Lockout-Tagout Ransomware: A Detection Method for Ransomware using Fuzzy Hashing and Clustering

Nitin Naik[1,2], Paul Jenkins[1,2], Jonathan Gillett[2], Haralambos Mouratidis[3], Kshirasagar Naik[4] and Jingping Song[5]

[1]School of Computing, University of Portsmouth, United Kingdom
[2]Faculty of Science and Technology, Bournemouth University, United Kingdom
[3]Centre for Secure, Intelligent and Usable Systems, University of Brighton, United Kingdom
[4]Department of Electrical and Computer Engineering, University of Waterloo, Canada
[5]Software College, Northeastern University, China
Email: {nitin.naik, paul.jenkins}@port.ac.uk, {nnaik, pjenkins, gillettj}@bournemouth.ac.uk,
h.mouratidis@brighton.ac.uk, snaik@uwaterloo.ca, songjp@swc.neu.edu.cn

***Abstract*—Ransomware attacks are a prevalent cybersecurity threat to every user and enterprise today. This is attributed to their polymorphic behaviour and dispersion of inexhaustible versions due to the same ransomware family or threat actor. A certain ransomware family or threat actor repeatedly utilises nearly same style or codebase to create a vast number of ransomware versions. Therefore, it is essential for users and enterprises to keep well-informed about this threat landscape and adopt proactive prevention strategies to minimise its spread and affects. This requires a technique to detect ransomware samples to determine the similarity and link with the known ransomware family or threat actor. Therefore, this paper presents a detection method for ransomware by employing a combination of a similarity preserving hashing method called fuzzy hashing and a clustering method. This detection method is applied on the collected WannaCry/WannaCryptor ransomware samples utilising a range of fuzzy hashing and clustering methods. The clustering results of various clustering methods are evaluated through the use of the internal evaluation indexes to determine the accuracy and consistency of their clustering results, thus the effective combination of fuzzy hashing and clustering method as applied to the particular ransomware corpus. The proposed detection method is a static analysis method, which requires fewer computational overheads and performs rapid comparative analysis with respect to other static analysis methods.**

***Index Terms*—Ransomware; Similarity Preserving Hashing; Fuzzy Hashing; SSDEEP; SDHASH; Clustering, K-Means; PAM; AGNES; DIANA, CLARA; WannaCry; WannaCryptor.**

I. Introduction

A ransomware attack is as an attempt to extort a user or enterprise by denying it access to its data by encrypting or locking it. Generally, ransomware encrypts data, however, it may apply different approach such as locking or erasing data. The concept of ransom related IT crime is an old one and was first discussed by *Donn Parker*, in the publication *Crime by Computer* [1] in 1976. However, ransomware infections have increased significantly in recent years, developing into one of the most significant and problematic cybercrime known, due to its polymorphism. WannaCry or WannaCryptor is such an example, having emerged in the last five years, causing a total loss of around $4 billion to both organisations and individuals [2], [3]. The extent of the loss by ransomware as demonstrated by research conducted by Cybersecurity Ventures which predicted the global damage due to ransomware would reach as high as $11.5 billion annually by 2019 [4]. The crux of this analysis is that "ransomware will have attacked a business every 14 seconds by the end of 2019" [4].

Detecting new or unknown ransomware requires a method that can process the ransomware corpus using unlabelled or generic labels (mostly mislabelled), satisfactorily [5]. Such detection methods are capable of finding matched samples and determining their degree of similarity, thus assisting further classification of samples into the most appropriate groups. In the first stage, fuzzy hashing can be used to find matched sample(s) with a degree of similarity [6]. Nonetheless, the classification of samples can be accomplished by either clustering or classification, but any classification technique necessitates accurate labelling of samples that is a tedious task because there is no universally accepted taxonomy [7]. Consequently, in the second stage, it is useful to cluster similar samples within the corpus [8]. Therefore, this paper presents a detection method for ransomware by employing a combination of a similarity preserving hashing method called fuzzy hashing and a clustering method. This detection method is applied on the collected WannaCry/WannaCryptor ransomware samples utilising fuzzy hashing methods SSDEEP [9], SDHASH [10] and clustering methods K-Mean [11], PAM [12], AGNES [13], DIANA [13], CLARA [12]. The clustering results of these clustering methods are evaluated through the use of three internal evaluation indexes, the Dunn Index [14], the Silhouette Index [15] and Connectivity Index [16], which are used to determine the accuracy and consistency of their clustering results, thus, the effective combination of fuzzy hashing and clustering method can be selected for the particular ransomware corpus. Later, the detection results can be used

for both advanced static and dynamic analysis of ransomware.

The paper is organised into the subsequent sections: Section II describes fuzzy hashing and its types SSDEEP and SDHASH methods; clustering and its types Partitioning-Based Clustering and Hierarchical Clustering. Section III explains the process of gathering WannaCry or WannaCryptor ransomware samples for the implementation of this proposed ransomware detection method. Section IV outlines the proposed detection method for ransomware using fuzzy hashing and clustering. Section V presents the experimental evaluation of the combination of different fuzzy hashing methods and clustering methods and their detection results. Finally, Section VI presents the summary of the paper and suggests some future enhancements.

