.882	0.943	0.151
5_fixed_threshold_from_end	0.649	0.966	0.953	0.908	0.909	0.924	0.966	0.098
6_fixed_threshold_5_skip	0.65	0.952	0.946	0.867	0.876	0.871	0.952	0.143
6_fixed_threshold_5_skip_from_end	0.627	0.914	0.908	0.854	0.855	0.844	0.914	0.175
6_fixed_threshold_from_end	0.693	0.947	0.938	0.89	0.891	0.894	0.947	0.113
6_fixed_threshold_10_skip_from_end	0.657	0.94	0.919	0.846	0.847	0.842	0.94	0.2
6_fixed_threshold_10_skip	0.608	0.943	0.927	0.856	0.863	0.877	0.943	0.157
7_fixed_threshold_from_end	0.68	0.956	0.945	0.892	0.893	0.889	0.956	0.113
7_fixed_threshold_5_skip_from_end	0.622	0.899	0.881	0.793	0.795	0.79	0.899	0.27
7_fixed_threshold_5_skip	0.635	0.906	0.9	0.822	0.823	0.821	0.906	0.206
7_fixed_threshold_10_skip_from_end	0.716	0.926	0.913	0.808	0.81	0.828	0.926	0.257
7_fixed_threshold_10_skip	0.626	0.904	0.886	0.777	0.784	0.79	0.904	0.252
8_fixed_threshold_5_skip	0.676	0.901	0.892	0.745	0.745	0.748	0.901	0.35
8_fixed_threshold_5_skip_from_end	0.609	0.825	0.798	0.67	0.674	0.672	0.825	0.398
8_fixed_threshold_from_end	0.674	0.926	0.909	0.839	0.842	0.849	0.926	0.173
8_fixed_threshold_10_skip_from_end	0.7	0.922	0.91	0.803	0.805	0.815	0.922	0.235
8_fixed_threshold_10_skip	0.592	0.882	0.863	0.773	0.776	0.769	0.882	0.294
9_fixed_threshold_from_end	0.716	0.882	0.858	0.755	0.76	0.728	0.882	0.251
9_fixed_threshold_10_skip_from_end	0.654	0.899	0.879	0.74	0.744	0.759	0.899	0.37
9_fixed_threshold_5_skip	0.661	0.855	0.824	0.681	0.684	0.696	0.855	0.403
9_fixed_threshold_5_skip_from_end	0.62	0.812	0.774	0.637	0.642	0.668	0.812	0.493
9_fixed_threshold_10_skip	0.695	0.943	0.925	0.835	0.84	0.846	0.943	0.245
10_fixed_threshold_10_skip	0.571	0.753	0.727	0.536	0.549	0.556	0.753	0.587
10_fixed_threshold_10_skip_from_end	0.682	0.908	0.894	0.784	0.786	0.791	0.908	0.307
10_fixed_threshold_5_skip	0.707	0.921	0.906	0.785	0.785	0.792	0.921	0.366
10_fixed_threshold_5_skip_from_end	0.63	0.802	0.776	0.633	0.64	0.65	0.802	0.501
10_fixed_threshold_from_end	0.69	0.888	0.861	0.722	0.729	0.698	0.888	0.337
15_fixed_threshold_5_skip	0.582	0.702	0.686	0.525	0.529	0.511	0.702	0.529
15_fixed_threshold_5_skip_from_end	0.712	0.869	0.842	0.581	0.586	0.591	0.869	0.514
15_fixed_threshold_from_end	0.669	0.821	0.795	0.638	0.638	0.651	0.821	0.494
15_fixed_threshold_10_skip	0.624	0.7	0.671	0.437	0.446	0.459	0.7	0.643
15_fixed_threshold_10_skip_from_end	0.596	0.7	0.66	0.34	0.351	0.368	0.7	0.716
20_fixed_threshold_from_end	0.72	0.812	0.78	0.571	0.589	0.578	0.812	0.547
20_fixed_threshold_5_skip_from_end	0.676	0.764	0.688	0.325	0.332	0.351	0.764	0.698
20_fixed_threshold_5_skip	0.749	0.712	0.668	0.39	0.399	0.383	0.712	0.653
20_fixed_threshold_10_skip_from_end	0.783	0.698	0.567	0.322	0.345	0.315	0.698	0.706
20_fixed_threshold_10_skip	0.643	0.749	0.701	0.367	0.388	0.358	0.749	0.701
30_fixed_threshold_10_skip	0.697	0.685	0.633	0.373	0.395	0.383	0.685	0.77
30_fixed_threshold_5_skip_from_end	0.806	0.613	0.552	0.284	0.317	0.345	0.613	0.646
30_fixed_threshold_from_end	0.702	0.838	0.779	0.372	0.397	0.379	0.838	0.652
30_fixed_threshold_10_skip_from_end	1	0.504	0.496	0.268	0.317	0.327	0.504	nan
30_fixed_threshold_5_skip	0.745	0.672	0.577	0.332	0.354	0.39	0.672	0.692
40_fixed_threshold_5_skip	0.686	0.723	0.666	0.315	0.343	0.363	0.723	0.766
40_fixed_threshold_10_skip_from_end	0.702	0.754	0.681	0.29	0.329	0.359	0.754	0.778
40_fixed_threshold_5_skip_from_end	0.721	0.728	0.651	0.262	0.295	0.431	0.728	0.809
40_fixed_threshold_from_end	0.742	0.766	0.68	0.346	0.377	0.363	0.766	0.77
40_fixed_threshold_10_skip	1	0.554	0.446	0.288	0.343	0.395	0.554	nan
100_fixed_threshold_10_skip_from_end	1	0.667	0.444	0.311	0.378	0.483	0.667	nan
100_fixed_threshold_from_end	1	0.66	0.447	0.324	0.388	0.473	0.66	nan
100_fixed_threshold_5_skip_from_end	1	0.67	0.454	0.324	0.389	0.481	0.67	nan
100_fixed_threshold_5_skip	1	0.67	0.454	0.324	0.389	0.481	0.67	nan
100_fixed_threshold_10_skip	1	0.667	0.444	0.311	0.378	0.483	0.667	nan

Figure 23: Min_20 Dataset - Experiment 3 Results - 2

Netflow Behavior

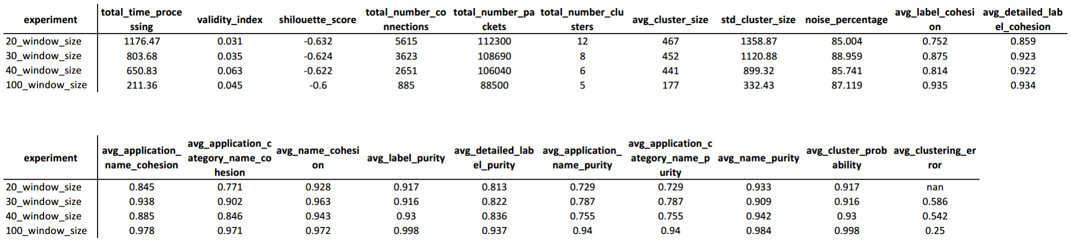
10000 Dataset



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
6_fixed_threshold	1887.98	0	-0.837	8468	50808	285	29	266.96	53.342	0.589	0.772
7_fixed_threshold	595.51	0.014	-0.705	4098	28686	81	50	325.49	71.767	0.689	0.848
8_fixed_threshold	523.73	0.121	-0.558	3954	31632	60	65	334.62	64.82	0.637	0.828
9_fixed_threshold	187.15	0.11	-0.453	2537	22833	16	158	455.37	71.581	0.78	0.849
10_fixed_threshold	172.49	0.007	-0.591	2453	24530	9	272	781.31	96.046	0.962	0.975
15_fixed_threshold	52.09	0	nan	1242	18630	1	1242	nan	100	1	1
20_fixed_threshold	18.26	0	nan	572	11440	1	572	nan	100	1	1
30_fixed_threshold	17.19	0	nan	524	15720	1	524	nan	100	1	1
40_fixed_threshold	18.92	0	nan	505	20200	1	505	nan	100	1	1
100_fixed_threshold	29.09	0	nan	455	45500	1	455	nan	100	1	1

experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
6_fixed_threshold	0.695	0.668	0.696	0.989	0.971	0.988	0.988	0.977	0.989	nan
7_fixed_threshold	0.84	0.748	0.782	0.969	0.949	0.958	0.959	0.946	0.969	0.09
8_fixed_threshold	0.795	0.702	0.784	0.987	0.955	0.963	0.963	0.96	0.987	0.075
9_fixed_threshold	0.869	0.824	0.856	0.981	0.873	0.895	0.897	0.905	0.981	0.239
10_fixed_threshold	0.973	0.962	0.972	0.977	0.853	0.855	0.855	0.855	0.977	0.338
15_fixed_threshold	1	1	1	0.945	0.465	0.462	0.509	0.452	0.945	nan
20_fixed_threshold	1	1	1	0.937	0.708	0.75	0.75	0.879	0.937	nan
30_fixed_threshold	1	1	1	0.968	0.773	0.788	0.788	0.937	0.968	nan
40_fixed_threshold	1	1	1	0.982	0.802	0.816	0.816	0.956	0.982	nan
100_fixed_threshold	1	1	1	0.996	0.888	0.899	0.899	0.982	0.996	nan

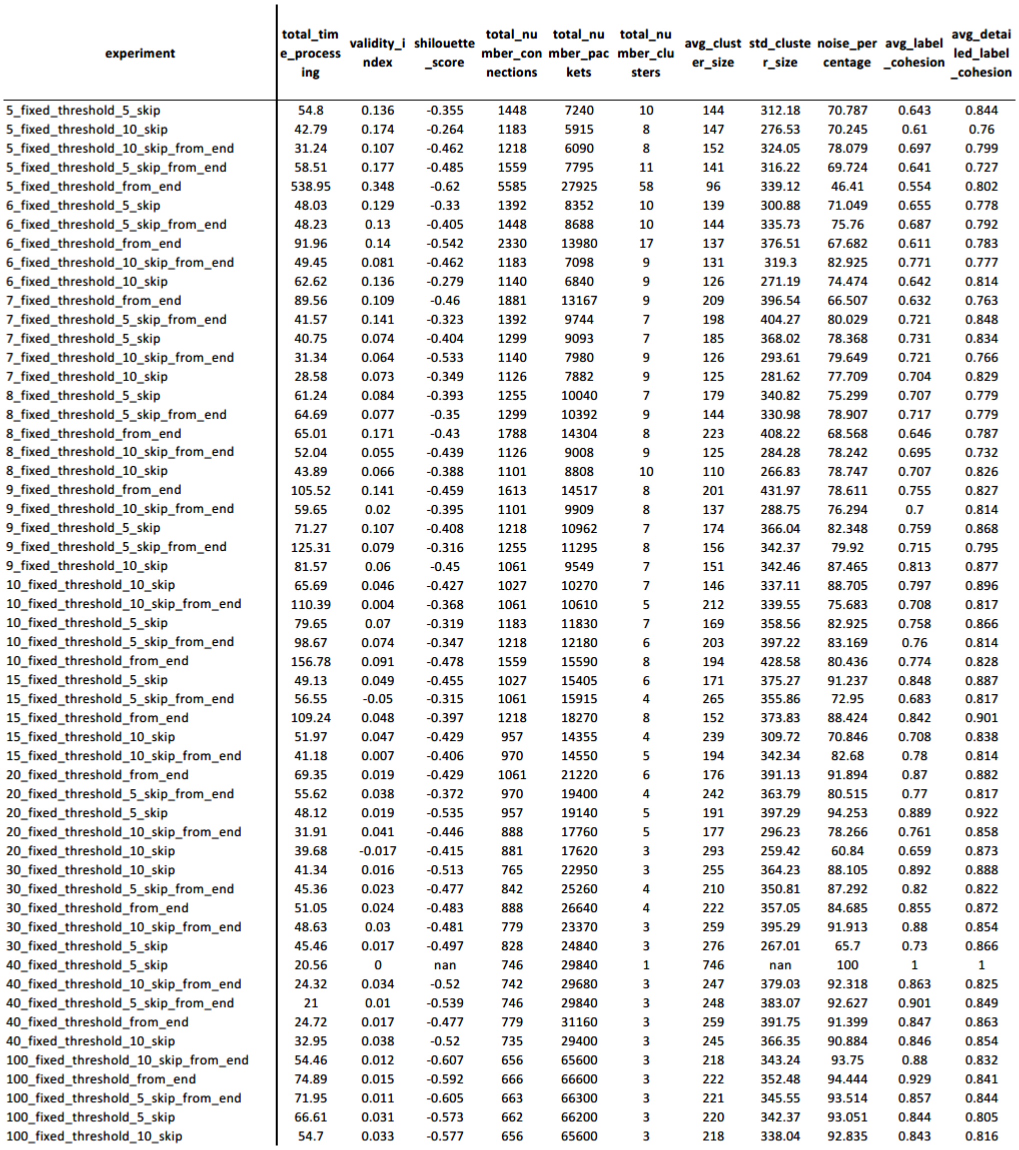
Figure 24: 10000 Dataset - Experiment 1 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
20_window_size	1176.47	0.031	-0.632	5615	112300	12	467	1358.87	85.004	0.752	0.859
30_window_size	803.68	0.035	-0.624	3623	108690	8	452	1120.88	88.959	0.875	0.923
40_window_size	650.83	0.063	-0.622	2651	106040	6	441	899.32	85.741	0.814	0.922
100_window_size	211.36	0.045	-0.6	885	88500	5	177	332.43	87.119	0.935	0.934

