Cyberthreat Hunting - Part 2: Tracking Ransomware Threat Actors using Fuzzy Hashing and Fuzzy C-Means Clustering

Nitin Naik[1], Paul Jenkins[1], Nick Savage[2] and Longzhi Yang[3]

[1]Defence School of Communications and Information Systems, Ministry of Defence, United Kingdom
[2]School of Computing, University of Portsmouth, United Kingdom
[3]Department of Computer and Information Sciences, Northumbria University, United Kingdom
Email: nitin.naik100@mod.gov.uk, paul.jenkins683@mod.gov.uk, nick.savage@port.ac.uk, longzhi.yang@northumbria.ac.uk

***Abstract*—Threat actors are constantly seeking new attack surfaces, with ransomeware being one the most successful attack vectors that have been used for financial gain. This has been achieved through the dispersion of unlimited polymorphic samples of ransomware whilst those responsible evade detection and hide their identity. Nonetheless, every ransomware threat actor adopts some similar style or uses some common patterns in their malicious code writing, which can be significant evidence contributing to their identification. The first step in attempting to identify the source of the attack is to cluster a large number of ransomware samples based on very little or no information about the samples, accordingly, their traits and signatures can be analysed and identified. Therefore, this paper proposes an efficient fuzzy analysis approach to cluster ransomware samples based on the combination of two fuzzy techniques fuzzy hashing and fuzzy c-means (FCM) clustering. Unlike other clustering techniques, FCM can directly utilise similarity scores generated by a fuzzy hashing method and cluster them into similar groups without requiring additional transformational steps to obtain distance among objects for clustering. Thus, it reduces the computational overheads by utilising fuzzy similarity scores obtained at the time of initial triaging of whether the sample is known or unknown ransomware. The performance of the proposed fuzzy method is compared against k-means clustering and the two fuzzy hashing methods SSDEEP and SDHASH which are evaluated based on their FCM clustering results to understand how the similarity score affects the clustering results.**

***Index Terms*—Ransomware; Similarity Preserving; Fuzzy Hashing; SSDEEP; SDHASH; Fuzzy C-means Clustering; FCM; WannaCry; WannaCryptor; Locky; Cerber; CryptoWall; Triaging; Context-Triggered Piecewise Hashing; CTPH.**

I. Introduction

Ransomware threat actors are using ransomware as a weapon to attack cyber infrastructure and exploit users financially. They conceal their identity by dispersing multiple polymorphic instances of ransomware, thereby presenting an advanced persistent threat [1], [2]. There are several categories of ransomware, which pose a severe threat such as WannaCry/WannaCryptor, Locky, Cerber, CryptoWall, Petya, Notpetya, GandCrab, Bad Rabbit and CryptoLocker [2], [3], [4]. However, every ransomware threat actor adopts some similar style or uses some common patterns in their malicious code writing [2]; therefore, through the use of classification or clustering techniques, these ransomware threat actors/groups or new ransomware families can be discovered, based on the information mined from the particular corpus of ransomware [5]. However, the success of any classification or clustering is dependent on the chosen metric for calculating the similarity distance or score amongst the objects [6]. As presented in [7], fuzzy hashing (in particular SDHASH) yielded superior triaging results for various ransomware corpora, therefore, this can be considered as a preferred similarity metric for any classification or clustering method.

Normally, most malware (here ransomware) corpus comes with unlabelled or generic labels (mostly mislabelled) [8]. However, the correct classification of any corpus requires precise labels and characteristics, which can be difficult to achieve due to the lack of a globally accepted ground truth [9]. Consequently, it is useful to cluster similar samples within the corpus as a pre-processing step for any advanced analysis or to improve the classification accuracy [10]. Therefore, this paper proposes an efficient fuzzy analysis approach to cluster ransomware, based on the combination of two fuzzy techniques: a fuzzy hashing method [11], [12] and fuzzy c-means (FCM) clustering method [13], [14], [15], [16]. This combination of two fuzzy methods reduces the computational overheads by utilising fuzzy similarity scores obtained at the time of initial triaging of whether the sample is known or unknown ransomware.