II. Background Information

A. Fuzzy Hashing

In security analysis, hashing is used to determine both the integrity and similarity of files under examination, the latter utilising cryptographic techniques and the former utilising fuzzy techniques. In malware investigation, when attempting to determine malware strains it is the similarity of sample which is of interest as often malware developers utilise similar code, leading to different variants [9]. In this type of analysis, generally, a file is divided into multiple blocks and a hash value is calculated for each block, the final step being the concatenation of all hash values of the blocks to generate the fuzzy hash value as shown in Fig. 1. Several factors are involved in determining the length of fuzzy hash value including the block size, the file size, and the output size of the selected hash function [17]. In contrast the complete file is hashed in cryptographic hashing with the output hash having a fixed size irrespective of input file size. There are different categories of fuzzy hashing techniques, classified as follows: Context-Triggered Piecewise Hashing (CTPH), Statistically-Improbable Features (SIF), Block-Based Hashing (BBH) and Block-Based Rebuilding (BBR) [18], [19], [20]. The comparison of files in forensic analysis, where known malware files are compared with unknown samples for the purpose of triaging and clustering of malware to identify new variants, requires an understanding of the degree of similarity between samples. This suggests the use of the similarity preserving characteristic of fuzzy hashing which is effective in forensic investigation when comparing new samples with existing malware families for their triage and clustering, in samples which have the same functionality, yet the same cryptographic values [21].

Generally, the similarity of samples can be measured based upon their syntactic or semantic levels [21]. At a syntactic level, two files are compared to find similarity on the basis of their byte sequence of data but not the context of data. Whereas at semantic level, two files are compared to find similarity on the basis of their context [21]. Fuzzy hashing is only utilised to find similarity between two files at syntactic level.

1) SSDEEP: The SSDEEP fuzzy hashing method was initially developed for finding spam emails [9]. This method divides a file into number of blocks based on the content of that file. The endpoint points of these blocks are determined

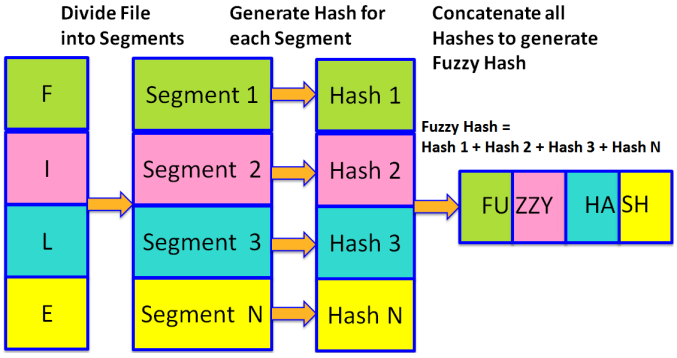

Fig. 1. Generation of Fuzzy Hash Value in Fuzzy Hashing Method

by a rolling hash method utilising the Adler32 function [17]. Generating the SSDEEP fuzzy hash value for the file, consists of calculating an individual hash value for each of block and concatenating these into a single hash value. Similarity between the two files is calculated by utilising Damerau-Levenshtein distance function.

2) SDHASH: The SDHSH fuzzy hashing method finds common and rare features in a file and matches the rare features in another file to determine the degree of similarity between the two files [10]. Generally a feature is a 64-byte string and is found using an entropy calculation. It employs the cryptographic hash function SHA-1 and Bloom filters to calculate the SDHASH fuzzy hash value of a file [22]. A Bloom filter is a space-efficient probabilistic data structure to find whether the element is definitely not present in the set or may be present in the set. Similarity between the two files is calculated by utilising a Hamming distance function.

B. Clustering

Clustering is a machine learning algorithm used to group data based upon their similarity or difference. It is utilised when no prior knowledge of the dataset is known, partitioning the data set and placing different objects into these groups based upon their similarity or difference.

1) Partitioning-Based Clustering: Partitioning based clustering is used to identify the number of partitions/groups in the dataset based on their similarity or difference. It is an iterative process beginning with random partitioning, relocating data items from one cluster to another in each subsequent iteration. Predominately, partitioning clustering methods require a pre-determined value for the number of clusters to find. The most commonly used partitioning clustering methods are k-means, k-mediods/pam and clara [11], [12].

2) Hierarchical Clustering: Hierarchical clustering is a substitution method of clustering used to identify similar groups in the dataset. Unlike partitioning clustering, it does not require a pre-determined cluster number to group the dataset. Commonly, hierarchical clustering results are illustrated in the form of tree structure called a dendrogram. It utilises a pairwise distance/proximity matrix between observations as clustering criteria. Hierarchical clustering is further classified into two main categories: agglomerative clustering and divisive

clustering. The most commonly used hierarchical clustering methods are agnes (agglomerative hierarchical) and diana (divisive hierarchical).