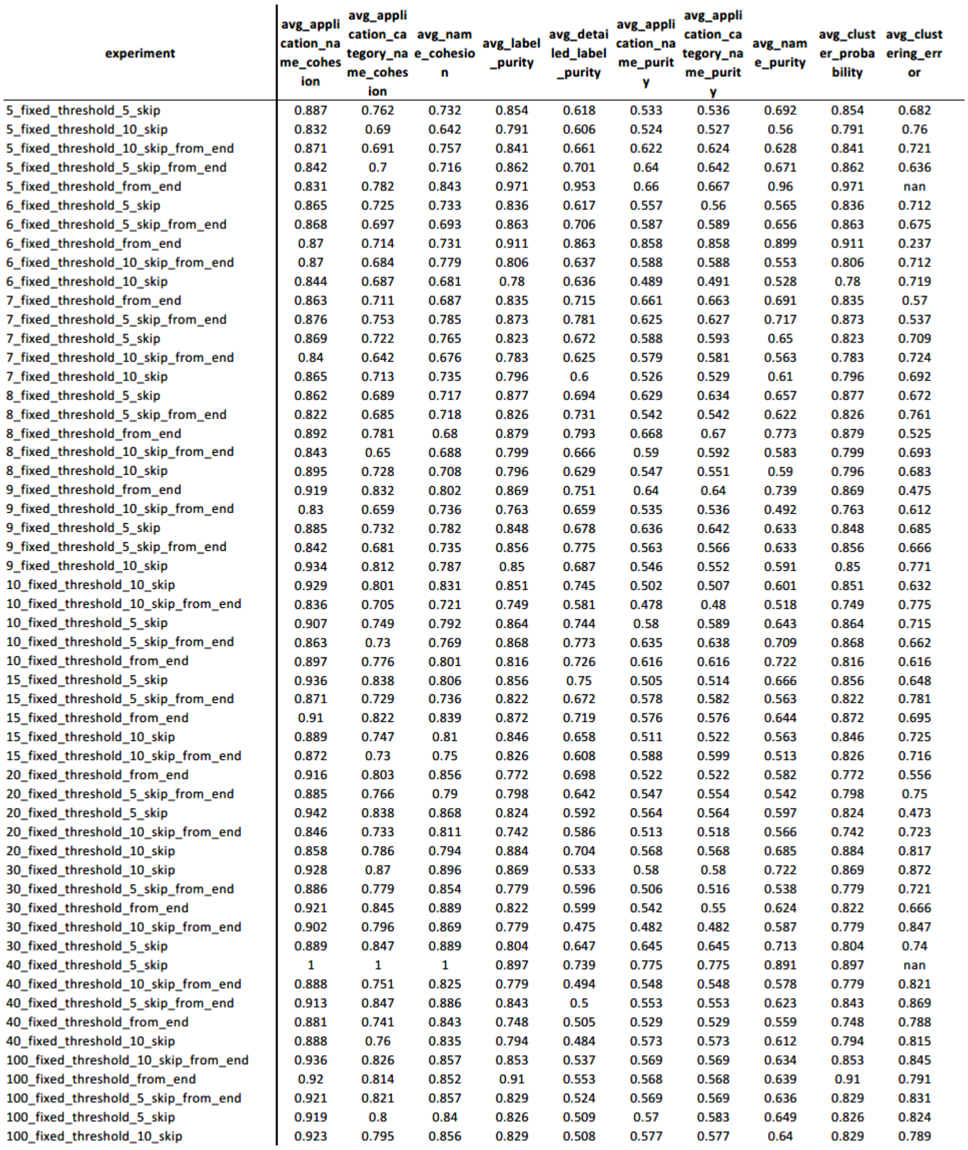
experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
20_window_size	0.845	0.771	0.928	0.917	0.813	0.729	0.729	0.933	0.917	nan
30_window_size	0.938	0.902	0.963	0.916	0.822	0.787	0.787	0.909	0.916	0.586
40_window_size	0.885	0.846	0.943	0.93	0.836	0.755	0.755	0.942	0.93	0.542
100_window_size	0.978	0.971	0.972	0.998	0.937	0.94	0.94	0.984	0.998	0.25

Figure 25: 10000 Dataset - Experiment 4 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
5_fixed_threshold_5_skip	54.8	0.136	-0.355	1448	7240	10	144	312.18	70.787	0.643	0.844
5_fixed_threshold_10_skip	42.79	0.174	-0.264	1183	5915	8	147	276.53	70.245	0.61	0.76
5_fixed_threshold_10_skip_from_end	31.24	0.107	-0.462	1218	6090	8	152	324.05	78.079	0.697	0.799
5_fixed_threshold_5_skip_from_end	58.51	0.177	-0.485	1559	7795	11	141	316.22	69.724	0.641	0.727
5_fixed_threshold_from_end	538.95	0.348	-0.62	5585	27925	58	96	339.12	46.41	0.554	0.802
6_fixed_threshold_5_skip	48.03	0.129	-0.33	1392	8352	10	139	300.88	71.049	0.655	0.778
6_fixed_threshold_5_skip_from_end	48.23	0.13	-0.405	1448	8688	10	144	335.73	75.76	0.687	0.792
6_fixed_threshold_from_end	91.96	0.14	-0.542	2330	13980	17	137	376.51	67.682	0.611	0.783
6_fixed_threshold_10_skip_from_end	49.45	0.081	-0.462	1183	7098	9	131	319.3	82.925	0.771	0.777
6_fixed_threshold_10_skip	62.62	0.136	-0.279	1140	6840	9	126	271.19	74.474	0.642	0.814
7_fixed_threshold_from_end	89.56	0.109	-0.46	1881	13167	9	209	396.54	66.507	0.632	0.763
7_fixed_threshold_5_skip_from_end	41.57	0.141	-0.323	1392	9744	7	198	404.27	80.029	0.721	0.848
7_fixed_threshold_5_skip	40.75	0.074	-0.404	1299	9093	7	185	368.02	78.368	0.731	0.834
7_fixed_threshold_10_skip_from_end	31.34	0.064	-0.533	1140	7980	9	126	293.61	79.649	0.721	0.766
7_fixed_threshold_10_skip	28.58	0.073	-0.349	1126	7882	9	125	281.62	77.709	0.704	0.829
8_fixed_threshold_5_skip	61.24	0.084	-0.393	1255	10040	7	179	340.82	75.299	0.707	0.779
8_fixed_threshold_5_skip_from_end	64.69	0.077	-0.35	1299	10392	9	144	330.98	78.907	0.717	0.779
8_fixed_threshold_from_end	65.01	0.171	-0.43	1788	14304	8	223	408.22	68.568	0.646	0.787
8_fixed_threshold_10_skip_from_end	52.04	0.055	-0.439	1126	9008	9	125	284.28	78.242	0.695	0.732
8_fixed_threshold_10_skip	43.89	0.066	-0.388	1101	8808	10	110	266.83	78.747	0.707	0.826
9_fixed_threshold_from_end	105.52	0.141	-0.459	1613	14517	8	201	431.97	78.611	0.755	0.827
9_fixed_threshold_10_skip_from_end	59.65	0.02	-0.395	1101	9909	8	137	288.75	76.294	0.7	0.814
9_fixed_threshold_5_skip	71.27	0.107	-0.408	1218	10962	7	174	366.04	82.348	0.759	0.868
9_fixed_threshold_5_skip_from_end	125.31	0.079	-0.316	1255	11295	8	156	342.37	79.92	0.715	0.795
9_fixed_threshold_10_skip	81.57	0.06	-0.45	1061	9549	7	151	342.46	87.465	0.813	0.877
10_fixed_threshold_10_skip	65.69	0.046	-0.427	1027	10270	7	146	337.11	88.705	0.797	0.896
10_fixed_threshold_10_skip_from_end	110.39	0.004	-0.368	1061	10610	5	212	339.55	75.683	0.708	0.817
10_fixed_threshold_5_skip	79.65	0.07	-0.319	1183	11830	7	169	358.56	82.925	0.758	0.866
10_fixed_threshold_5_skip_from_end	98.67	0.074	-0.347	1218	12180	6	203	397.22	83.169	0.76	0.814
10_fixed_threshold_from_end	156.78	0.091	-0.478	1559	15590	8	194	428.58	80.436	0.774	0.828
15_fixed_threshold_5_skip	49.13	0.049	-0.455	1027	15405	6	171	375.27	91.237	0.848	0.887
15_fixed_threshold_5_skip_from_end	56.55	-0.05	-0.315	1061	15915	4	265	355.86	72.95	0.683	0.817
15_fixed_threshold_from_end	109.24	0.048	-0.397	1218	18270	8	152	373.83	88.424	0.842	0.901
15_fixed_threshold_10_skip	51.97	0.047	-0.429	957	14355	4	239	309.72	70.846	0.708	0.838
15_fixed_threshold_10_skip_from_end	41.18	0.007	-0.406	970	14550	5	194	342.34	82.68	0.78	0.814
20_fixed_threshold_from_end	69.35	0.019	-0.429	1061	21220	6	176	391.13	91.894	0.87	0.882
20_fixed_threshold_5_skip_from_end	55.62	0.038	-0.372	970	19400	4	242	363.79	80.515	0.77	0.817
20_fixed_threshold_5_skip	48.12	0.019	-0.535	957	19140	5	191	397.29	94.253	0.889	0.922
20_fixed_threshold_10_skip_from_end	31.91	0.041	-0.446	888	17760	5	177	296.23	78.266	0.761	0.858
20_fixed_threshold_10_skip	39.68	-0.017	-0.415	881	17620	3	293	259.42	60.84	0.659	0.873
30_fixed_threshold_10_skip	41.34	0.016	-0.513	765	22950	3	255	364.23	88.105	0.892	0.888
30_fixed_threshold_5_skip_from_end	45.36	0.023	-0.477	842	25260	4	210	350.81	87.292	0.82	0.822
30_fixed_threshold_from_end	51.05	0.024	-0.483	888	26640	4	222	357.05	84.685	0.855	0.872
30_fixed_threshold_10_skip_from_end	48.63	0.03	-0.481	779	23370	3	259	395.29	91.913	0.88	0.854
30_fixed_threshold_5_skip	45.46	0.017	-0.497	828	24840	3	276	267.01	65.7	0.73	0.866
40_fixed_threshold_5_skip	20.56	0	nan	746	29840	1	746	nan	100	1	1
40_fixed_threshold_10_skip_from_end	24.32	0.034	-0.52	742	29680	3	247	379.03	92.318	0.863	0.825
40_fixed_threshold_5_skip_from_end	21	0.01	-0.539	746	29840	3	248	383.07	92.627	0.901	0.849
40_fixed_threshold_from_end	24.72	0.017	-0.477	779	31160	3	259	391.75	91.399	0.847	0.863
40_fixed_threshold_10_skip	32.95	0.038	-0.52	735	29400	3	245	366.35	90.884	0.846	0.854
100_fixed_threshold_10_skip_from_end	54.46	0.012	-0.607	656	65600	3	218	343.24	93.75	0.88	0.832
100_fixed_threshold_from_end	74.89	0.015	-0.592	666	66600	3	222	352.48	94.444	0.929	0.841
100_fixed_threshold_5_skip_from_end	71.95	0.011	-0.605	663	66300	3	221	345.55	93.514	0.857	0.844
100_fixed_threshold_5_skip	66.61	0.031	-0.573	662	66200	3	220	342.37	93.051	0.844	0.805
100_fixed_threshold_10_skip	54.7	0.033	-0.577	656	65600	3	218	338.04	92.835	0.843	0.816