As this is the first time, the combination of fuzzy hashing and fuzzy c-means clustering methods are used together for ransomware clustering, naturally, it is compared with the most common k-means clustering method to determine the clustering performance of the proposed FCM based fuzzy analysis approach. Furthermore, the two fuzzy hashing methods SSDEEP and SDHASH are evaluated based on their FCM clustering results to understand how the similarity score affects the clustering results. Evaluating the clustering results is a challenging task due to different perspectives and the unavailability of standard labels. Thus, for the rigorous comparison,

four evaluation metrics (particularly for FCM) Fuzzy Silhouette Index, Partition Coefficient, Modified Partition Coefficient and Partition Entropy are calculated and compared for both SSDEEP and SDHASH based FCM clustering method.

The background information and data collection process is already explained in the first part of the paper [7]. The rest of the paper is divided into the following sections: Section II presents the proposed fuzzy analysis approach for the analysis of ransomware corpus. Section III presents the experimental evaluation of FCM clustering and k-means clustering results. Section IV presents the experimental evaluation of SSDEEP and SDHASH based FCM clustering results. Finally, Section V concludes the paper with the possible future enhancement.

II. Proposed Fuzzy Analysis Approach for Ransomware

The proposed fuzzy analysis approach to cluster ransomware is based on the combination of two fuzzy techniques: a fuzzy hashing method [11], [12] and FCM clustering method [13], [14], [15], [16], in which any fuzzy hashing method and any variation of FCM can be employed. Furthermore, as a pre-processing step - unpacking the ransomware corpus may or may not require the use of an unpacking tool depending on whether the ransomware corpus is already unpacked or not, which is normally the preliminary requirement for most of malware analysis [17], [18], [19], [20]. Examining thousands of samples of ransomware families, they exhibit various degrees of similarity with each other and identifying their degree of similarity using a fuzzy hashing method would be beneficial for clustering and directly placing into similar (classification) groups albeit with little information available. Unlike other clustering techniques, FCM can directly utilise similarity scores generated by a fuzzy hashing method and cluster them into similar groups without requiring an additional transformational step to obtain distance amongst objects for clustering. Thus, it reduces the computational overheads by utilising fuzzy similarity scores obtained at the time of initial triaging whether the sample is known or unknown ransomware. Moreover, a fuzzy hashing method performs two tasks grouping the samples into known and unknown categories and assigning the fuzzy similarity score to the known samples in the range of 0 to 1 (or 0 to 100%), later this value is directly utilised as the degree of membership for the FCM process and used to expedite the clustering operation. Most fuzzy hashes are compact in size, which can save memory and other computational resources in comparison to other analysis methods.

The initial triaging results of four ransomware corpora using SSDEEP and SDHASH fuzzy hashing methods are the basis of FCM clustering which were analysed in the first part of the paper [7]. The employed dataset of 200 samples of four categories of ransomware WannaCry, Locky, Cerber and CryptoWall with 50 samples of each was collected from two sources *Hybrid Analysis* [21] and *Malshare* [22]. The verification of samples and their further analysis is performed

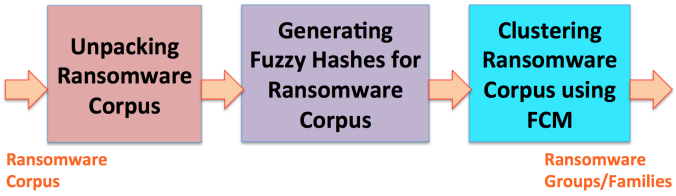

Fig. 1. Fuzzy Analysis Approach for the analysis of Ransomware Corpus

based on the information obtained from *VirusTotal* [23] and mostly based on the discretion of authors [24], [25], [26], [27].

