Resident Data Pattern Analysis Using Sector Clustering for Storage Drive Forensics

Nitesh Bharadwaj, Upasna Singh, Gaurav Gupta

Chapter 8

RESIDENT DATA PATTERN ANALYSIS USING SECTOR CLUSTERING FOR STORAGE DRIVE FORENSICS

Nitesh Bharadwaj, Upasna Singh and Gaurav Gupta

Abstract Storage drives are huge reservoirs of digital evidence. The acquisition and examination of storage drives for evidentiary artifacts require enormous amounts of manual effort and computing resources, leading to huge case backlogs. This chapter describes a forensic triage methodology that leverages random sampling and unsupervised clustering to provide insights about the regions of interest on a storage drive. The number of sector samples to be evaluated during triage for legitimate inferences to be drawn about drive content is also discussed. Experiments involving storage drives of various capacities illustrate the effectiveness and utility of the extracted patterns for rapid drive triage.

Keywords: Large storage drives, random sector sampling, unsupervised clustering

1. Introduction

The rapid growth of storage capacity in computers and electronic devices has severely affected the timeliness of digital forensic investigations. The volume of data encountered in investigations is relentlessly advancing beyond the processing capabilities of digital forensic practitioners and traditional forensic tools [17]. As a result, huge backlogs of cases exist in forensic laboratories around the world. The immediate solution is not to modify well-defined digital forensic procedures, but to make evidence processing strategies more efficient and effective.

This research leverages random sector sampling and unsupervised clustering in the first step of a forensic examination, namely triage, to render evidence processing more efficient and effective. The idea is to perform a quick forensic survey that provides insights about resident data and data patterns on storage media. The data patterns assist in

rapidly identifying forensically-significant and insignificant regions on the media. A region is a collection of similar types of contiguous sectors on a storage drive, which can be broadly classified based on their non-null (significant) or null (insignificant) content. The significant regions include human-readable, executable, compressed and encrypted content, as well as non-null sectors, all of which are important in investigations. The insignificant regions include negligibly-important null, empty and unallocated sectors. Clearly, the identification, preservation and examination of significant regions and the elimination of insignificant regions from further processing can save enormous amounts of resources.

This chapter describes a forensic triage methodology that leverages random sector sampling and unsupervised clustering to provide insights about the regions of interest on a storage drive. The methodology rapidly explores media for resident data patterns, identifies forensically-significant and insignificant regions and makes inferences about the resident data content. The number of sector samples that need to be evaluated to make legitimate inferences about drive content is discussed. Experiments involving storage drives of various capacities illustrate the effectiveness and utility of the extracted patterns for rapid drive triage.

2. Background and Related Work

Richard and Roussev [15] have discussed the difficulties involved in processing large volumes of digital evidence. They highlighted the need for novel techniques for evidence acquisition and analysis. Garfinkel [7] notes that the massive capacity of storage devices, diversity of hardware interfaces, operating systems and file formats, large quantities of devices per case, use of anti-forensic strategies, proliferation of remote cloud storage and legal challenges are contributing to a "coming digital forensic crisis." He also discusses research directions that could help mitigate the coming crisis.

Beebe [1] emphasizes the need to address data volume and scalability issues in digital forensics using selective acquisition and effective computational and analytical approaches (e.g., data-mining-based search, file classification and graphical processing units). Quick and Choo [13] have identified the need to leverage data reduction, data mining and intelligence analysis to advance digital forensic capabilities.

Bharadwaj and Singh [3] have highlighted the key challenges and gaps (e.g., evidence examination delays, resource constraints, data heterogeneity, preservation costs, and methods and tool development) that impact digital forensics. It is imperative to develop advanced forensically-

sound techniques and tools that can support the rapid and efficient processing of large volumes of digital evidence.

2.1 Triage

Triage refers to a partial forensic examination conducted under limited time and resource constraints [17].

Garfinkel [8] has advocated the use of random sector sampling in a triage method to achieve fast drive analysis. He demonstrated its effectiveness at identifying digital media content and detecting whether or not a drive was wiped properly. Random sampling has been utilized very effectively by the New South Wales Police Force in discovery processes involving child abuse material; the application of random sampling significantly reduced case backlogs [10].