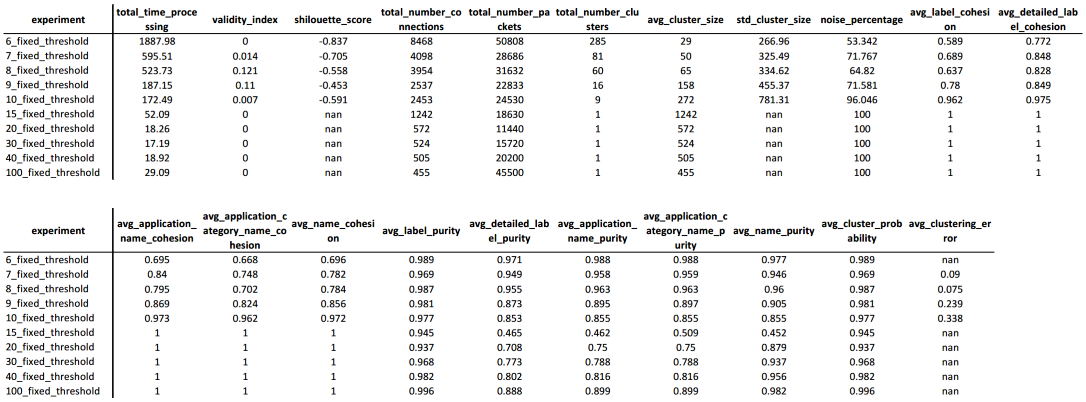
Figure 26: 10000 Dataset - Experiment 3 Results - 1



experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold_5_skip	0.887	0.762	0.732	0.854	0.618	0.533	0.536	0.692	0.854	0.682
5_fixed_threshold_10_skip	0.832	0.69	0.642	0.791	0.606	0.524	0.527	0.56	0.791	0.76
5_fixed_threshold_10_skip_from_end	0.871	0.691	0.757	0.841	0.661	0.622	0.624	0.628	0.841	0.721
5_fixed_threshold_5_skip_from_end	0.842	0.7	0.716	0.862	0.701	0.64	0.642	0.671	0.862	0.636
5_fixed_threshold_from_end	0.831	0.782	0.843	0.971	0.953	0.66	0.667	0.96	0.971	nan
6_fixed_threshold_5_skip	0.865	0.725	0.733	0.836	0.617	0.557	0.56	0.565	0.836	0.712
6_fixed_threshold_5_skip_from_end	0.868	0.697	0.693	0.863	0.706	0.587	0.589	0.656	0.863	0.675
6_fixed_threshold_from_end	0.87	0.714	0.731	0.911	0.863	0.858	0.858	0.899	0.911	0.237
6_fixed_threshold_10_skip_from_end	0.87	0.684	0.779	0.806	0.637	0.588	0.588	0.553	0.806	0.712
6_fixed_threshold_10_skip	0.844	0.687	0.681	0.78	0.636	0.489	0.491	0.528	0.78	0.719
7_fixed_threshold_from_end	0.863	0.711	0.687	0.835	0.715	0.661	0.663	0.691	0.835	0.57
7_fixed_threshold_5_skip_from_end	0.876	0.753	0.785	0.873	0.781	0.625	0.627	0.717	0.873	0.537
7_fixed_threshold_5_skip	0.869	0.722	0.765	0.823	0.672	0.588	0.593	0.65	0.823	0.709
7_fixed_threshold_10_skip_from_end	0.84	0.642	0.676	0.783	0.625	0.579	0.581	0.563	0.783	0.724
7_fixed_threshold_10_skip	0.865	0.713	0.735	0.796	0.6	0.526	0.529	0.61	0.796	0.692
8_fixed_threshold_5_skip	0.862	0.689	0.717	0.877	0.694	0.629	0.634	0.657	0.877	0.672
8_fixed_threshold_5_skip_from_end	0.822	0.685	0.718	0.826	0.731	0.542	0.542	0.622	0.826	0.761
8_fixed_threshold_from_end	0.892	0.781	0.68	0.879	0.793	0.668	0.67	0.773	0.879	0.525
8_fixed_threshold_10_skip_from_end	0.843	0.65	0.688	0.799	0.666	0.59	0.592	0.583	0.799	0.693
8_fixed_threshold_10_skip	0.895	0.728	0.708	0.796	0.629	0.547	0.551	0.59	0.796	0.683
9_fixed_threshold_from_end	0.919	0.832	0.802	0.869	0.751	0.64	0.64	0.739	0.869	0.475
9_fixed_threshold_10_skip_from_end	0.83	0.659	0.736	0.763	0.659	0.535	0.536	0.492	0.763	0.612
9_fixed_threshold_5_skip	0.885	0.732	0.782	0.848	0.678	0.636	0.642	0.633	0.848	0.685
9_fixed_threshold_5_skip_from_end	0.842	0.681	0.735	0.856	0.775	0.563	0.566	0.633	0.856	0.666
9_fixed_threshold_10_skip	0.934	0.812	0.787	0.85	0.687	0.546	0.552	0.591	0.85	0.771
10_fixed_threshold_10_skip	0.929	0.801	0.831	0.851	0.745	0.502	0.507	0.601	0.851	0.632
10_fixed_threshold_10_skip_from_end	0.836	0.705	0.721	0.749	0.581	0.478	0.48	0.518	0.749	0.775
10_fixed_threshold_5_skip	0.907	0.749	0.792	0.864	0.744	0.58	0.589	0.643	0.864	0.715
10_fixed_threshold_5_skip_from_end	0.863	0.73	0.769	0.868	0.773	0.635	0.638	0.709	0.868	0.662
10_fixed_threshold_from_end	0.897	0.776	0.801	0.816	0.726	0.616	0.616	0.722	0.816	0.616
15_fixed_threshold_5_skip	0.936	0.838	0.806	0.856	0.75	0.505	0.514	0.666	0.856	0.648
15_fixed_threshold_5_skip_from_end	0.871	0.729	0.736	0.822	0.672	0.578	0.582	0.563	0.822	0.781
15_fixed_threshold_from_end	0.91	0.822	0.839	0.872	0.719	0.576	0.576	0.644	0.872	0.695
15_fixed_threshold_10_skip	0.889	0.747	0.81	0.846	0.658	0.511	0.522	0.563	0.846	0.725
15_fixed_threshold_10_skip_from_end	0.872	0.73	0.75	0.826	0.608	0.588	0.599	0.513	0.826	0.716
20_fixed_threshold_from_end	0.916	0.803	0.856	0.772	0.698	0.522	0.522	0.582	0.772	0.556
20_fixed_threshold_5_skip_from_end	0.885	0.766	0.79	0.798	0.642	0.547	0.554	0.542	0.798	0.75
20_fixed_threshold_5_skip	0.942	0.838	0.868	0.824	0.592	0.564	0.564	0.597	0.824	0.473
20_fixed_threshold_10_skip_from_end	0.846	0.733	0.811	0.742	0.586	0.513	0.518	0.566	0.742	0.723
20_fixed_threshold_10_skip	0.858	0.786	0.794	0.884	0.704	0.568	0.568	0.685	0.884	0.817
30_fixed_threshold_10_skip	0.928	0.87	0.896	0.869	0.533	0.58	0.58	0.722	0.869	0.872
30_fixed_threshold_5_skip_from_end	0.886	0.779	0.854	0.779	0.596	0.506	0.516	0.538	0.779	0.721
30_fixed_threshold_from_end	0.921	0.845	0.889	0.822	0.599	0.542	0.55	0.624	0.822	0.666
30_fixed_threshold_10_skip_from_end	0.902	0.796	0.869	0.779	0.475	0.482	0.482	0.587	0.779	0.847
30_fixed_threshold_5_skip	0.889	0.847	0.889	0.804	0.647	0.645	0.645	0.713	0.804	0.74
40_fixed_threshold_5_skip	1	1	1	0.897	0.739	0.775	0.775	0.891	0.897	nan
40_fixed_threshold_10_skip_from_end	0.888	0.751	0.825	0.779	0.494	0.548	0.548	0.578	0.779	0.821
40_fixed_threshold_5_skip_from_end	0.913	0.847	0.886	0.843	0.5	0.553	0.553	0.623	0.843	0.869
40_fixed_threshold_from_end	0.881	0.741	0.843	0.748	0.505	0.529	0.529	0.559	0.748	0.788
40_fixed_threshold_10_skip	0.888	0.76	0.835	0.794	0.484	0.573	0.573	0.612	0.794	0.815
100_fixed_threshold_10_skip_from_end	0.936	0.826	0.857	0.853	0.537	0.569	0.569	0.634	0.853	0.845
100_fixed_threshold_from_end	0.92	0.814	0.852	0.91	0.553	0.568	0.568	0.639	0.91	0.791
100_fixed_threshold_5_skip_from_end	0.921	0.821	0.857	0.829	0.524	0.569	0.569	0.636	0.829	0.831
100_fixed_threshold_5_skip	0.919	0.8	0.84	0.826	0.509	0.57	0.583	0.649	0.826	0.824
100_fixed_threshold_10_skip	0.923	0.795	0.856	0.829	0.508	0.577	0.577	0.64	0.829	0.789

Figure 27: 10000 Dataset - Experiment 3 Results - 2

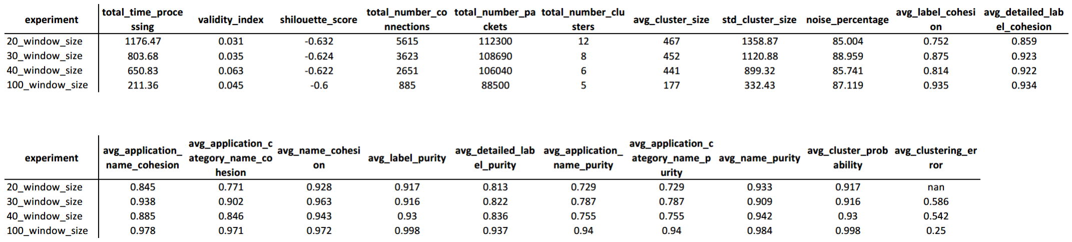
Min 20 Dataset



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
6_fixed_threshold	1887.98	0	-0.837	8468	50808	285	29	266.96	53.342	0.589	0.772
7_fixed_threshold	595.51	0.014	-0.705	4098	28686	81	50	325.49	71.767	0.689	0.848
8_fixed_threshold	523.73	0.121	-0.558	3954	31632	60	65	334.62	64.82	0.637	0.828
9_fixed_threshold	187.15	0.11	-0.453	2537	22833	16	158	455.37	71.581	0.78	0.849
10_fixed_threshold	172.49	0.007	-0.591	2453	24530	9	272	781.31	96.046	0.962	0.975
15_fixed_threshold	52.09	0	nan	1242	18630	1	1242	nan	100	1	1
20_fixed_threshold	18.26	0	nan	572	11440	1	572	nan	100	1	1
30_fixed_threshold	17.19	0	nan	524	15720	1	524	nan	100	1	1
40_fixed_threshold	18.92	0	nan	505	20200	1	505	nan	100	1	1
100_fixed_threshold	29.09	0	nan	455	45500	1	455	nan	100	1	1

experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
6_fixed_threshold	0.695	0.668	0.696	0.989	0.971	0.988	0.988	0.977	0.989	nan
7_fixed_threshold	0.84	0.748	0.782	0.969	0.949	0.958	0.959	0.946	0.969	0.09
8_fixed_threshold	0.795	0.702	0.784	0.987	0.955	0.963	0.963	0.96	0.987	0.075
9_fixed_threshold	0.869	0.824	0.856	0.981	0.873	0.895	0.897	0.905	0.981	0.239
10_fixed_threshold	0.973	0.962	0.972	0.977	0.853	0.855	0.855	0.855	0.977	0.338
15_fixed_threshold	1	1	1	0.945	0.465	0.462	0.509	0.452	0.945	nan
20_fixed_threshold	1	1	1	0.937	0.708	0.75	0.75	0.879	0.937	nan
30_fixed_threshold	1	1	1	0.968	0.773	0.788	0.788	0.937	0.968	nan
40_fixed_threshold	1	1	1	0.982	0.802	0.816	0.816	0.956	0.982	nan
100_fixed_threshold	1	1	1	0.996	0.888	0.899	0.899	0.982	0.996	nan