III. Collecting WannaCry/WannaCryptor Ransomware Samples

A Ransomware attack is a nefarious attack to extort money from victims which is a more sophisticated tactic than the DDoS attack to extort money [23], [24], [25]. It causes loss of money and reputational damage to the business and sometimes potentially permanent loss of data. Ransomware attack could be a minor or severe depending on the category of ransomware, nonetheless, certain types of ransomware are nefarious in terms of their intention. Such ransomware are the priority for this investigation such as WannaCry or WannaCryptor ransomware is one of the most significant variants of ransomware in recently and is selected for this study [3], [26], [27], [28]. The most labour intensive task was the collection of credible samples of the WannaCry ransomware. As a result of this process, it was decided to collect a reasonable number of WannaCry or WannaCryptor ransomware samples which could be easily investigated manually. All the WannaCry samples were gathered from two sources *Hybrid Analysis* [29] and *Malshare* [30] and their analysis were performed through the information acquired by *VirusTotal* [31]. The main difficulty was to verify the credibility of the collected ransomware samples that they were very likely to be WannaCry ransomware samples. The credibility of samples was evaluated through the criteria set on the basis of the result of various detection engines on VirusTotal, which was greater than or equal to 40, means minimum 40 detection engines on VirusTotal diagnosed the particular sample as ransomware/malware. To verify that they were WannaCry or WannaCryptor ransomware, they were manually checked on every detection engine, where a number of the engines identified a sample as a WannaCry or WannaCryptor ransomware. Nevertheless, this ransomware verification process was complex, and mainly dependent on the discretion of authors [32], [33], [34], [35], [36]. The selection process was lengthy and demanding, consequently, 112 samples of WannaCry or WannaCryptor ransomware were selected after each sample was fully analysed manually.

IV. Proposed Detection Method for Ransomware

The proposed detection method for ransomware is a two-stage process where, at the first stage, a fuzzy hashing method identifies the similarity amongst the samples and generates similarity scores, at the second stage a suitable clustering method is employed to organise similar samples into one group. However, this detection method may or may not require an additional stage of unpacking the ransomware samples using an unpacking tool; it is dependent on collected ransomware samples. If samples are not unpacked then it is the initial requisite for ransomware analysis. Fuzzy hashing is used for initial triaging, where it matches collected samples and determines the percentage similarity of samples with known samples if they are matched with any known sample. When

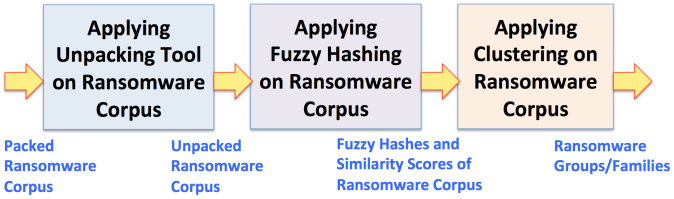

Fig. 2. A Detection Method for Ransomware

samples are matched, the similarity score can be further used in clustering operation to find their groups or families [37]. Later, based on these similarity scores, the closeness of the samples is computed to arrange them into similar groups by utilising the preferred clustering method [38]. Fuzzy hashing is one of the most efficient methods and requires fewer overheads for processing a significant number of ransomware samples. Essentially, Fuzzy hashing is beneficial due to its smaller hash size, resulting lower memory and computational resources as compared to alternative analysis techniques. Clustering is the natural choice when very little or no information is available about the new ransomware samples. The effective combination of a fuzzy hashing and a clustering method can be a respectable option for the initial analysis over other existing detection methods as it is tested in this paper. Subsequently, the detection results of this method can also be used for both advanced static and dynamic analysis of ransomware.

V. Experimental Evaluation of Different Clustering Methods and Their Detection Results

For the proposed detection method, five clustering methods *kmeans, pam, agnes, diana* and *clara* are analysed to determine the accuracy and consistency of their clustering results, thus, the most efficient clustering method can be selected. Clustering is an unsupervised technique as there are no standard labels available, therefore, it is generally evaluated by using internal indexes [38]. To avoid any bias in the results, three evaluation indexes are utilised to build an accurate assessment of the data and cluster quality for both fuzzy hashing and clustering methods. The three internal evaluation indexes Dunn Index, Silhouette Index and Connectivity Index are implemented in *clValid* package of **R** [16], [39], [40], [41], [40]. The Dunn Index has a value between *Zero* and ∞, here the greater value of the Dunn Index represents more accurate clustering results [14]. The Silhouette Index has a value in the interval $[-1, 1]$, here the greater value of the Silhouette Index represents more accurate clustering results [15]. The Connectivity Index has a value between *Zero* and ∞, here the smaller value of the Connectivity Index represents more accurate clustering results [16]. For all the clustering methods, the values of all the three indexes are computed for the range of clusters from 2 to 6 which is sufficient for a small dataset, checking the quality and consistency of the clustering results based on all these indexes. Finally the collective evaluation results of three internal clustering indexes for each clustering is compared for the optimal cluster size based on the ground truth.