Random sampling has been used to rapidly assess storage media and identify 4 KiB blocks identical to target data [5, 21]. Taguchi [21] has developed a confidence model to handle situations where no traces of target data are identified using a sector sampling approach. Canceill [5] has provided insights on how sector sampling can assist in storage drive analysis. He demonstrates that random sampling is an adaptive and scalable method for fast drive analysis. Since the selected 4 KiB blocks were evaluated in an overlapping manner, most of the sectors (512 bytes) had to be processed multiple times. Additional sector processing introduces computational loads that result in evidence processing delays.

Bharadwaj and Singh [3] have identified the number of sector samples that needs to be analyzed on an entire drive or in regions of storage to identify sectors with content identical to the target data. In these and other triaging methodologies, information about the desired target files must be available. However, the methodology proposed in this chapter does not have this constraint. The methodology leverages random sector sampling and clustering to gain insights about the regions of interest on a drive. Prioritizing the consideration of significant regions realizes substantial savings in evidence processing resources.

2.2 Data Reduction

An alternative approach to triage is data reduction. Roussev and Quates [16] have employed similarity digests for forensic triage. They show that the scope of an investigation can be narrowed by ignoring known excludable files during the acquisition and examination phases. Quick and Choo [14] have presented an approach that enhances the traditional forensic process by imaging a selection of key files such as registry, Internet history, log, picture and video files.

Digital forensic practitioners typically have complete access to suspects' data during investigations. However, Verma et al. [22, 23] argue that privacy preservation and completeness of investigations are incompatible with each other. They proposed a method for finding the most relevant pieces of evidence while preserving data privacy in a manner that increases investigative efficiency without negatively impacting evidence integrity and admissibility.

Beebe and Clark [2] discuss the benefits of applying data mining in digital forensic investigations. However, limited published work incorporates data mining and other techniques to reduce the effort involved in preserving and examining large volumes of digital evidence [12].

In contrast, the methodology proposed in this chapter does not rely on the collection of essential files; instead, regions of interest are identified by intelligently evaluating randomly-selected sector samples from a drive. The evaluation draws on clustering techniques that determine the significant regions based on the features selected for each random sector. These significant regions are targeted for selective acquisition and examination instead of processing all the drive sectors.

2.3 Clustering

Clustering has been employed in data mining and unsupervised learning applications to identify and understand data patterns in unlabeled, high-dimensional data. Clustering groups data using similarity measures based on centroid, hierarchical, expectation maximization and density techniques. Each clustering technique has its own advantages and disadvantages in terms of cluster quality, efficiency in handling noisy data and computational complexity. The efficiency and effectiveness of clustering techniques are dependent on the features selected for evaluation. This work employs simple centroid and density based clustering techniques to determine forensically-significant regions on evidentiary drives.

Centroid-Based Clustering. The k-means clustering technique computes the centroid of a cluster as the mean of the feature vectors assigned to the cluster. The technique requires the number of clusters to be specified in advance. It divides W samples into k disjoint clusters such that a distance function computed as the sum of squares of the intra-cluster distances to the centroid of the cluster is minimized. The distance function is given by:

$$Distance\ Function\ = \sum_{j=1}^{k} \sum_{i=1}^{W} \|w_i^{(j)} - c_j\|^2 \quad (1)$$

where c_j is the centroid of cluster j and $w_i \in W$.

In this work, the distinction between the significant and insignificant regions on a drive is formulated by considering three clusters ($k = 3$) that broadly represent three distinct types of data.

Density-Based Clustering. The density-based spatial clustering of applications with noise (DBSCAN) technique considers a cluster to be an area of high density separated by low-density samples. The clusters identified using this technique can be of any shape (non-linear boundaries), yielding different results compared with k-means and other linear clustering algorithms.

Two user-defined parameters *minimum_samples* and *eps* determine the density of samples needed to form a cluster. Higher *minimum_samples* and lower *eps* values indicate higher densities while lower *minimum_samples* and higher *eps* values indicate lower densities.

A sample in a dataset is called a core sample when other neighboring samples (*minimum_samples*) exist within a radius or distance of *eps*. Thus, prior information about the number of clusters is not required. The number of clusters is estimated based on the *minimum_samples* and *eps* parameter values.

2.4 Extracted Features

Extracted features or metrics can provide valuable insights about digital evidence. In this work, insights about forensically-significant regions on storage media are obtained using two derived metrics: (i) ASCII score; and (ii) entropy value.

ASCII Score. The greater the amount of text or human-readable ASCII bytes contained in a data unit, the greater the probability of it containing directly understandable information [19]. A sector is considered to be the smallest data unit on a drive. It is recommended that small sectors or blocks (e.g., 512 bytes) be considered because file blocks should efficiently map to drive sectors [24]. Hence, the standard size of a data unit (sector) considered in this research is 512 bytes.