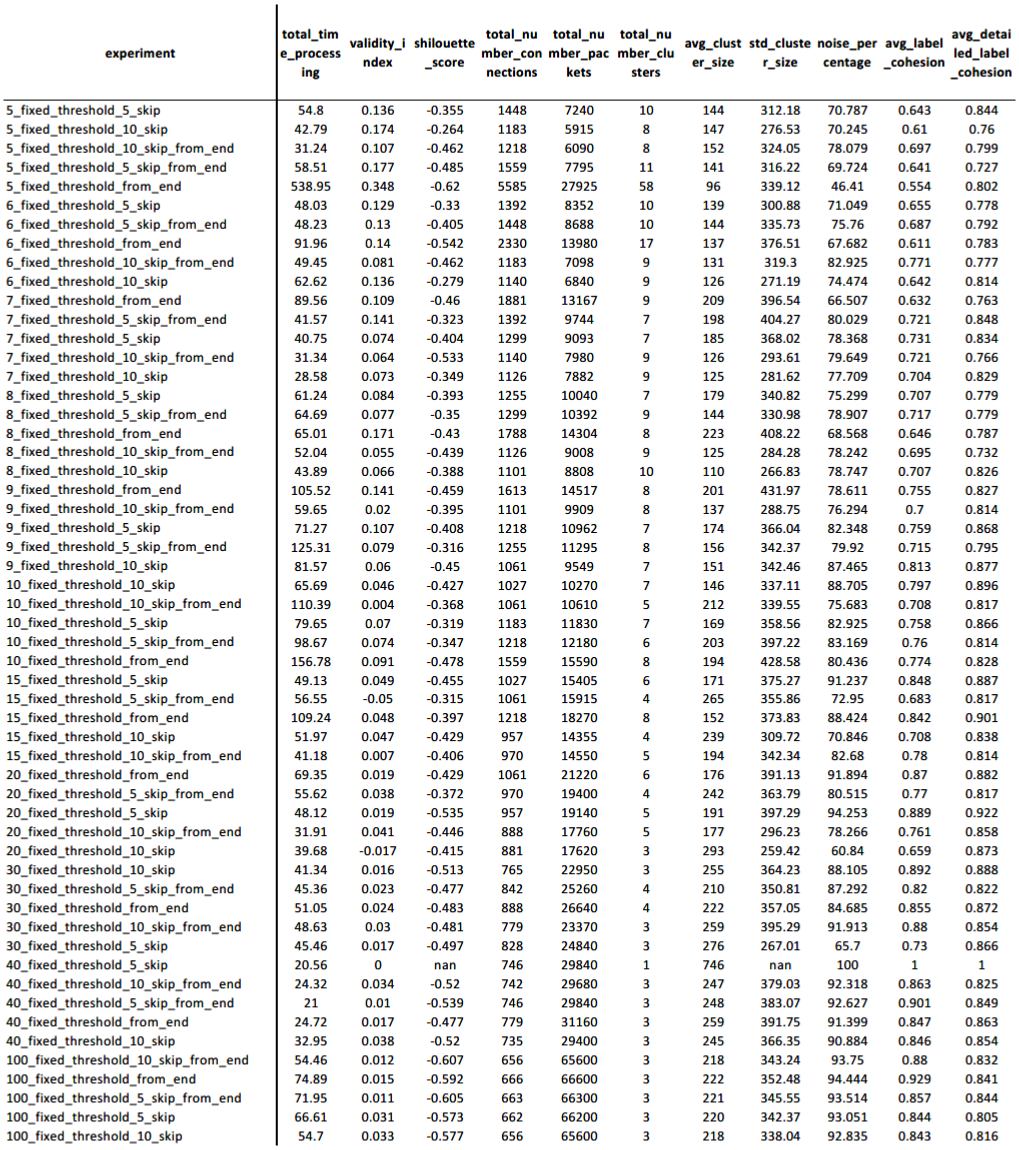
Figure 28: Min 20 - Experiment 1 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
20_window_size	1176.47	0.031	-0.632	5615	112300	12	467	1358.87	85.004	0.752	0.859
30_window_size	803.68	0.035	-0.624	3623	108690	8	452	1120.88	88.959	0.875	0.923
40_window_size	650.83	0.063	-0.622	2651	106040	6	441	899.32	85.741	0.814	0.922
100_window_size	211.36	0.045	-0.6	885	88500	5	177	332.43	87.119	0.935	0.934

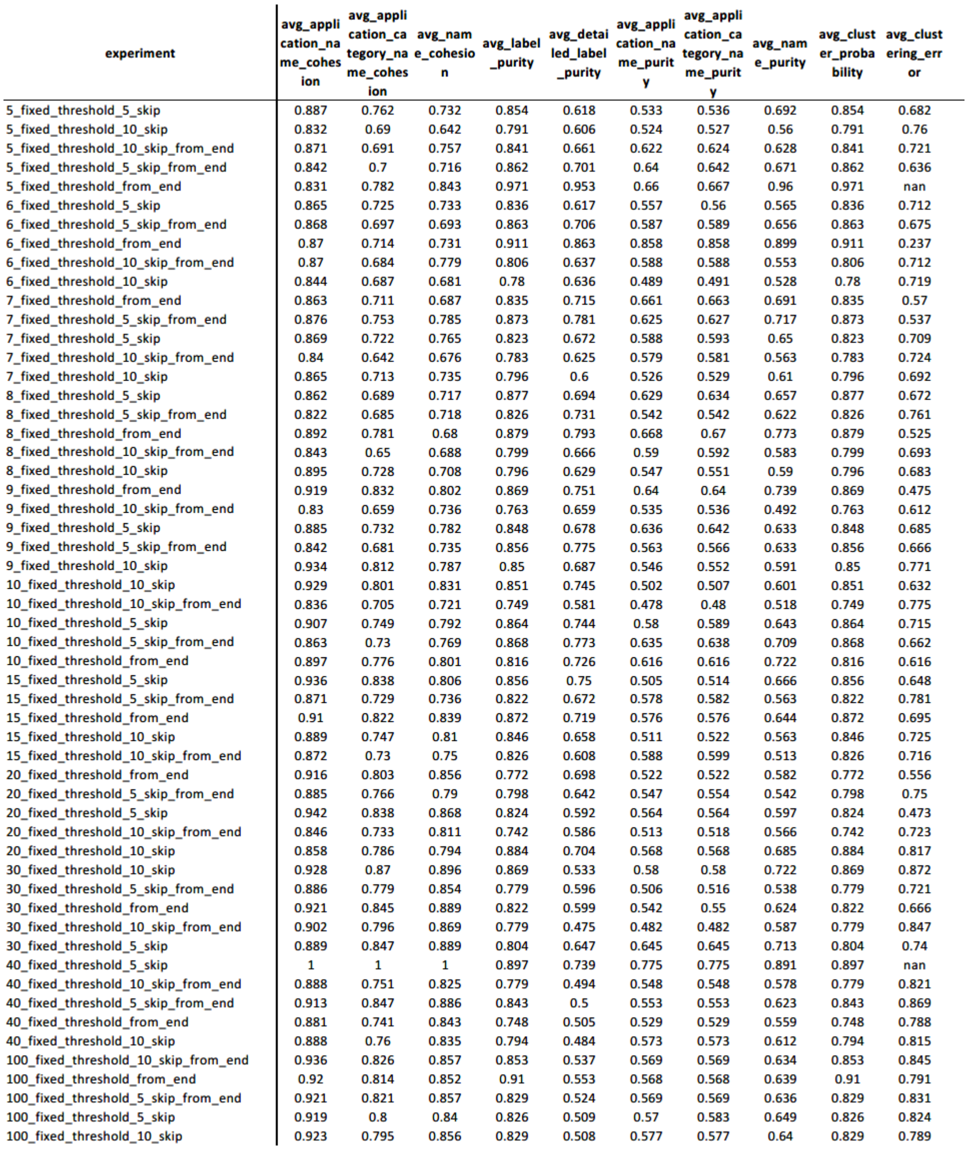
experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
20_window_size	0.845	0.771	0.928	0.917	0.813	0.729	0.729	0.933	0.917	nan
30_window_size	0.938	0.902	0.963	0.916	0.822	0.787	0.787	0.909	0.916	0.586
40_window_size	0.885	0.846	0.943	0.93	0.836	0.755	0.755	0.942	0.93	0.542
100_window_size	0.978	0.971	0.972	0.998	0.937	0.94	0.94	0.984	0.998	0.25

Figure 29: Min 20 - Experiment 4 Results



experiment	total_time_processing	validity_index	shilouette_score	total_number_connections	total_number_packets	total_number_clusters	avg_cluster_size	std_cluster_size	noise_percentage	avg_label_cohesion	avg_detailed_label_cohesion
5_fixed_threshold_5_skip	54.8	0.136	-0.355	1448	7240	10	144	312.18	70.787	0.643	0.844
5_fixed_threshold_10_skip	42.79	0.174	-0.264	1183	5915	8	147	276.53	70.245	0.61	0.76
5_fixed_threshold_10_skip_from_end	31.24	0.107	-0.462	1218	6090	8	152	324.05	78.079	0.697	0.799
5_fixed_threshold_5_skip_from_end	58.51	0.177	-0.485	1559	7795	11	141	316.22	69.724	0.641	0.727
5_fixed_threshold_from_end	538.95	0.348	-0.62	5585	27925	58	96	339.12	46.41	0.554	0.802
6_fixed_threshold_5_skip	48.03	0.129	-0.33	1392	8352	10	139	300.88	71.049	0.655	0.778
6_fixed_threshold_5_skip_from_end	48.23	0.13	-0.405	1448	8688	10	144	335.73	75.76	0.687	0.792
6_fixed_threshold_from_end	91.96	0.14	-0.542	2330	13980	17	137	376.51	67.682	0.611	0.783
6_fixed_threshold_10_skip_from_end	49.45	0.081	-0.462	1183	7098	9	131	319.3	82.925	0.771	0.777
6_fixed_threshold_10_skip	62.62	0.136	-0.279	1140	6840	9	126	271.19	74.474	0.642	0.814
7_fixed_threshold_from_end	89.56	0.109	-0.46	1881	13167	9	209	396.54	66.507	0.632	0.763
7_fixed_threshold_5_skip_from_end	41.57	0.141	-0.323	1392	9744	7	198	404.27	80.029	0.721	0.848
7_fixed_threshold_5_skip	40.75	0.074	-0.404	1299	9093	7	185	368.02	78.368	0.731	0.834
7_fixed_threshold_10_skip_from_end	31.34	0.064	-0.533	1140	7980	9	126	293.61	79.649	0.721	0.766
7_fixed_threshold_10_skip	28.58	0.073	-0.349	1126	7882	9	125	281.62	77.709	0.704	0.829
8_fixed_threshold_5_skip	61.24	0.084	-0.393	1255	10040	7	179	340.82	75.299	0.707	0.779
8_fixed_threshold_5_skip_from_end	64.69	0.077	-0.35	1299	10392	9	144	330.98	78.907	0.717	0.779
8_fixed_threshold_from_end	65.01	0.171	-0.43	1788	14304	8	223	408.22	68.568	0.646	0.787
8_fixed_threshold_10_skip_from_end	52.04	0.055	-0.439	1126	9008	9	125	284.28	78.242	0.695	0.732
8_fixed_threshold_10_skip	43.89	0.066	-0.388	1101	8808	10	110	266.83	78.747	0.707	0.826
9_fixed_threshold_from_end	105.52	0.141	-0.459	1613	14517	8	201	431.97	78.611	0.755	0.827
9_fixed_threshold_10_skip_from_end	59.65	0.02	-0.395	1101	9909	8	137	288.75	76.294	0.7	0.814
9_fixed_threshold_5_skip	71.27	0.107	-0.408	1218	10962	7	174	366.04	82.348	0.759	0.868
9_fixed_threshold_5_skip_from_end	125.31	0.079	-0.316	1255	11295	8	156	342.37	79.92	0.715	0.795
9_fixed_threshold_10_skip	81.57	0.06	-0.45	1061	9549	7	151	342.46	87.465	0.813	0.877
10_fixed_threshold_10_skip	65.69	0.046	-0.427	1027	10270	7	146	337.11	88.705	0.797	0.896
10_fixed_threshold_10_skip_from_end	110.39	0.004	-0.368	1061	10610	5	212	339.55	75.683	0.708	0.817
10_fixed_threshold_5_skip	79.65	0.07	-0.319	1183	11830	7	169	358.56	82.925	0.758	0.866
10_fixed_threshold_5_skip_from_end	98.67	0.074	-0.347	1218	12180	6	203	397.22	83.169	0.76	0.814
10_fixed_threshold_from_end	156.78	0.091	-0.478	1559	15590	8	194	428.58	80.436	0.774	0.828
15_fixed_threshold_5_skip	49.13	0.049	-0.455	1027	15405	6	171	375.27	91.237	0.848	0.887
15_fixed_threshold_5_skip_from_end	56.55	-0.05	-0.315	1061	15915	4	265	355.86	72.95	0.683	0.817
15_fixed_threshold_from_end	109.24	0.048	-0.397	1218	18270	8	152	373.83	88.424	0.842	0.901
15_fixed_threshold_10_skip	51.97	0.047	-0.429	957	14355	4	239	309.72	70.846	0.708	0.838
15_fixed_threshold_10_skip_from_end	41.18	0.007	-0.406	970	14550	5	194	342.34	82.68	0.78	0.814
20_fixed_threshold_from_end	69.35	0.019	-0.429	1061	21220	6	176	391.13	91.894	0.87	0.882
20_fixed_threshold_5_skip_from_end	55.62	0.038	-0.372	970	19400	4	242	363.79	80.515	0.77	0.817
20_fixed_threshold_5_skip	48.12	0.019	-0.535	957	19140	5	191	397.29	94.253	0.889	0.922
20_fixed_threshold_10_skip_from_end	31.91	0.041	-0.446	888	17760	5	177	296.23	78.266	0.761	0.858
20_fixed_threshold_10_skip	39.68	-0.017	-0.415	881	17620	3	293	259.42	60.84	0.659	0.873
30_fixed_threshold_10_skip	41.34	0.016	-0.513	765	22950	3	255	364.23	88.105	0.892	0.888
30_fixed_threshold_5_skip_from_end	45.36	0.023	-0.477	842	25260	4	210	350.81	87.292	0.82	0.822
30_fixed_threshold_from_end	51.05	0.024	-0.483	888	26640	4	222	357.05	84.685	0.855	0.872
30_fixed_threshold_10_skip_from_end	48.63	0.03	-0.481	779	23370	3	259	395.29	91.913	0.88	0.854
30_fixed_threshold_5_skip	45.46	0.017	-0.497	828	24840	3	276	267.01	65.7	0.73	0.866
40_fixed_threshold_5_skip	20.56	0	nan	746	29840	1	746	nan	100	1	1
40_fixed_threshold_10_skip_from_end	24.32	0.034	-0.52	742	29680	3	247	379.03	92.318	0.863	0.825
40_fixed_threshold_5_skip_from_end	21	0.01	-0.539	746	29840	3	248	383.07	92.627	0.901	0.849
40_fixed_threshold_from_end	24.72	0.017	-0.477	779	31160	3	259	391.75	91.399	0.847	0.863
40_fixed_threshold_10_skip	32.95	0.038	-0.52	735	29400	3	245	366.35	90.884	0.846	0.854
100_fixed_threshold_10_skip_from_end	54.46	0.012	-0.607	656	65600	3	218	343.24	93.75	0.88	0.832
100_fixed_threshold_from_end	74.89	0.015	-0.592	666	66600	3	222	352.48	94.444	0.929	0.841
100_fixed_threshold_5_skip_from_end	71.95	0.011	-0.605	663	66300	3	221	345.55	93.514	0.857	0.844
100_fixed_threshold_5_skip	66.61	0.031	-0.573	662	66200	3	220	342.37	93.051	0.844	0.805
100_fixed_threshold_10_skip	54.7	0.033	-0.577	656	65600	3	218	338.04	92.835	0.843	0.816