III. Experimental Evaluation of Fuzzy C-Means Clustering and K-Means Clutering Results

Evaluating the performance of FCM clustering, based on SSDEEP and SDHASH fuzzy similarity scores, its clustering results were compared with k-means clustering utilising the same SSDEEP and SDHASH fuzzy similarity scores. This assessment was based on the measurement of the internal accuracy of clusters by employing the Silhouette Coefficient which is commonly used for the evaluation a clustering method. The average Silhouette Coefficient was computed utilising Euclidean Distance for both FCM and k-means clustering methods, where the greater value of Silhouette Coefficient represents more accurate clustering results. Both k-means [28] and FCM [29] clustering methods are implemented in R.

A. Comparative Evaluation of FCM Clustering and K-Means Clustering based on SSDEEP Similarity Scores

In the first evaluation, all the four SSDEEP results of WannaCry, Locky, Cerber and CryptoWall ransomware [7] were utilised for both FCM and k-means clustering methods. The average Silhouette Coefficient was computed for the range of clusters from 2 to 7 checking the consistency in performance of FCM and k-means clustering methods. The comparative results of the average Silhouette Coefficient based on SSDEEP similarity scores for the four ransomware categories are shown in Table I. Overall, for the four ransomware categories, FCM clustering generated improved clustering results in comparison to k-means clustering with only a few exceptions. This reflects the natural alignment of the two fuzzy methods utilised together.

B. Comparative Evaluation of FCM Clustering and K-Means Clustering based on SDHASH Similarity Scores

In the second evaluation, all four SDHASH results of WannaCry, Locky, Cerber and CryptoWall ransomware [7] were utilised for both FCM and k-means clustering methods. Similarly, the average Silhouette Coefficient was computed for the same range of clusters from 2 to 7 checking the consistency in performance of FCM and k-means clustering methods. The comparative results of the average Silhouette Coefficient based on SDHASH similarity scores for the four ransomware categories are shown in Table II. Here, for the first WannaCry ransomware category, both clustering methods are generating almost similar results; whereas, for the second

TABLE I
SILHOUETTE COEFFICIENT FOR FCM AND K-MEANS CLUTERING BASED ON SSDEEP SIMILARITY SCORES FOR RANSOMWARE CORPORA

Ransomware	Clustering	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
WannaCry	K-Means Clustering	0.51	0.43	0.34	0.49	0.4	0.36
	FCM Clustering	0.51	0.42	0.51	0.56	0.56	0.52
Locky	K-Means Clustering	0.47	0.55	0.71	0.47	0.51	0.61
	FCM Clustering	0.47	0.6	0.71	0.75	0.83	0.83
Cerber	K-Means Clustering	0.36	0.29	0.46	0.71	0.45	0.41
	FCM Clustering	0.36	0.51	0.46	0.57	0.69	0.65
CryptoWall	K-Means Clustering	0.37	0.44	0.31	0.77	0.86	0.8
	FCM Clustering	0.37	0.55	0.64	0.77	0.86	0.86

Locky ransomware category, FCM is generating improved clustering results. However, for the last two Cerber and CryptoWall ransomware categories, the results are inconsistent and inconclusive for both FCM and k-means. Furthermore, the important issue is all the values of Silhouette Coefficient are quite low for both clustering methods, which reflects the poor quality of clustering, irrespective of the underlying clustering method. This may perhaps indicate the poor data quality (i.e. insignificant similarity scores), generated by SDHASH fuzzy hashing method. Overall, FCM clustering has still generated better or similar clustering results in comparison to k-means clustering with one major exception of Cerber ransomware category. This reflects the natural alignment of the two fuzzy methods may perhaps also dependent on the quality of data (i.e. similarity scores).