Traditionally, the ASCII score is the ratio of the number of ASCII bytes to the total number of bytes in a file [19]. However, in this work, the ASCII score is evaluated for every randomly-selected sector instead of a specific file. This may assist a digital forensic practitioner in examining even minute details instantaneously from the drive, such as keywords, credit card details, email, phone numbers and other information that can be directly recorded and understood by the practitioner. Moreover, the ASCII score can help exclude sectors containing little or no human-

readable information. If an investigative scenario requires the analysis of plaintext or directly-readable information, then sectors with high ASCII scores should be analyzed first (highest priority) because it becomes much easier to extract useful information that could provide important leads when dealing with a large volume of data.

Entropy Value. Entropy specifies the amount of uncertainty of an unknown or random quantity. It is computed by summing the frequency of each observed byte value in a fixed-length data block and then computing an entropy value. Lyda and Hamrock [11] compute the entropy value based on bytes (`00` to `FF`) in a file using `bintropy`, a binary-file entropy analysis tool that enables practitioners to conveniently and quickly identify encrypted and packed malware.

In this work, an entropy value is computed for bytes in every randomly-selected sector on a drive. The entropy value is low for sectors that are less compressed (e.g., text files) and high for compressed file fragments [19]. Encrypted data also has a high entropy value.

The entropy value $E($s of a randomly-selected sector s is given by:

$$E(s) = -\sum_{b=1}^{m} P(b) \log_2 P(b) \tag{2}$$

where $P(b)$ is the probability of the frequency of the b^{th} byte information in sector s that consists of a series of m bytes. Alternatively, the entropy value can be viewed as considering all the values that a byte b in a sector s can take, and $P(b)$ is the probability of the frequency of each occurring byte in the randomly-selected sector s.

Randomly-selected sectors are easily classified as null or non-null sectors based on their content [4]. In this work, the entropy metric is used to identify sectors with human-readable, multimedia (images, audio and video), encoded, compressed, encrypted or executable content. Hence, the entropy value is used in addition to the ASCII score in sector evaluations.

A sector that contains only zero or null bytes is referred to as a null-sector. A null-sector has the lowest entropy value of zero. The ASCII score for a null-sector would be high. However, in this work, a null sector is considered to have an ASCII score of zero due to the absence of relevant information. A sector that contains information other than null bytes is referred to as a non-null sector. Non-null sectors have plaintext (direct human readable), multimedia, encoded, compressed, encrypted or executable content.

Table 1. Random sector categories based on the ASCII scores and entropy values.

Range of ASCII Score (x)	Range of Entropy Value (y)	Assumed Sector Category
0	0	Null data
$0.6 \leq x \leq 1.0$	$0 < y \leq 4.8$	Plaintext data
$0 < x < 0.6$	$4.8 < y \leq 8.0$	Compressed/encrypted data

Lyda and Hamrock [11] statistically evaluated a large set of packed and encrypted malware files based on the entropy of their bytes. They classified them into four categories of files: (i) plaintext; (ii) native; (iii) packed; and (iv) encrypted executable.

2.5 Assumptions

The proposed methodology assumes that three categories of data exist: (i) null data; (ii) plaintext data; and (iii) compressed/encrypted data. In general, it is easy to discriminate between null and non-null sectors. However, it is difficult to differentiate between resident and deleted data in the absence of the original filesystem or prior information.

The proposed methodology employs two metrics, ASCII score and entropy value, to determine forensically-significant regions on storage media. Storage media is considered to correspond to a bulk data volume, possibly without a legitimate filesystem and metadata, as in the case of deleted, altered or corrupted filesystem information or a formatted drive. The methodology is also applicable to forensic images with raw formats such as DD, IMG and RAW.

Table 1 shows how ASCII scores and entropy values are used to categorize randomly-selected sectors as containing null, plaintext and compressed/encrypted (encoded) data [11, 19]. As mentioned above, null data has an ASCII score of zero and an entropy value of zero. Since readable (plaintext) file fragments always have high ASCII scores and low entropy values, the ranges for this category are set to [0.6, 1.0] and (0, 4.8], respectively. Finally, the ASCII score and entropy value ranges for compressed/encrypted data are set to (0.0, 0.6) and (4.8, 8.0], respectively.