Figure 30: Min 20 - Experiment 3 Results - 1

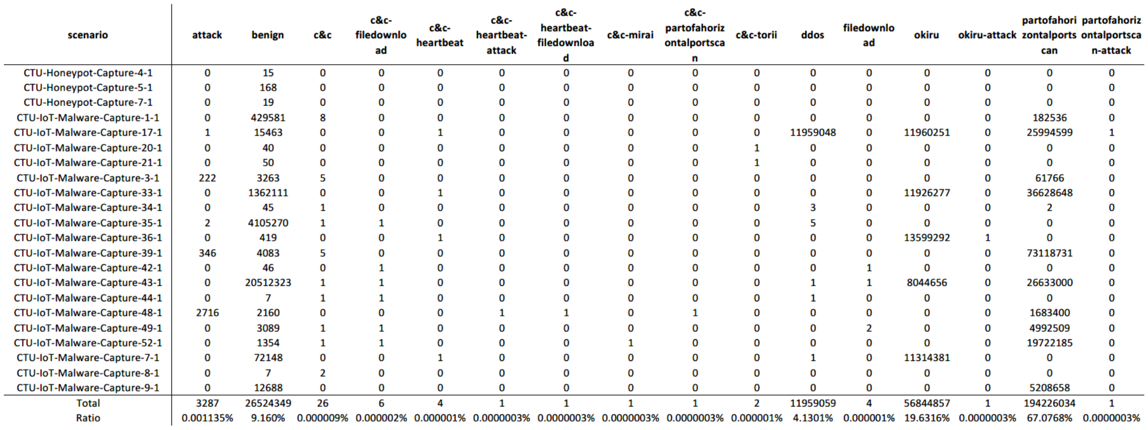


experiment	avg_application_name_cohesion	avg_application_category_name_cohesion	avg_name_cohesion	avg_label_purity	avg_detailed_label_purity	avg_application_name_purity	avg_application_category_name_purity	avg_name_purity	avg_cluster_probability	avg_clustering_error
5_fixed_threshold_5_skip	0.887	0.762	0.732	0.854	0.618	0.533	0.536	0.692	0.854	0.682
5_fixed_threshold_10_skip	0.832	0.69	0.642	0.791	0.606	0.524	0.527	0.56	0.791	0.76
5_fixed_threshold_10_skip_from_end	0.871	0.691	0.757	0.841	0.661	0.622	0.624	0.628	0.841	0.721
5_fixed_threshold_5_skip_from_end	0.842	0.7	0.716	0.862	0.701	0.64	0.642	0.671	0.862	0.636
5_fixed_threshold_from_end	0.831	0.782	0.843	0.971	0.953	0.66	0.667	0.96	0.971	nan
6_fixed_threshold_5_skip	0.865	0.725	0.733	0.836	0.617	0.557	0.56	0.565	0.836	0.712
6_fixed_threshold_5_skip_from_end	0.868	0.697	0.693	0.863	0.706	0.587	0.589	0.656	0.863	0.675
6_fixed_threshold_from_end	0.87	0.714	0.731	0.911	0.863	0.858	0.858	0.899	0.911	0.237
6_fixed_threshold_10_skip_from_end	0.87	0.684	0.779	0.806	0.637	0.588	0.588	0.553	0.806	0.712
6_fixed_threshold_10_skip	0.844	0.687	0.681	0.78	0.636	0.489	0.491	0.528	0.78	0.719
7_fixed_threshold_from_end	0.863	0.711	0.687	0.835	0.715	0.661	0.663	0.691	0.835	0.57
7_fixed_threshold_5_skip_from_end	0.876	0.753	0.785	0.873	0.781	0.625	0.627	0.717	0.873	0.537
7_fixed_threshold_5_skip	0.869	0.722	0.765	0.823	0.672	0.588	0.593	0.65	0.823	0.709
7_fixed_threshold_10_skip_from_end	0.84	0.642	0.676	0.783	0.625	0.579	0.581	0.563	0.783	0.724
7_fixed_threshold_10_skip	0.865	0.713	0.735	0.796	0.6	0.526	0.529	0.61	0.796	0.692
8_fixed_threshold_5_skip	0.862	0.689	0.717	0.877	0.694	0.629	0.634	0.657	0.877	0.672
8_fixed_threshold_5_skip_from_end	0.822	0.685	0.718	0.826	0.731	0.542	0.542	0.622	0.826	0.761
8_fixed_threshold_from_end	0.892	0.781	0.68	0.879	0.793	0.668	0.67	0.773	0.879	0.525
8_fixed_threshold_10_skip_from_end	0.843	0.65	0.688	0.799	0.666	0.59	0.592	0.583	0.799	0.693
8_fixed_threshold_10_skip	0.895	0.728	0.708	0.796	0.629	0.547	0.551	0.59	0.796	0.683
9_fixed_threshold_from_end	0.919	0.832	0.802	0.869	0.751	0.64	0.64	0.739	0.869	0.475
9_fixed_threshold_10_skip_from_end	0.83	0.659	0.736	0.763	0.659	0.535	0.536	0.492	0.763	0.612
9_fixed_threshold_5_skip	0.885	0.732	0.782	0.848	0.678	0.636	0.642	0.633	0.848	0.685
9_fixed_threshold_5_skip_from_end	0.842	0.681	0.735	0.856	0.775	0.563	0.566	0.633	0.856	0.666
9_fixed_threshold_10_skip	0.934	0.812	0.787	0.85	0.687	0.546	0.552	0.591	0.85	0.771
10_fixed_threshold_10_skip	0.929	0.801	0.831	0.851	0.745	0.502	0.507	0.601	0.851	0.632
10_fixed_threshold_10_skip_from_end	0.836	0.705	0.721	0.749	0.581	0.478	0.48	0.518	0.749	0.775
10_fixed_threshold_5_skip	0.907	0.749	0.792	0.864	0.744	0.58	0.589	0.643	0.864	0.715
10_fixed_threshold_5_skip_from_end	0.863	0.73	0.769	0.868	0.773	0.635	0.638	0.709	0.868	0.662
10_fixed_threshold_from_end	0.897	0.776	0.801	0.816	0.726	0.616	0.616	0.722	0.816	0.616
15_fixed_threshold_5_skip	0.936	0.838	0.806	0.856	0.75	0.505	0.514	0.666	0.856	0.648
15_fixed_threshold_5_skip_from_end	0.871	0.729	0.736	0.822	0.672	0.578	0.582	0.563	0.822	0.781
15_fixed_threshold_from_end	0.91	0.822	0.839	0.872	0.719	0.576	0.576	0.644	0.872	0.695
15_fixed_threshold_10_skip	0.889	0.747	0.81	0.846	0.658	0.511	0.522	0.563	0.846	0.725
15_fixed_threshold_10_skip_from_end	0.872	0.73	0.75	0.826	0.608	0.588	0.599	0.513	0.826	0.716
20_fixed_threshold_from_end	0.916	0.803	0.856	0.772	0.698	0.522	0.522	0.582	0.772	0.556
20_fixed_threshold_5_skip_from_end	0.885	0.766	0.79	0.798	0.642	0.547	0.554	0.542	0.798	0.75
20_fixed_threshold_5_skip	0.942	0.838	0.868	0.824	0.592	0.564	0.564	0.597	0.824	0.473
20_fixed_threshold_10_skip_from_end	0.846	0.733	0.811	0.742	0.586	0.513	0.518	0.566	0.742	0.723
20_fixed_threshold_10_skip	0.858	0.786	0.794	0.884	0.704	0.568	0.568	0.685	0.884	0.817
30_fixed_threshold_10_skip	0.928	0.87	0.896	0.869	0.533	0.58	0.58	0.722	0.869	0.872
30_fixed_threshold_5_skip_from_end	0.886	0.779	0.854	0.779	0.596	0.506	0.516	0.538	0.779	0.721
30_fixed_threshold_from_end	0.921	0.845	0.889	0.822	0.599	0.542	0.55	0.624	0.822	0.666
30_fixed_threshold_10_skip_from_end	0.902	0.796	0.869	0.779	0.475	0.482	0.482	0.587	0.779	0.847
30_fixed_threshold_5_skip	0.889	0.847	0.889	0.804	0.647	0.645	0.645	0.713	0.804	0.74
40_fixed_threshold_5_skip	1	1	1	0.897	0.739	0.775	0.775	0.891	0.897	nan
40_fixed_threshold_10_skip_from_end	0.888	0.751	0.825	0.779	0.494	0.548	0.548	0.578	0.779	0.821
40_fixed_threshold_5_skip_from_end	0.913	0.847	0.886	0.843	0.5	0.553	0.553	0.623	0.843	0.869
40_fixed_threshold_from_end	0.881	0.741	0.843	0.748	0.505	0.529	0.529	0.559	0.748	0.788
40_fixed_threshold_10_skip	0.888	0.76	0.835	0.794	0.484	0.573	0.573	0.612	0.794	0.815
100_fixed_threshold_10_skip_from_end	0.936	0.826	0.857	0.853	0.537	0.569	0.569	0.634	0.853	0.845
100_fixed_threshold_from_end	0.92	0.814	0.852	0.91	0.553	0.568	0.568	0.639	0.91	0.791
100_fixed_threshold_5_skip_from_end	0.921	0.821	0.857	0.829	0.524	0.569	0.569	0.636	0.829	0.831
100_fixed_threshold_5_skip	0.919	0.8	0.84	0.826	0.509	0.57	0.583	0.649	0.826	0.824
100_fixed_threshold_10_skip	0.923	0.795	0.856	0.829	0.508	0.577	0.577	0.64	0.829	0.789