IV. EXPERIMENTAL EVALUATION OF SSDEEP AND SDHASH BASED FUZZY C-MEANS CLUSTERING RESULTS

The previous section demonstrated the slightly better performance of the FCM clustering method in comparison with the k-means clustering method. However, in the case of SDHASH, the lower values of the average Silhouette Coefficient in two ransomware categories Cerber and CryptoWall require further investigation into their clustering results to gain more insight into the clustering process and factors affecting it. Since clustering is an unsupervised method and mostly evaluated by internal metrics due the unavailability of standard class labels. However, instead of using only one internal evaluation metric which may not be sufficiently conclusive; here four metrics are utilised for a rigorous evaluation of data and cluster quality for both SSDEEP and SDHASH fuzzy hashing methods which can lead to improvement of the proposed fuzzy approach. The four evaluation metrics (particularly for FCM) Fuzzy Silhouette Index (FHI), Partition Coefficient (PC), Modified Partition Coefficient (MPC) and Partition Entropy (PE) are implemented in *fclust* package of **R** [30]. The first three evaluation metrics FHI, PC and MPC, a higher value represents improved clustering results and the last evaluation metric PE, a lower value represents improved clustering results. For all four selected ransomware categories, the values of all the four metrics are computed for the range of clusters from 2 to 7 checking the quality and consistency of the FCM clustering results based on all four metrics. This evaluation was deemed necessary for two reasons: 1) verifying FCM clustering results based on the comparative better value of all the four evaluation metrics and selecting the optimal clustering arrangement; and 2) verifying the optimal clustering result generated by FCM (using SSDEEP/SDHASH) for all the four ransomware categories with the optimal clustering result based on the ground truth (which is derived manually from the direct observation and thorough analysis of each ransomware corpus).

A. Evaluation Metrics for SSDEEP based FCM Clustering Results

Tables III to VI illustrate the evaluation metric results of FCM clustering based on SSDEEP similarity scores for the range of clusters from 2 to 7. In each result, the optimal clustering arrangement was selected based on the overall consideration of the four evaluation metrics. The optimal clustering result of FCM for all the ransomware categories WannaCry, Locky, Cerber and CryptoWall was verified with the optimal clustering result based on the ground truth to check whether it reflects the actual clustering arrangement or not (see Table XI). The optimal clustering result of FCM for the first WannaCry ransomware (i.e. cluster size = 2) and last CryptoWall ransomware (i.e. cluster size = 6) are verified as the actual clustering arrangements whereas the other two results for the Locky and Cerber are different from the actual clustering arrangements (i.e. cluster size = 3), perhaps reflecting the issue of data (similarity scores) in these clustering arrangements.

B. Evaluation Metrics for SDHASH based FCM Clustering Results

Tables VII to X illustrate the evaluation metrics results of FCM clustering based on SDHASH similarity scores for the same range of clusters from 2 to 7. Similarly, in each result, the optimal clustering arrangement was selected based on the overall consideration of the four evaluation metrics. In addition, the optimal clustering result of FCM for all the ransomware categories WannaCry, Locky, Cerber and CryptoWall was again verified with the optimal clustering result based on the ground truth to check whether it reflects the actual clustering arrangement or not (see Table XI). The results are quite similar to the previous SSDEEP results, where the optimal clustering result of FCM for the first WannaCry ransomware (i.e. cluster size = 2) and last CryptoWall ransomware (i.e. cluster size =

TABLE II
SILHOUETTE COEFFICIENT FOR FCM AND K-MEANS CLUTERING BASED ON SDHASH SIMILARITY SCORES FOR RANSOMWARE CORPORA

Ransomware	Clustering	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
WannaCry	K-Means Clustering	0.84	0.85	0.84	0.34	0.5	0.52
	FCM Clustering	0.85	0.85	0.84	0.85	0.5	0.52
Locky	K-Means Clustering	0.38	0.39	0.34	0.5	0.5	0.53
	FCM Clustering	0.38	0.49	0.55	0.56	0.56	0.56
Cerber	K-Means Clustering	0.25	0.25	0.2	0.21	0.15	0.18
	FCM Clustering	0.18	0.18	0.18	0.18	0.25	0.26
CryptoWall	K-Means Clustering	0.16	0.4	0.2	0.28	0.41	0.41
	FCM Clustering	0.31	0.25	0.3	0.41	0.41	0.41