A. Comparative Evaluation of Clustering Methods based on SSDEEP Similarity Scores

This experiment has utilised the SSDEEP similarity scores for clustering the collected WannaCry/WannaCryptor ransomware samples. The clustering results of all the clustering methods are given in Tables I to III and Fig. 3, for the range of clusters from 2 to 6 using three internal indexes Dunn, Silhouette and Connectivity. The evaluation results show that three clustering methods kmean, agnes and diana provide the best results for all three internal indexes but for different cluster size. In case of the Dunn index and Connectivity indexes, the cluster size is two as shown in Tables I and III respectively, whereas for the Silhouette index, the cluster size is six as shown in Table II. Later, the collective evaluation results of three internal clustering indexes for each clustering is compared for the cluster size two (see Table IV), which is the optimal cluster size based on the ground truth. This optimal cluster size is determined based on the manual analysis of WannaCry/WannaCryptor ransomware samples. The collective evaluation results show that similar three clustering methods kmeans, agnes (agglomerative hierarchical) and diana (divisive hierarchical) performed well in comparison to others clustering methods.

B. Comparative Evaluation of Clustering Methods based on SDHASH Similarity Scores

This experiment has utilised SDHASH similarity scores for clustering the collected WannaCry/WannaCryptor ransomware samples. The clustering results of all the clustering methods are given in Tables V to VII and Fig. 4, for the range of clusters from 2 to 6 using three internal indexes Dunn, Silhouette and Connectivity. The evaluation results show that there is no single method providing the best results for all three internal indexes. In case of the Dunn index and Silhouette indexes, three clustering methods kmean, agnes and diana have produced the best result for the cluster size three as shown in Tables V and VI respectively. While for Connectivity index, two clustering methods pam and clara have produced the best result for the cluster size two as shown in VII. Later, the collective evaluation results of three internal clustering indexes for each clustering is compared for the cluster size two (see Table VIII), which is the optimal cluster size based on the ground truth. As mentioned earlier, this optimal cluster size is determined based on the manual analysis of WannaCry/WannaCryptor ransomware samples. The collective evaluation results show that two clustering methods agnes (agglomerative hierarchical) and diana (divisive hierarchical) performed well in comparison to other clustering methods.

Both clustering methods agnes (agglomerative hierarchical) and diana (divisive hierarchical) have one similarity that they are a hierarchical clustering method, thus, it suggests that hierarchical clustering is the most suitable method for both fuzzy hashing methods SSDEEP and SDHASH. Nonetheless, agglomerative clustering is good at identifying small clusters, while divisive clustering is good at identifying large clusters. Depending on the nature of data and number of clusters, a preferred hierarchical clustering method can be selected.

C. Comparative Evaluation of SSDEEP and SDHASH Fuzzy Hashing Methods

In this detection method, the first stage of fuzzy hashing is crucial for generating improved clustering results and detection rates. Therefore, the results of the two fuzzy hashing methods SSDEEP and SDHASH are compared to evaluate their effectiveness for the ransomware corpus. As stated previously, all the 112 WannaCry ransomware samples were methodically confirmed as a ransomware sample, therefore, this experiment does not consider the event of false positive for both SSDEEP and SDHASH methods. Consequently, the main aim of this experiment was to detect the similarity among WannaCry/WannaCryptor ransomware samples and record the degree of similarity amongst them. Both SSDEEP and SDHASH methods performed well and detected similarity for the majority of the WannaCry/WannaCryptor ransomware samples. However, the detection results of the two fuzzy hashing methods indicate two main differences between them. Firstly, SDHASH detected similarity in more samples (108/112) than SSDEEP (104/112). Secondly, SDHASH detection results showed several insignificant values in the degree of similarity between the samples whereas SSDEEP detection results indicated largely significant values in the degree of similarity between the samples. The similarity scores of a fuzzy hashing method was the only key factor affecting the clustering results for this detection method, which is reflected in the evaluation of the clustering results based on these two methods. Both fuzzy hashing methods are completely different with respect to their working and both have their advantages and limitations as SDHASH can detect similarity in large samples but that may affect the clustering results due to weak similarity scores, whereas SSDEEP may not detect as many similar samples as SDHASH but can generate a different clustering results due to the consideration of only strong similarity scores.