Figure 31: Min 20 - Experiment 3 Results - 2

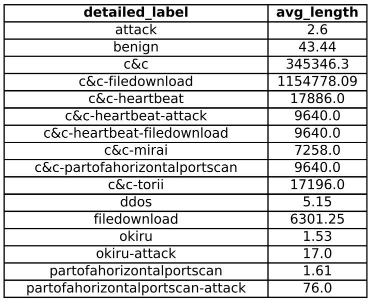
7.5 Dataset Information

Original Dataset



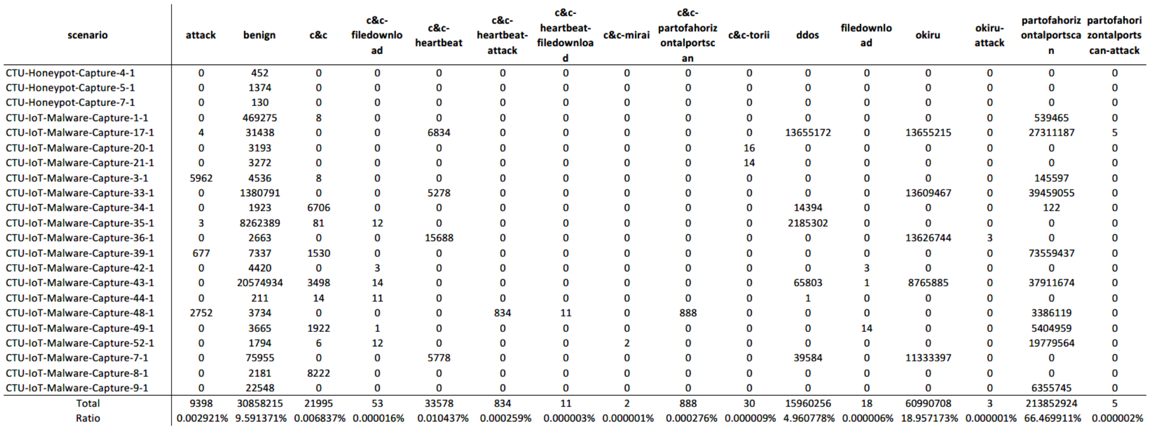
scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	168	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	429581	8	0	0	0	0	0	0	0	0	0	0	0	182536	0
CTU-IoT-Malware-Capture-17-1	1	15463	0	0	1	0	0	0	0	0	11959048	0	11960251	0	25994599	1
CTU-IoT-Malware-Capture-20-1	0	40	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	50	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	222	3263	5	0	0	0	0	0	0	0	0	0	0	0	61766	0
CTU-IoT-Malware-Capture-33-1	0	1362111	0	0	1	0	0	0	0	0	0	0	11926277	0	36628648	0
CTU-IoT-Malware-Capture-34-1	0	45	1	0	0	0	0	0	0	0	3	0	0	0	2	0
CTU-IoT-Malware-Capture-35-1	2	4105270	1	1	0	0	0	0	0	0	5	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	419	0	0	1	0	0	0	0	0	0	0	13599292	1	0	0
CTU-IoT-Malware-Capture-39-1	346	4083	5	0	0	0	0	0	0	0	0	0	0	0	73118731	0
CTU-IoT-Malware-Capture-42-1	0	46	0	1	0	0	0	0	0	0	0	1	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	20512323	1	1	0	0	0	0	0	0	1	1	8044656	0	26633000	0
CTU-IoT-Malware-Capture-44-1	0	7	1	1	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	2716	2160	0	0	0	1	1	0	1	0	0	0	0	0	1683400	0
CTU-IoT-Malware-Capture-49-1	0	3089	1	1	0	0	0	0	0	0	0	2	0	0	4992509	0
CTU-IoT-Malware-Capture-52-1	0	1354	1	1	0	0	0	1	0	0	0	0	0	0	19722185	0
CTU-IoT-Malware-Capture-7-1	0	72148	0	0	1	0	0	0	0	0	1	0	11314381	0	0	0
CTU-IoT-Malware-Capture-8-1	0	7	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	12688	0	0	0	0	0	0	0	0	0	0	0	0	5208658	0
Total	3287	26524349	26	6	4	1	1	1	1	2	11959059	4	56844857	1	194226034	1
Ratio	0.001135%	9.160%	0.000009%	0.000002%	0.000001%	0.0000003%	0.0000003%	0.0000003%	0.0000003%	0.000001%	4.1301%	0.000001%	19.6316%	0.0000003%	67.0768%	0.0000003%

Figure 32: Detailed Label Distribution - Unidirectional Behavior



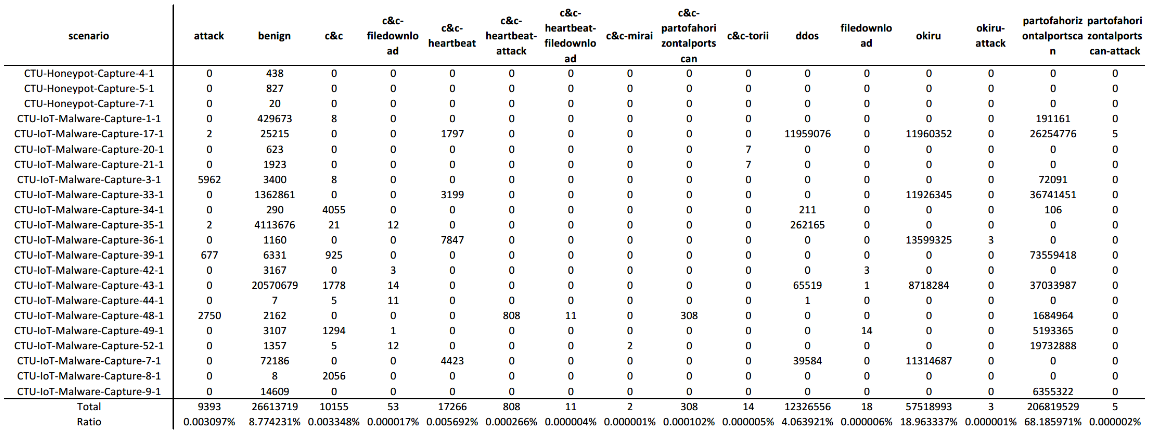
detailed_label	avg_length
attack	2.6
benign	43.44
c&c	345346.3
c&c-filedownload	1154778.09
c&c-heartbeat	17886.0
c&c-heartbeat-attack	9640.0
c&c-heartbeat-filedownload	9640.0
c&c-mirai	7258.0
c&c-partofahorizontalportscan	9640.0
c&c-torii	17196.0
ddos	5.15
filedownload	6301.25
okiru	1.53
okiru-attack	17.0
partofahorizontalportscan	1.61
partofahorizontalportscan-attack	76.0

Figure 33: Detailed Label Average Length - Unidirectional Behavior



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	452	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	1374	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	130	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	469275	8	0	0	0	0	0	0	0	0	0	0	0	539465	0
CTU-IoT-Malware-Capture-17-1	4	31438	0	0	6834	0	0	0	0	0	13655172	0	13655215	0	27311187	5
CTU-IoT-Malware-Capture-20-1	0	3193	0	0	0	0	0	0	0	16	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	3272	0	0	0	0	0	0	0	14	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	5962	4536	8	0	0	0	0	0	0	0	0	0	0	0	145597	0
CTU-IoT-Malware-Capture-33-1	0	1380791	0	0	5278	0	0	0	0	0	0	0	13609467	0	39459055	0
CTU-IoT-Malware-Capture-34-1	0	1923	6706	0	0	0	0	0	0	0	14394	0	0	0	122	0
CTU-IoT-Malware-Capture-35-1	3	8262389	81	12	0	0	0	0	0	0	2185302	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	2663	0	0	15688	0	0	0	0	0	0	0	13626744	3	0	0
CTU-IoT-Malware-Capture-39-1	677	7337	1530	0	0	0	0	0	0	0	0	0	0	0	73559437	0
CTU-IoT-Malware-Capture-42-1	0	4420	0	3	0	0	0	0	0	0	0	3	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	20574934	3498	14	0	0	0	0	0	0	65803	1	8765885	0	37911674	0
CTU-IoT-Malware-Capture-44-1	0	211	14	11	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	2752	3734	0	0	0	834	11	0	888	0	0	0	0	0	3386119	0
CTU-IoT-Malware-Capture-49-1	0	3665	1922	1	0	0	0	0	0	0	0	14	0	0	5404959	0
CTU-IoT-Malware-Capture-52-1	0	1794	6	12	0	0	0	2	0	0	0	0	0	0	19779564	0
CTU-IoT-Malware-Capture-7-1	0	75955	0	0	5778	0	0	0	0	0	39584	0	11333397	0	0	0
CTU-IoT-Malware-Capture-8-1	0	2181	8222	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	22548	0	0	0	0	0	0	0	0	0	0	0	0	6355745	0
Total	9398	30858215	21995	53	33578	834	11	2	888	30	15960256	18	60990708	3	213852924	5
Ratio	0.002921%	9.591371%	0.006837%	0.000016%	0.010437%	0.000259%	0.000003%	0.000001%	0.000276%	0.000009%	4.960778%	0.000006%	18.957173%	0.000001%	66.469911%	0.000002%

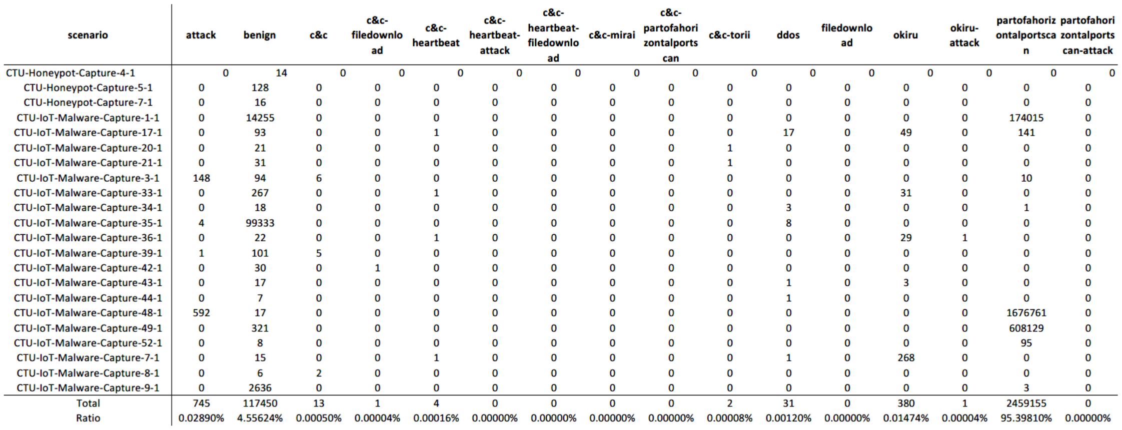
Figure 34: Detailed Label Distribution - IoT-23 Behavior



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	438	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	827	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	429673	8	0	0	0	0	0	0	0	0	0	0	0	191161	0
CTU-IoT-Malware-Capture-17-1	2	25215	0	0	1797	0	0	0	0	0	11959076	0	11960352	0	26254776	5
CTU-IoT-Malware-Capture-20-1	0	623	0	0	0	0	0	0	0	7	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	1923	0	0	0	0	0	0	0	7	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	5962	3400	8	0	0	0	0	0	0	0	0	0	0	0	72091	0
CTU-IoT-Malware-Capture-33-1	0	1362861	0	0	3199	0	0	0	0	0	0	0	11926345	0	36741451	0
CTU-IoT-Malware-Capture-34-1	0	290	4055	0	0	0	0	0	0	0	211	0	0	0	106	0
CTU-IoT-Malware-Capture-35-1	2	4113676	21	12	0	0	0	0	0	0	262165	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	1160	0	0	7847	0	0	0	0	0	0	0	13599325	3	0	0
CTU-IoT-Malware-Capture-39-1	677	6331	925	0	0	0	0	0	0	0	0	0	0	0	73559418	0
CTU-IoT-Malware-Capture-42-1	0	3167	0	3	0	0	0	0	0	0	0	3	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	20570679	1778	14	0	0	0	0	0	0	65519	1	8718284	0	37033987	0
CTU-IoT-Malware-Capture-44-1	0	7	5	11	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	2750	2162	0	0	0	808	11	0	308	0	0	0	0	0	1684964	0
CTU-IoT-Malware-Capture-49-1	0	3107	1294	1	0	0	0	0	0	0	0	14	0	0	5193365	0
CTU-IoT-Malware-Capture-52-1	0	1357	5	12	0	0	0	2	0	0	0	0	0	0	19732888	0
CTU-IoT-Malware-Capture-7-1	0	72186	0	0	4423	0	0	0	0	0	39584	0	11314687	0	0	0
CTU-IoT-Malware-Capture-8-1	0	8	2056	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	14609	0	0	0	0	0	0	0	0	0	0	0	0	6355322	0
Total	9393	26613719	10155	53	17266	808	11	2	308	14	12326556	18	57518993	3	206819529	5
Ratio	0.003097%	8.774231%	0.003348%	0.000017%	0.005692%	0.000266%	0.000004%	0.000001%	0.000102%	0.000005%	4.063921%	0.000006%	18.963337%	0.000001%	68.185971%	0.000002%