TABLE III
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SSDEEP FOR WANNACRY RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.7224137	0.7370338	0.7833303	0.7372618	0.7760049	0.7921996
Partition Coefficient	0.7501411	0.637614	0.6347659	0.6179419	0.6327981	0.6468907
Modified Partition Coefficient	0.5002822	0.456421	0.5130212	0.5224274	0.5593577	0.5880392
Partition Entropy	0.4026222	0.6397044	0.7175874	0.7998569	0.8224821	0.822943

TABLE IV
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SSDEEP FOR LOCKY RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.6340325	0.8188398	0.8780175	0.9149925	0.9401008	0.982355
Partition Coefficient	0.6614655	0.721327	0.7788498	0.8455779	0.8942035	0.8597256
Modified Partition Coefficient	0.322931	0.5819905	0.705133	0.8069724	0.8730442	0.8363466
Partition Entropy	0.5085802	0.5008109	0.4347904	0.3331637	0.2460767	0.3055177

TABLE V
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SSDEEP FOR CERBER RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.5833714	0.7339973	0.8002152	0.7285703	0.8659088	0.7798352
Partition Coefficient	0.6660919	0.6583039	0.6870257	0.5246157	0.7842891	0.8108148
Modified Partition Coefficient	0.3321837	0.4874558	0.5827009	0.4057697	0.7411469	0.7792839
Partition Entropy	0.5099735	0.6152036	0.6248182	0.9757595	0.477397	0.433583

6) are verified as the actual clustering arrangements whereas the other two results for the Locky and Cerber are different from the actual clustering arrangements (i.e. cluster size = 3), perhaps again reflecting the issue of data (similarity scores) in these clustering arrangements. Moreover, the evaluation results for the ransomware categories Cerber and CryptoWall are suboptimal and reflect relatively the lowest values of evaluation metrics which may further affect the clustering arrangements. After analysing the SDHASH results manually, it was discovered that several SDHASH similarity scores were trivial and in the range of 1 to 10%, this could have affected the evaluation metrics and quality of clustering.

V. CONCLUSION

This paper proposed an efficient fuzzy analysis approach to cluster ransomware based on the combination of two fuzzy techniques: a fuzzy hashing method and fuzzy c-means (FCM) clustering method. The analysis, in which a fuzzy hashing method generates similarity scores for the unpacked ransomware corpus, then utilized by FCM to cluster ransomware samples into similar groups without requiring an additional transformational steps, obtaining distance amongst objects to produce the clustering. Consequently, it reduces the computational overheads by utilising fuzzy similarity scores obtained at the time of initial triaging of whether the sample is known or unknown ransomware. The performance of the proposed fuzzy method is compared against the k-means clustering method based on the average Silhouette Coefficient for the range of clusters from 2 to 7. This evaluation demonstrated a slightly better performance of the FCM-based proposed fuzzy analysis method in comparison with the k-means clustering method with only a few exceptions. Later, the two fuzzy hashing methods SSDEEP and SDHASH are evaluated based on their FCM clustering results to understand how the similarity score affects the clustering results. Interestingly, both fuzzy hashing based clustering results are quite similar in terms of providing the optimal clustering results. However, this evaluation found some of the lowest values of performance metrics for the SDHASH method, which indicated the poor data quality (i.e. insignificant similarity scores). This was verified through the

TABLE VI
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SSDEEP FOR CRYPTOWALL RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.6150487	0.7675016	0.846787	0.9110777	0.953994	0.9869859
Partition Coefficient	0.5806928	0.6725729	0.735166	0.8633428	0.9181526	0.8835326
Modified Partition Coefficient	0.1613856	0.5088593	0.6468881	0.8291785	0.9017831	0.8641214
Partition Entropy	0.606458	0.5736678	0.5180916	0.2962557	0.1989172	0.2528264

TABLE VII
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SDHASH FOR WANNACRY RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.9738676	0.9178563	0.9108147	0.907304	0.6234609	0.6746477
Partition Coefficient	0.9535457	0.9623356	0.9719547	0.9822674	0.8657797	0.8642533
Modified Partition Coefficient	0.9370913	0.9435033	0.9626063	0.9718343	0.8389356	0.8416288
Partition Entropy	0.08731693	0.07931977	0.07089287	0.08797412	0.2449191	0.2708055