VI. Conclusion

This paper proposed an efficient detection method for ransomware by employing a combination of a fuzzy hashing and a clustering method. This detection method applied on the collected WannaCry/WannaCryptor ransomware samples utilising fuzzy hashing methods SSDEEP, SDHASH and clustering methods K-Mean, PAM, AGNES, DIANA, CLARA. The clustering results of all these clustering methods were evaluated through the use of three internal evaluation indexes Dunn Index, Silhouette Index and Connectivity Index, to determine the accuracy and consistency of their clustering results, thus, the most effective combination of fuzzy hashing and clustering method can be selected for the particular ransomware corpus. Later, the detection results can be used for both advanced static and dynamic analysis of ransomware. The clustering results showed that two clustering methods AGNES and DIANA performed well in comparison to other methods. Moreover, both SSDEEP and SDHASH methods performed well and

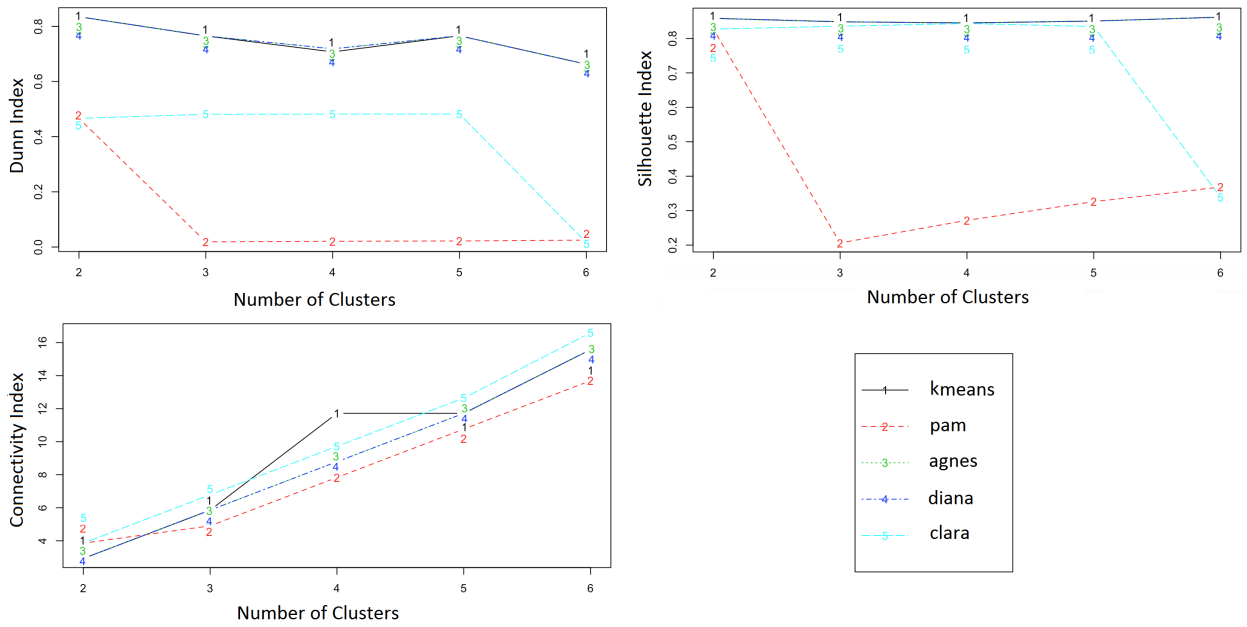

Fig. 3. Dunn Index, Silhouette Index and Connectivity Index Graphs for Different Clustering Methods based on SSDEEP Similarity Scores for WannaCry Ransomware

TABLE I
DUNN INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SSDEEP FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	**Cluster Size=2**	**Cluster Size=3**	**Cluster Size=4**	**Cluster Size=5**	**Cluster Size=6**
kmeans	**0.8348**	0.7653	0.7074	0.7658	0.6628
pam	0.4661	0.0189	0.021	0.0219	0.0249
agnes	**0.8348**	0.7653	0.7185	0.7658	0.6628
diana	**0.8348**	0.7653	0.7185	0.7658	0.6628
clara	0.4661	0.4815	0.4818	0.4818	0.0164

Note: The Dunn Index has a value between *Zero* and ∞, where the greater value of the Dunn Index represents more accurate clustering results [14].

TABLE II
SILHOUETTE INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SSDEEP FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	**Cluster Size=2**	**Cluster Size=3**	**Cluster Size=4**	**Cluster Size=5**	**Cluster Size=6**
kmeans	0.8584	0.8483	0.8451	0.8504	**0.8618**
pam	0.8271	0.2063	0.272	0.3261	0.3685
agnes	0.8584	0.8483	0.8447	0.8504	**0.8618**
diana	0.8584	0.8483	0.8447	0.8504	**0.8618**
clara	0.8271	0.8356	0.8431	0.8346	0.3388

Note: The Silhouette Index has a value in the interval $[-1, 1]$, where the greater value of the Silhouette Index represents more accurate clustering results [15].

TABLE III
CONNECTIVITY INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SSDEEP FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	Cluster Size=2	Cluster Size=3	Cluster Size=4	Cluster Size=5	Cluster Size=6
kmeans	**2.929**	5.8579	11.7159	11.7159	15.5738
pam	3.8579	4.9052	7.8341	10.7631	13.6921
agnes	**2.929**	5.8579	8.7869	11.7159	15.5738
diana	**2.929**	5.8579	8.7869	11.7159	15.5738
clara	3.8579	6.7869	9.7159	12.6448	16.5956

Note: The Connectivity Index has a value between *Zero* and ∞, where the smaller value of the Connectivity Index represents more accurate clustering results [16].