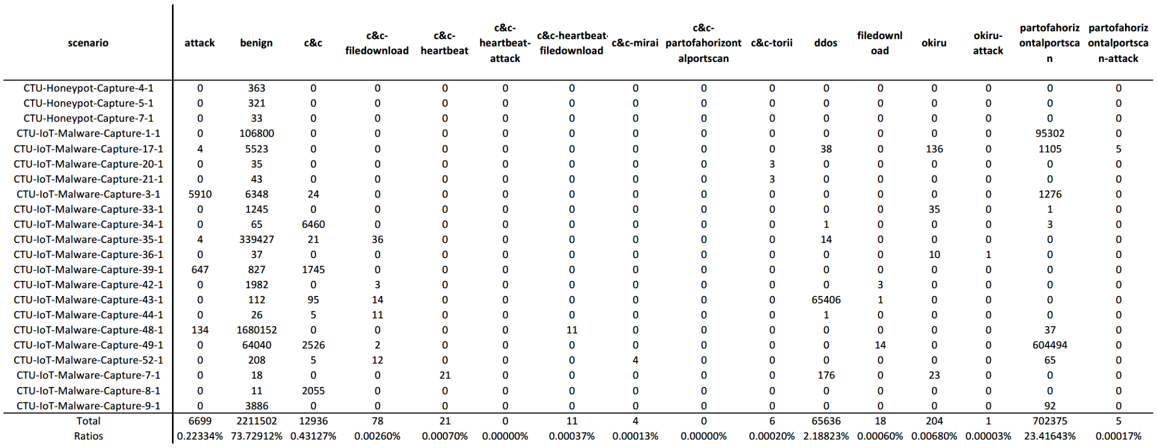
Figure 35: Detailed Label Distribution - Netflow Behavior

Filtered Datasets



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	14255	0	0	0	0	0	0	0	0	0	0	0	0	174015	0
CTU-IoT-Malware-Capture-17-1	0	93	0	0	1	0	0	0	0	0	17	0	49	0	141	0
CTU-IoT-Malware-Capture-20-1	0	21	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	31	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	148	94	6	0	0	0	0	0	0	0	0	0	0	0	10	0
CTU-IoT-Malware-Capture-33-1	0	267	0	0	1	0	0	0	0	0	0	0	31	0	0	0
CTU-IoT-Malware-Capture-34-1	0	18	0	0	0	0	0	0	0	0	3	0	0	0	1	0
CTU-IoT-Malware-Capture-35-1	4	99333	0	0	0	0	0	0	0	0	8	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	22	0	0	1	0	0	0	0	0	0	0	29	1	0	0
CTU-IoT-Malware-Capture-39-1	1	101	5	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	30	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	17	0	0	0	0	0	0	0	0	1	0	3	0	0	0
CTU-IoT-Malware-Capture-44-1	0	7	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	592	17	0	0	0	0	0	0	0	0	0	0	0	0	1676761	0
CTU-IoT-Malware-Capture-49-1	0	321	0	0	0	0	0	0	0	0	0	0	0	0	608129	0
CTU-IoT-Malware-Capture-52-1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	95	0
CTU-IoT-Malware-Capture-7-1	0	15	0	0	1	0	0	0	0	0	1	0	268	0	0	0
CTU-IoT-Malware-Capture-8-1	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	2636	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Total	745	117450	13	1	4	0	0	0	0	2	31	0	380	1	2459155	0
Ratio	0.02890%	4.55624%	0.00050%	0.00004%	0.00016%	0.00000%	0.00000%	0.00000%	0.00000%	0.00008%	0.00120%	0.00000%	0.01474%	0.00004%	95.39810%	0.00000%

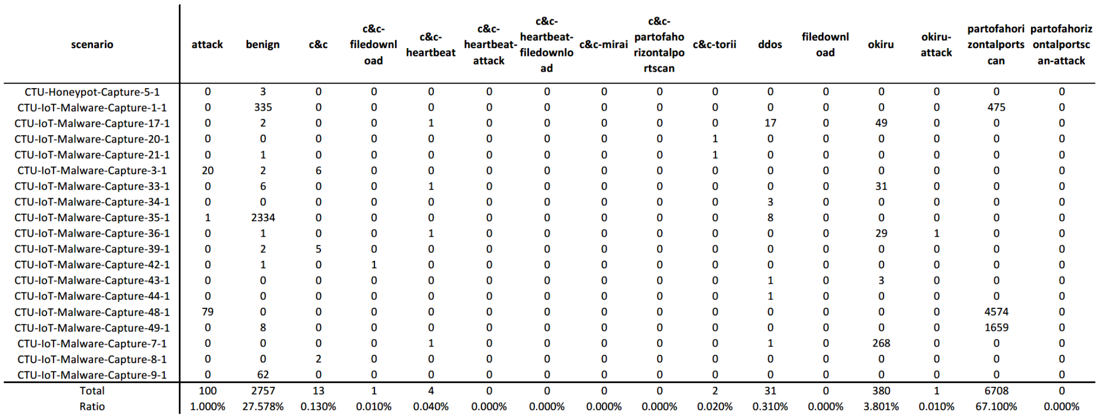
Figure 36: Detailed Label Distribution - Filtered Dataset - Unidirectional Behavior



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	363	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	321	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	106800	0	0	0	0	0	0	0	0	0	0	0	0	95302	0
CTU-IoT-Malware-Capture-17-1	4	5523	0	0	0	0	0	0	0	0	38	0	136	0	1105	5
CTU-IoT-Malware-Capture-20-1	0	35	0	0	0	0	0	0	0	3	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	43	0	0	0	0	0	0	0	3	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	5910	6348	24	0	0	0	0	0	0	0	0	0	0	0	1276	0
CTU-IoT-Malware-Capture-33-1	0	1245	0	0	0	0	0	0	0	0	0	0	35	0	1	0
CTU-IoT-Malware-Capture-34-1	0	65	6460	0	0	0	0	0	0	0	1	0	0	0	3	0
CTU-IoT-Malware-Capture-35-1	4	339427	21	36	0	0	0	0	0	0	14	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	37	0	0	0	0	0	0	0	0	0	0	10	1	0	0
CTU-IoT-Malware-Capture-39-1	647	827	1745	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	1982	0	3	0	0	0	0	0	0	0	3	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	112	95	14	0	0	0	0	0	0	65406	1	0	0	0	0
CTU-IoT-Malware-Capture-44-1	0	26	5	11	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	134	1680152	0	0	0	0	11	0	0	0	0	0	0	0	37	0
CTU-IoT-Malware-Capture-49-1	0	64040	2526	2	0	0	0	0	0	0	0	14	0	0	604494	0
CTU-IoT-Malware-Capture-52-1	0	208	5	12	0	0	0	4	0	0	0	0	0	0	65	0
CTU-IoT-Malware-Capture-7-1	0	18	0	0	21	0	0	0	0	0	176	0	23	0	0	0
CTU-IoT-Malware-Capture-8-1	0	11	2055	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	3886	0	0	0	0	0	0	0	0	0	0	0	0	92	0
Total	6699	2211502	12936	78	21	0	11	4	0	6	65636	18	204	1	702375	5
Ratios	0.22334%	73.72912%	0.43127%	0.00260%	0.00070%	0.00000%	0.00037%	0.00013%	0.00000%	0.00020%	2.18823%	0.00060%	0.00680%	0.00003%	23.41643%	0.00017%

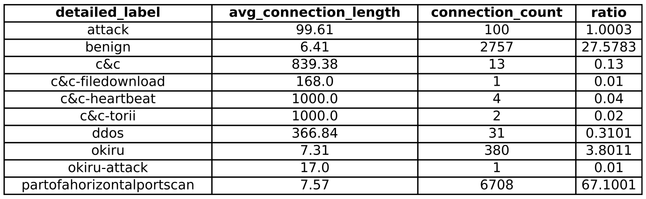
Figure 37: Detailed Label Distribution - Filtered Dataset - Netflow Behavior

Balanced Datasets - Unidirectional Behavior



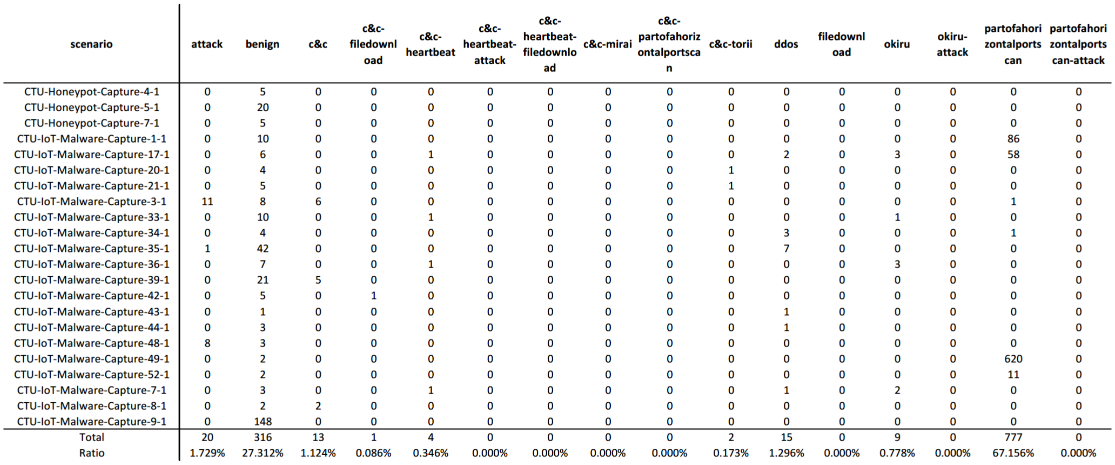
scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-5-1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	335	0	0	0	0	0	0	0	0	0	0	0	0	475	0
CTU-IoT-Malware-Capture-17-1	0	2	0	0	1	0	0	0	0	0	17	0	49	0	0	0
CTU-IoT-Malware-Capture-20-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	20	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-33-1	0	6	0	0	1	0	0	0	0	0	0	0	31	0	0	0
CTU-IoT-Malware-Capture-34-1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
CTU-IoT-Malware-Capture-35-1	1	2334	0	0	0	0	0	0	0	0	8	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	1	0	0	1	0	0	0	0	0	0	0	29	1	0	0
CTU-IoT-Malware-Capture-39-1	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0
CTU-IoT-Malware-Capture-44-1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	79	0	0	0	0	0	0	0	0	0	0	0	0	0	4574	0
CTU-IoT-Malware-Capture-49-1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	1659	0
CTU-IoT-Malware-Capture-7-1	0	0	0	0	1	0	0	0	0	0	1	0	268	0	0	0
CTU-IoT-Malware-Capture-8-1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	100	2757	13	1	4	0	0	0	0	2	31	0	380	1	6708	0
Ratio	1.000%	27.578%	0.130%	0.010%	0.040%	0.000%	0.000%	0.000%	0.000%	0.020%	0.310%	0.000%	3.801%	0.010%	67.100%	0.000%

Figure 38: Detailed Label Distribution - Unidirectional Behavior - 10000 Dataset



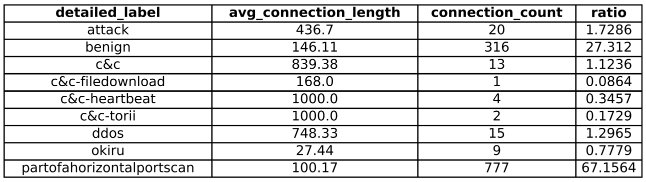
detailed_label	avg_connection_length	connection_count	ratio
attack	99.61	100	1.0003
benign	6.41	2757	27.5783
c&c	839.38	13	0.13
c&c-filedownload	168.0	1	0.01
c&c-heartbeat	1000.0	4	0.04
c&c-torii	1000.0	2	0.02
ddos	366.84	31	0.3101
okiru	7.31	380	3.8011
okiru-attack	17.0	1	0.01
partofahorizontalportscan	7.57	6708	67.1001