TABLE VIII
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SDHASH FOR LOCKY RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.6566241	0.7355261	0.8044129	0.9636842	0.964037	0.9643856
Partition Coefficient	0.5 0.6255529	0.6413573	0.5875071	0.5420436	0.5080622	0.4814996
Modified Partition Coefficient	—	0.4383293	0.5218098	0.4843838	0.4504524	0.4260726
Partition Entropy	0.6931472	0.6633887	0.7031208	0.8353462	0.9756238	1.101317

TABLE IX
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SDHASH FOR CERBER RANSOMWARE CORPUS

Performance Matrix	Cluster Size = 2	Cluster Size = 3	Cluster Size = 4	Cluster Size = 5	Cluster Size = 6	Cluster Size = 7
Fuzzy Silhouette Index	0.4015527	0.6113267	0.6031268	0.255693	0.6966099	0.7795199
Partition Coefficient	0.5	0.3333333	0.25	0.2	0.2220473	0.2795594
Modified Partition Coefficient	—	—	—	—	0.06645678	0.1594859
Partition Entropy	0.6931472	1.098612	1.386294	1.609438	1.65455	1.605168

analysis of the SDHASH results manually. It was discovered that several SDHASH similarity scores are trivial and in the range of 1 to 10%, this could have affected the evaluation metrics and quality of clustering.

For future improvements, it is important to evaluate the proposed method with a threshold value of the similarity scores to establish the effect of these similarity scores and further enhance the fuzzy hashing algorithm itself, thus increasing the clustering performance of FCM. This proposed fuzzy analysis approach could be automated by generating sparse fuzzy rules based on the best results of FCM [31] and employing an adaptive fuzzy rule interpolation technique [32], [33], [34], [35]. Moreover, this sparse fuzzy rule base can be updated dynamically by employing dynamic fuzzy rule interpolation (D-FRI) method [36], [37], [38], [39], [40], [41], [42]. After further enhancing and automating this fuzzy analysis approach, it can be tested in different security frameworks [43].

ACKNOWLEDGEMENT

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

REFERENCES

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

[2] K. Savage, P. Coogan, and H. Lau, "The evolution of ransomware - Symantec," pp. 1–57, 2015.

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

[4] Malwarebytes. (2019) Ransomware. [Online]. Available: https://www.malwarebytes.com/ransomware/

[5] 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 *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[6] N. Naik, P. Jenkins, N. Savage, and V. Katos, "Big data security analysis approach using computational intelligence techniques in R for desktop users," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016.

[7] 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.

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

[9] 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.

TABLE X
FCM CLUSTERING PERFORMANCE EVALUATION BASED ON SDHASH FOR CRYPTOWALL RANSOMWARE CORPUS

Performance Matrix	**Cluster Size = 2**	**Cluster Size = 3**	**Cluster Size = 4**	**Cluster Size = 5**	**Cluster Size = 6**	**Cluster Size = 7**
Fuzzy Silhouette Index	0.573102	0.4108554	0.9998093	0.4144373	0.9998094	0.9998094
Partition Coefficient	0.5	0.3333333	0.3603401	0.2	0.2746634	0.2505365
Modified Partition Coefficient	—	—	—	—	0.1295961	0.125626
Partition Entropy	0.6931472	1.098612	1.170856	1.609438	1.539534	1.68124

TABLE XI
EVALUATION OF THE OPTIMAL CLUSTERING RESULTS OF FCM

Ransomware Corpus	Optimal Cluster Size based on the Ground Truth	Optimal Cluster Size based on the FCM + SSDEEP	Optimal Cluster Size based on the FCM + SDHASH
WannaCry	2	2	2
Locky	3	6	4
Cerber	3	6	7
CryptoWall	6	6	6

[10] 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 (cset 15)*, vol. 5, no. 1. USENIX Association Washington, DC, 2015, p. 52.