TABLE IV
COLLECTIVE EVALUATION RESULTS FOR ALL THE INDEXES FOR DIFFERENT CLUSTERING METHODS BASED ON THE SSDEEP FUZZY HASHING METHOD FOR CLUSTER SIZE 2 (WHICH IS THE OPTIMAL CLUSTER SIZE BASED ON THE GROUND TRUTH)

Clustering Index	kmeans	pam	agnes	diana	clara
Dunn Index	**0.8348**	0.4661	**0.8348**	**0.8348**	0.4661
Silhouette Index	**0.8584**	0.8271	**0.8584**	**0.8584**	0.8271
Connectivity Index	**2.929**	3.8579	**2.929**	**2.929**	3.8579

Note: Here *kmeans, agnes* and *diana* have produced the most accurate results for this particular experiment based on the ground truth.

TABLE V
DUNN INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SDHASH FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	Cluster Size=2	Cluster Size=3	Cluster Size=4	Cluster Size=5	Cluster Size=6
kmeans	0.7073	**5.8244**	0.3756	0.2502	0.0959
pam	0.0284	0.0397	0.3756	0.0729	0.1427
agnes	0.9756	**5.8244**	0.3756	0.3756	0.3085
diana	0.9756	**5.8244**	0.3756	0.2429	0.3163
clara	0.0284	0.0397	0.3756	0.1445	0.1479

Note: The Dunn Index has a value between *Zero* and ∞, where the greater value of the Dunn Index represents more accurate clustering results [14].

detected similarity for most of the WannaCry/WannaCryptor ransomware samples. However, SDHASH detected similarity in more samples including several insignificant similarity scores, whereas SSDEEP could not detect as many similar samples as SDHASH with only considered strong similarity scores. This requires further investigation to determine their effects on the accuracy of clustering results. Moreover, in the future, it is essential to evaluate the proposed detection method on a larger sample of WannaCry/WannaCryptor ransomware and on other types of ransomware.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of *Hybrid-Analysis.com, Malshare.com* and *VirusTotal.com* for this research work.

REFERENCES

[1] D. B. Parker and D. Parker, *Crime by computer*. Scribner New York, 1976.

[2] S. Cobb. (2018) RANSOMWARE: An enterprise perspective. [Online]. Available: https://www.eset.com/us/business/resources/white-papers/ransomware-an-enterprise-perspective/

[3] J. M. Ehrenfeld, "Wannacry, cybersecurity and health information technology: A time to act," *Journal of medical systems*, vol. 41, no. 7, p. 104, 2017.

[4] S. Morgan. (2018) Global Ransomware damage costs predicted to exceed $8 Billion in 2018. [Online]. Available: https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-exceed-8-billion-in-2018

[5] L. Nataraj. (2013) Clustering a Malware Corpus. [Online]. Available: https://sarvamblog.blogspot.com/2013/04/clustering-malware-corpus.html

[6] N. Naik, P. Jenkins, and N. Savage, "A ransomware detection method using fuzzy hashing for mitigating the risk of occlusion of information systems," in *2019 IEEE International Symposium on Systems Engineering (ISSE)*, 2019.

[7] P. Li, L. Liu, D. Gao, and M. K. Reiter, "On challenges in evaluating malware clustering," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2010, pp. 238–255.

[8] Y. Li, S. C. Sundaramurthy, A. G. Bardas, X. Ou, D. Caragea, X. Hu, and J. Jang, "Experimental study of fuzzy hashing in malware clustering analysis," in *8th Workshop on Cyber Security Experimentation and Test*, vol. 5, no. 1. USENIX Association Washington, DC, 2015, p. 52.

[9] J. Kornblum, "Identifying almost identical files using context triggered piecewise hashing," *Digital investigation*, vol. 3, pp. 91–97, 2006.

[10] V. Roussev, "Data fingerprinting with similarity digests," in *IFIP International Conference on Digital Forensics*. Springer, 2010, pp. 207–226.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on*