Figure 39: Detailed Label Connections Summary - Unidirectional Behavior - 10000 Dataset



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	10	0	0	0	0	0	0	0	0	0	0	0	0	86	0
CTU-IoT-Malware-Capture-17-1	0	6	0	0	1	0	0	0	0	0	2	0	3	0	58	0
CTU-IoT-Malware-Capture-20-1	0	4	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	11	8	6	0	0	0	0	0	0	0	0	0	0	0	1	0
CTU-IoT-Malware-Capture-33-1	0	10	0	0	1	0	0	0	0	0	0	0	1	0	0	0
CTU-IoT-Malware-Capture-34-1	0	4	0	0	0	0	0	0	0	0	3	0	0	0	1	0
CTU-IoT-Malware-Capture-35-1	1	42	0	0	0	0	0	0	0	0	7	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	7	0	0	1	0	0	0	0	0	0	0	3	0	0	0
CTU-IoT-Malware-Capture-39-1	0	21	5	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-44-1	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	8	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-49-1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	620	0
CTU-IoT-Malware-Capture-52-1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	11	0
CTU-IoT-Malware-Capture-7-1	0	3	0	0	1	0	0	0	0	0	1	0	2	0	0	0
CTU-IoT-Malware-Capture-8-1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	148	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	20	316	13	1	4	0	0	0	0	2	15	0	9	0	777	0
Ratio	1.729%	27.312%	1.124%	0.086%	0.346%	0.000%	0.000%	0.000%	0.000%	0.173%	1.296%	0.000%	0.778%	0.000%	67.156%	0.000%

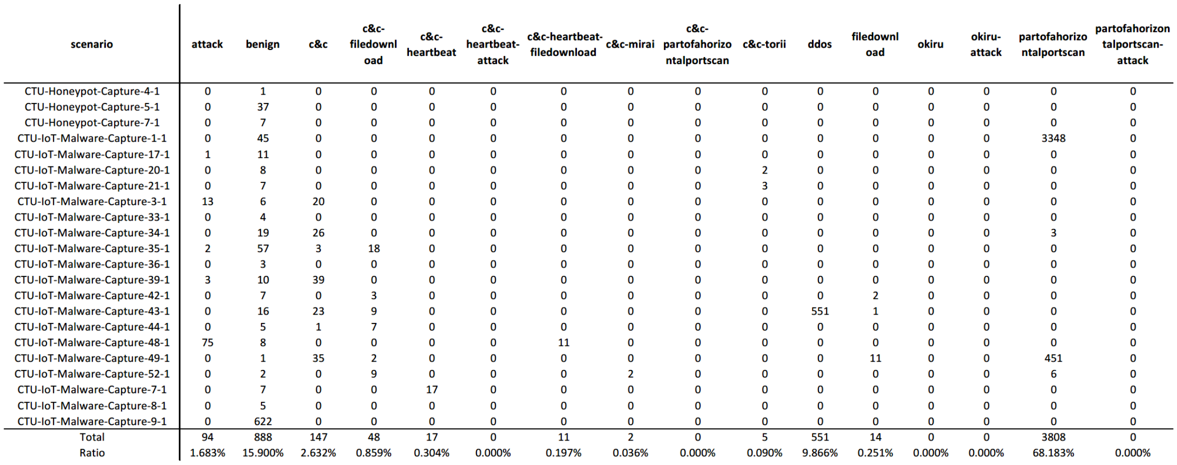
Figure 40: Detailed Label Distribution - Unidirectional Behavior - Min 20 Dataset



detailed_label	avg_connection_length	connection_count	ratio
attack	436.7	20	1.7286
benign	146.11	316	27.312
c&c	839.38	13	1.1236
c&c-filedownload	168.0	1	0.0864
c&c-heartbeat	1000.0	4	0.3457
c&c-torii	1000.0	2	0.1729
ddos	748.33	15	1.2965
okiru	27.44	9	0.7779
partofahorizontalportscan	100.17	777	67.1564

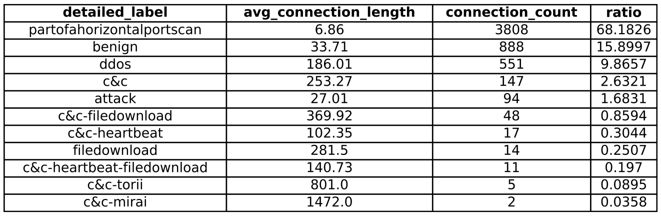
Figure 41: Detailed Label Connections Summary - Unidirectional Behavior - Min 20 Dataset

Balanced Datasets - Netflow Behavior



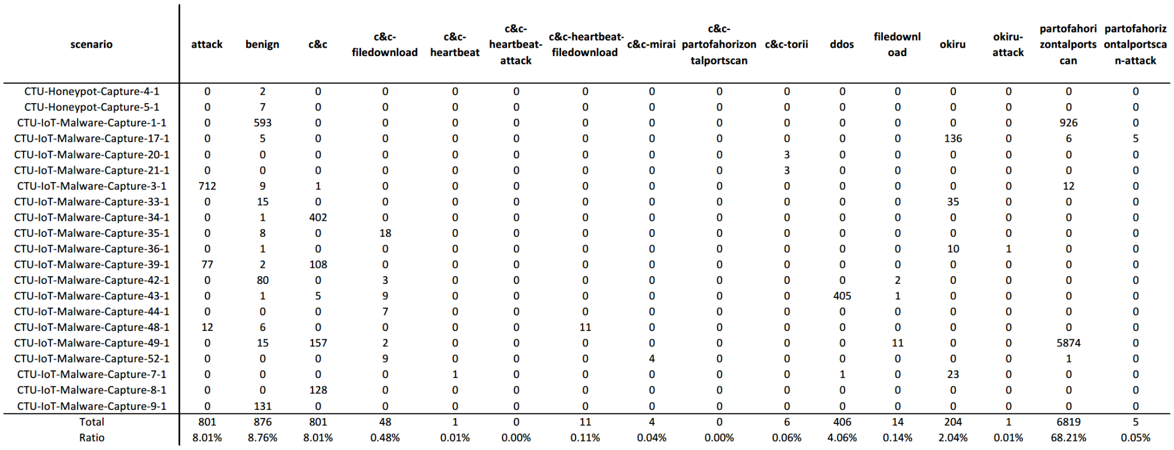
scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-7-1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	45	0	0	0	0	0	0	0	0	0	0	0	0	3348	0
CTU-IoT-Malware-Capture-17-1	1	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-20-1	0	8	0	0	0	0	0	0	0	2	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	7	0	0	0	0	0	0	0	3	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	13	6	20	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-33-1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-34-1	0	19	26	0	0	0	0	0	0	0	0	0	0	0	3	0
CTU-IoT-Malware-Capture-35-1	2	57	3	18	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-39-1	3	10	39	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	7	0	3	0	0	0	0	0	0	0	2	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	16	23	9	0	0	0	0	0	0	551	1	0	0	0	0
CTU-IoT-Malware-Capture-44-1	0	5	1	7	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	75	8	0	0	0	0	11	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-49-1	0	1	35	2	0	0	0	0	0	0	0	11	0	0	451	0
CTU-IoT-Malware-Capture-52-1	0	2	0	9	0	0	0	2	0	0	0	0	0	0	6	0
CTU-IoT-Malware-Capture-7-1	0	7	0	0	17	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-8-1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	622	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	94	888	147	48	17	0	11	2	0	5	551	14	0	0	3808	0
Ratio	1.683%	15.900%	2.632%	0.859%	0.304%	0.000%	0.197%	0.036%	0.000%	0.090%	9.866%	0.251%	0.000%	0.000%	68.183%	0.000%

Figure 42: Detailed Label Distribution - Netflow Behavior - Min 20 Dataset



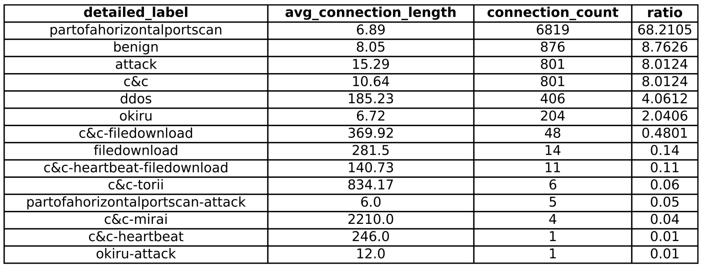
detailed_label	avg_connection_length	connection_count	ratio
partofahorizontalportscan	6.86	3808	68.1826
benign	33.71	888	15.8997
ddos	186.01	551	9.8657
c&c	253.27	147	2.6321
attack	27.01	94	1.6831
c&c-filedownload	369.92	48	0.8594
c&c-heartbeat	102.35	17	0.3044
filedownload	281.5	14	0.2507
c&c-heartbeat-filedownload	140.73	11	0.197
c&c-torii	801.0	5	0.0895
c&c-mirai	1472.0	2	0.0358

Figure 43: Detailed Label Connections Summary - Netflow Behavior - Min 20 Dataset



scenario	attack	benign	c&c	c&c-filedownload	c&c-heartbeat	c&c-heartbeat-attack	c&c-heartbeat-filedownload	c&c-mirai	c&c-partofahorizontalportscan	c&c-torii	ddos	filedownload	okiru	okiru-attack	partofahorizontalportscan	partofahorizontalportscan-attack
CTU-Honeypot-Capture-4-1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-Honeypot-Capture-5-1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-1-1	0	593	0	0	0	0	0	0	0	0	0	0	0	0	926	0
CTU-IoT-Malware-Capture-17-1	0	5	0	0	0	0	0	0	0	0	0	0	136	0	6	5
CTU-IoT-Malware-Capture-20-1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
CTU-IoT-Malware-Capture-21-1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
CTU-IoT-Malware-Capture-3-1	712	9	1	0	0	0	0	0	0	0	0	0	0	0	12	0
CTU-IoT-Malware-Capture-33-1	0	15	0	0	0	0	0	0	0	0	0	0	35	0	0	0
CTU-IoT-Malware-Capture-34-1	0	1	402	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-35-1	0	8	0	18	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-36-1	0	1	0	0	0	0	0	0	0	0	0	0	10	1	0	0
CTU-IoT-Malware-Capture-39-1	77	2	108	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-42-1	0	80	0	3	0	0	0	0	0	0	0	2	0	0	0	0
CTU-IoT-Malware-Capture-43-1	0	1	5	9	0	0	0	0	0	0	405	1	0	0	0	0
CTU-IoT-Malware-Capture-44-1	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-48-1	12	6	0	0	0	0	11	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-49-1	0	15	157	2	0	0	0	0	0	0	0	11	0	0	5874	0
CTU-IoT-Malware-Capture-52-1	0	0	0	9	0	0	0	4	0	0	0	0	0	0	1	0
CTU-IoT-Malware-Capture-7-1	0	0	0	0	1	0	0	0	0	0	1	0	23	0	0	0
CTU-IoT-Malware-Capture-8-1	0	0	128	0	0	0	0	0	0	0	0	0	0	0	0	0
CTU-IoT-Malware-Capture-9-1	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	801	876	801	48	1	0	11	4	0	6	406	14	204	1	6819	5
Ratio	8.01%	8.76%	8.01%	0.48%	0.01%	0.00%	0.11%	0.04%	0.00%	0.06%	4.06%	0.14%	2.04%	0.01%	68.21%	0.05%

Figure 44: Detailed Label Distribution - Netflow Behavior - 10000 Dataset



detailed_label	avg_connection_length	connection_count	ratio
partofahorizontalportscan	6.89	6819	68.2105
benign	8.05	876	8.7626
attack	15.29	801	8.0124
c&c	10.64	801	8.0124
ddos	185.23	406	4.0612
okiru	6.72	204	2.0406
c&c-filedownload	369.92	48	0.4801
filedownload	281.5	14	0.14
c&c-heartbeat-filedownload	140.73	11	0.11
c&c-torii	834.17	6	0.06
partofahorizontalportscan-attack	6.0	5	0.05
c&c-mirai	2210.0	4	0.04
c&c-heartbeat	246.0	1	0.01
okiru-attack	12.0	1	0.01

Figure 45: Detailed Label Connections Summary - Netflow Behavior - 10000 Dataset