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

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

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[14] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE transactions on fuzzy systems*, vol. 1, no. 2, pp. 98–110, 1993.

[15] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE transactions on fuzzy systems*, vol. 13, no. 4, pp. 517–530, 2005.

[16] M.-S. Yang and K.-L. Wu, "Unsupervised possibilistic clustering," *Pattern Recognition*, vol. 39, no. 1, pp. 5–21, 2006.

[17] N. Naik, P. Jenkins, R. Cooke, D. Ball, A. Foster, and Y. Jin, "Augmented windows fuzzy firewall for preventing denial of service attack," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[18] 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.

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

[20] ——, "Securing digital identities in the cloud by selecting an apposite federated identity management from saml, oauth and openid connect," in *11th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2017, pp. 163–174.

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

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

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

[24] 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 *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[25] 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.

[26] 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.

[27] N. Naik and P. Jenkins, "Discovering hackers by stealth: Predicting fingerprinting attacks on honeypot systems," in *IEEE International Symposium on Systems Engineering (ISSE)*, 2018.

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

[29] Z. Cebeci, F. Yildiz, A. T. Kavlak, C. Cebeci, and H. Onder, "Package PPCLUST- Probabilistic and possibilistic cluster analysis," *CRAN R studio*, 2019.

[30] P. Giordani and M. B. Ferraro, "Package FCLUST: Fuzzy Clustering," *CRAN R studio*, 2015.

[31] Y. Tan, H. P. H. Shum, F. Chao, V. Vijayakumar, and L. Yang, "Curvature-based sparse rule base generation for fuzzy rule interpolation," *Journal of Intelligent & Fuzzy Systems*, Feb. 2019. [Online]. Available: https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169978

[32] L. Yang, F. Chao, and Q. Shen, "Generalized adaptive fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 839–853, Aug 2017.

[33] L. Yang and Q. Shen, "Adaptive fuzzy interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 6, pp. 1107–1126, Dec 2011.

[34] J. Li, L. Yang, Y. Qu, and G. Sexton, "An extended takagi–sugeno–kang inference system (tsk+) with fuzzy interpolation and its rule base generation," *Soft Computing*, vol. 22, no. 10, pp. 3155–3170, May 2018. [Online]. Available: https://doi.org/10.1007/s00500-017-2925-8

[35] L. Yang and Q. Shen, "Closed form fuzzy interpolation," *Fuzzy Sets and Systems*, vol. 225, pp. 1 – 22, 2013, theme: Fuzzy Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011413001486

[36] N. Naik, C. Shang, Q. Shen, and P. Jenkins, "D-FRI-CiscoFirewall: Dynamic fuzzy rule interpolation for Cisco ASA Firewall," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019.

[37] N. Naik, R. Diao, and Q. Shen, "Dynamic fuzzy rule interpolation and its application to intrusion detection," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 1878–1892, 2018.

[38] ——, "Genetic algorithm-aided dynamic fuzzy rule interpolation," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2014, pp. 2198–2205.

[39] N. Naik, R. Diao, C. Quek, and Q. Shen, "Towards dynamic fuzzy rule interpolation," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2013, pp. 1–7.

[40] N. Naik, R. Diao, C. Shang, Q. Shen, and P. Jenkins, "D-FRI-WinFirewall: Dynamic fuzzy rule interpolation for Windows Firewall," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[41] N. Naik, C. Shang, Q. Shen, and P. Jenkins, "Intelligent dynamic honeypot enabled by dynamic fuzzy rule interpolation," in *The 4th IEEE International Conference on Data Science and Systems (DSS-2018)*. IEEE, 2018, pp. 1520–1527.

[42] ——, "Vigilant dynamic honeypot assisted by dynamic fuzzy rule interpolation," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018.

[43] N. Elisa, L. Yang, F. Chao, and Y. Cao, "A framework of blockchain-based secure and privacy-preserving e-government system," *Wireless Networks*, Dec 2018. [Online]. Available: https://doi.org/10.1007/s11276-018-1883-0