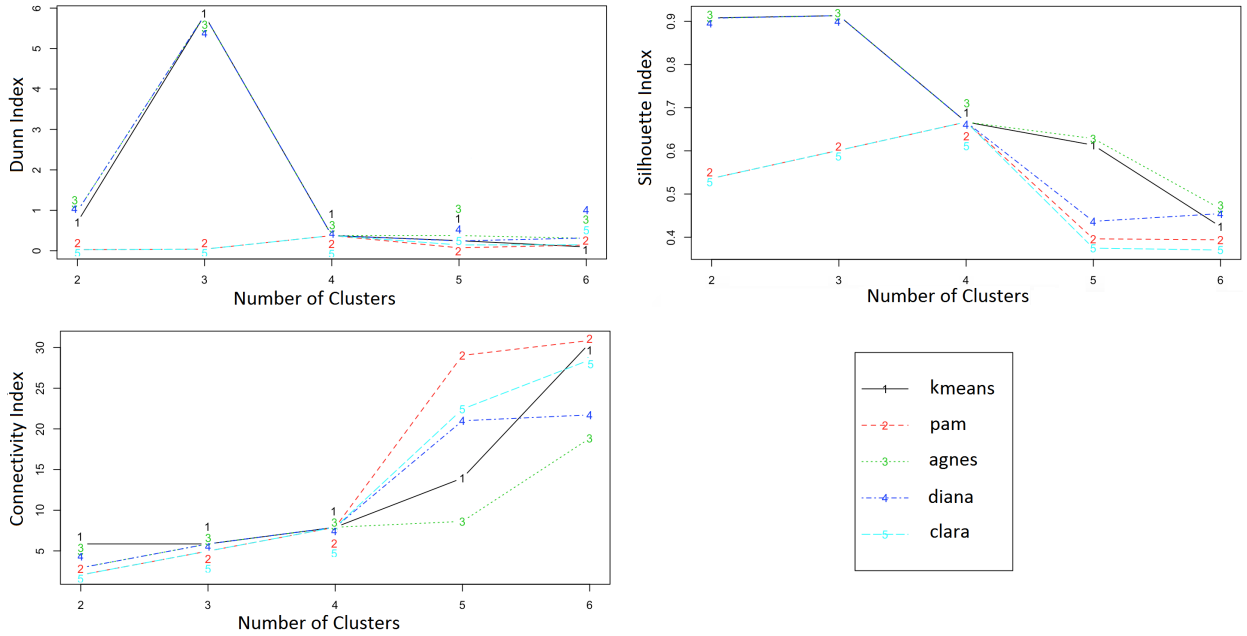

Fig. 4. Dunn Index, Silhouette Index and Connectivity Index Graphs for Different Clustering Methods based on SDHASH Similarity Scores for WannaCry Ransomware

TABLE VI
SILHOUETTE INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SDHASH FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	**Cluster Size=2**	**Cluster Size=3**	**Cluster Size=4**	**Cluster Size=5**	**Cluster Size=6**
kmeans	0.9079	**0.9131**	0.6668	0.6132	0.4244
pam	0.5362	0.6012	0.6668	0.3963	0.394
agnes	0.9072	**0.9131**	0.6668	0.6285	0.4658
diana	0.9072	**0.9131**	0.6668	0.4367	0.4546
clara	0.5362	0.6012	0.6668	0.3747	0.3703

Note: The Silhouette Index has a value in the interval $[-1, 1]$, where the greater value of the Silhouette Index represents more accurate clustering results [15].

TABLE VII
CONNECTIVITY INDEX EVALUATION RESULTS FOR DIFFERENT CLUSTERING METHODS BASED ON THE SDHASH FUZZY HASHING METHOD FOR WANNACRY RANSOMWARE CORPUS

Clustering Method	**Cluster Size=2**	**Cluster Size=3**	**Cluster Size=4**	**Cluster Size=5**	**Cluster Size=6**
kmeans	5.8579	5.8579	7.9075	13.9548	30.5365
pam	**2.0496**	4.9786	7.9075	29.0361	30.8623
agnes	2.929	5.8579	7.9075	8.6298	18.8377
diana	2.929	5.8579	7.9075	20.9992	21.7214
clara	**2.0496**	4.9786	7.9075	22.4321	28.4794

Note: The Connectivity Index has a value between *Zero* and ∞, where the smaller value of the Connectivity Index represents more accurate clustering results [16].

TABLE VIII
COLLECTIVE EVALUATION RESULTS OF ALL THE INDEXES FOR DIFFERENT CLUSTERING METHODS BASED ON THE SDHASH FUZZY HASHING METHOD FOR CLUSTER SIZE 2 (WHICH IS THE OPTIMAL CLUSTER SIZE BASED ON THE GROUND TRUTH)

Clustering Index	**kmeans**	**pam**	**agnes**	**diana**	**clara**
Connectivity Index	5.8579	2.0496	**2.929**	**2.929**	2.0496
Dunn Index	0.7073	0.0284	**0.9756**	**0.9756**	0.0284
Silhouette Index	0.9079	0.5362	**0.9072**	**0.9072**	0.5362

Note: Both *agnes (agglomerative hierarchical)* and *diana (divisive hierarchical)* are a type of hierarchical clustering, which has produced the most accurate results for this particular experiment.

mathematical statistics and probability, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[12] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, Inc, 1990.

[13] A. Struyf, M. Hubert, P. Rousseeuw *et al.*, "Clustering in an object-oriented environment," *Journal of Statistical Software*, vol. 1, no. 4, pp. 1–30, 1997.

[14] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[15] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[16] G. Brock, V. Pihur, S. Datta, S. Datta *et al.*, "clValid, an R package for cluster validation," *Journal of Statistical Software (Brock et al., March 2008)*, 2011.

[17] A. Tridgell, "Efficient algorithms for sorting and synchronization," Ph.D. dissertation, Australian National University Canberra, 1999.

[18] F. Breitinger and H. Baier, "A fuzzy hashing approach based on random sequences and hamming distance," in *Annual ADFSL Conference on Digital Forensics, Security and Law. 15*, 2012. [Online]. Available: https://commons.erau.edu/adfsl/2012/wednesday/15

[19] C. Sadowski and G. Levin, "Simhash: Hash-based similarity detection," 2007. [Online]. Available: www.webrankinfo.com/dossiers/wp-content/uploads/simhash.pdff

[20] V. Gayoso Martínez, F. Hernández Álvarez, and L. Hernández Encinas, "State of the art in similarity preserving hashing functions," 2014. [Online]. Available: http://digital.csic.es/bitstream/10261/135120/1/Similarity_preserving_Hashing_functions.pdf

[21] N. Naik, P. Jenkins, N. Savage, and L. Yang, "Cyberthreat Hunting-Part 1: Triaging Ransomware using Fuzzy Hashing, Import Hashing and YARA Rules," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019.

[22] V. Roussev, "An evaluation of forensic similarity hashes," *digital investigation*, vol. 8, pp. S34–S41, 2011.

[23] N. Naik, P. Jenkins, R. Cooke, D. Ball, A. Foster, and Y. Jin, "Augmented windows fuzzy firewall for preventing denial of service attack," in *2017 IEEE International Conference on Fuzzy Systems*, 2017, pp. 1–6.

[24] N. Naik and P. Jenkins, "Fuzzy reasoning based windows firewall for preventing denial of service attack," in *IEEE International Conference on Fuzzy Systems*, 2016, pp. 759–766.

[25] ——, "Enhancing windows firewall security using fuzzy reasoning," in *IEEE International Conference on Dependable, Autonomic and Secure Computing*, 2016, pp. 263–269.

[26] R. Richardson and M. North, "Ransomware: Evolution, mitigation and prevention," *International Management Review*, vol. 13, no. 1, pp. 10–21, 2017.

[27] K. Cabaj, P. Gawkowski, K. Grochowski, and D. Osojca, "Network activity analysis of Cryptowall ransomware," *Przeglad Elektrotechniczny*, vol. 91, no. 11, pp. 201–204, 2015.

[28] Y. Klijnsma. (2019) The history of Cryptowall: a large scale cryptographic ransomware threat. [Online]. Available: https://www.cryptowalltracker.org/

[29] Hybrid-Analysis. (2019) Hybrid Analysis. [Online]. Available: https://www.hybrid-analysis.com/

[30] Malshare. (2019) A free Malware repository providing researchers access to samples, malicious feeds, and YARA results. [Online]. Available: https://malshare.com/index.php

[31] VirusTotal. (2019) Virustotal. [Online]. Available: https://www.virustotal.com/#/home/upload

[32] N. Naik, P. Jenkins, B. Kerby, J. Sloane, and L. Yang, "Fuzzy logic aided intelligent threat detection in cisco adaptive security appliance 5500 series firewalls," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[33] N. Naik, P. Jenkins, R. Cooke, and L. Yang, "Honeypots that bite back: A fuzzy technique for identifying and inhibiting fingerprinting attacks on low interaction honeypots," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[34] N. Naik, P. Jenkins, and N. Savage, "Threat-aware honeypot for discovering and predicting fingerprinting attacks using principal components analysis," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018.

[35] N. Naik and P. Jenkins, "A fuzzy approach for detecting and defending against spoofing attacks on low interaction honeypots," in *21st International Conference on Information Fusion*. IEEE, 2018, pp. 904–910.

[36] ——, "Discovering hackers by stealth: Predicting fingerprinting attacks on honeypot systems," in *2018 IEEE International Symposium on Systems Engineering (ISSE)*, 2018.

[37] N. Naik, P. Jenkins, N. Savage, L. Yang, K. Naik, and J. Song, "Augmented YARA rules fused with fuzzy hashing in ransomware triaging," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019.

[38] N. Naik, P. Jenkins, N. Savage, and L. Yang, "Cyberthreat Hunting-Part 2: Tracking Ransomware Threat Actors using Fuzzy Hashing and Fuzzy C-Means Clustering," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019.

[39] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, "Package CLUSTER- Finding groups in data: Cluster analysis," *CRAN R studio*, 2018.

[40] C. Fraley and A. E. Raftery, "Model-based methods of classification: using the MCLUST software in chemometrics," *Journal of Statistical Software*, vol. 18, no. 6, pp. 1–13, 2007.

[41] R. Wehrens, "KOHONEN: Supervised and Unsupervised Self-Organising Maps," *R package version*, vol. 2, no. 2, 2007.