Many clustering algorithms, including k-means, require the number of clusters to be known *a priori*. Therefore, the proposed methodology assumes that the maximum number of clusters is three. DBSCAN clustering is highly dependent on the *minimum_samples* and *eps* parameter

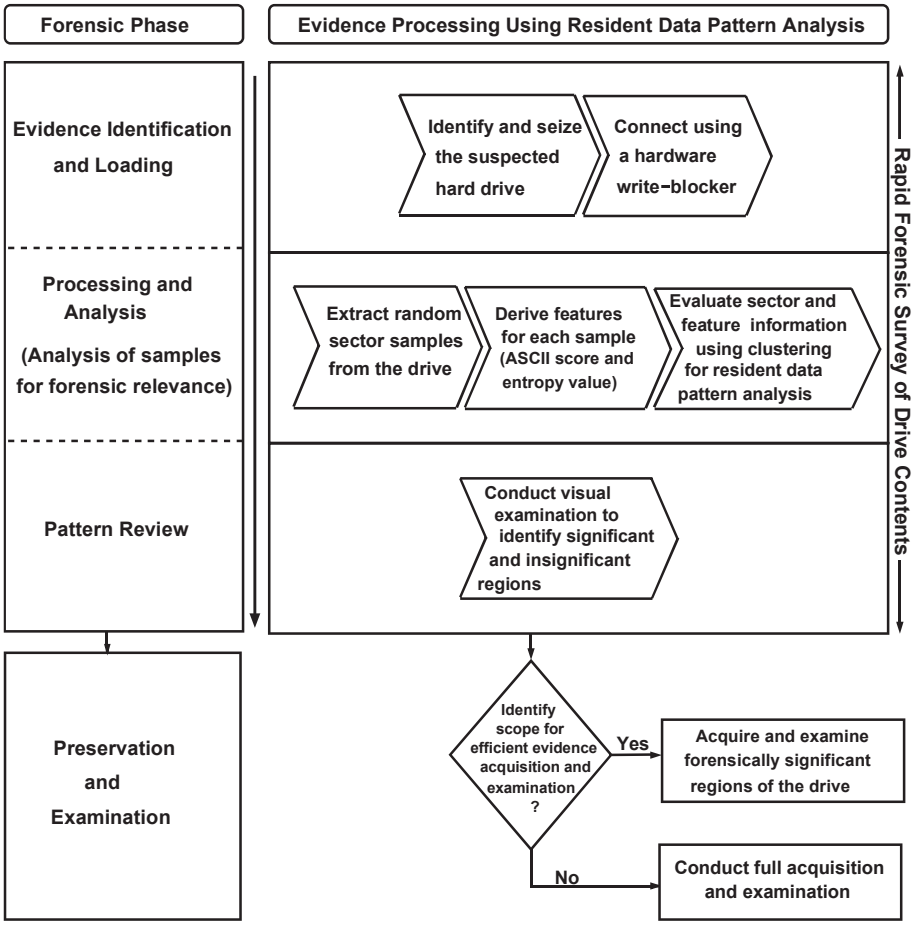


Figure 1. Proposed methodology for efficient evidence analysis.

values. The proposed methodology typically uses $minimum\_samples = 10$ and $eps = 0.5$. However, explicit mentions are made when different values of these parameters are employed.

3. Proposed Methodology

Figure 1 presents the proposed methodology for efficient evidence analysis. The storage media drive is assumed to be mounted on the investigator's computer.

The process begins with the random extraction of a specified number of sector samples from across the drive. Important features such as the ASCII score, entropy value and sector category are computed and

recorded for each extracted sample. The sector versus feature map is examined using k-means and DBSCAN clustering to gain insights about the significant and insignificant regions on the drive. Although fewer sector samples are examined, it is still possible to obtain a good idea of the distribution and characteristics of data on the drive. When a reasonable quantity of insignificant sectors exist on the drive, their elimination from consideration reduces the subsequent analysis effort. When the drive data patterns reveal that the number of significant sectors is large or the drive is completely filled with data, it is advisable to proceed with a full forensic acquisition followed by the exhaustive examination of artifacts.

The random sector samples are selected based on the accessible sector count. This information can be obtained using utilities such as `fdisk` and `hdparm`. In random sampling, an arbitrary number between the first and last sector number is generated, which is then recorded for further evaluation. During analysis, care is taken care to ensure that the random samples are fetched without replacement. Specifically, no sector should be selected multiple times; this is accomplished by maintaining a record of the previously-selected sectors. When a previously-selected sector is identified, it is dropped in favor of a new sector despite an increase in the evaluation load. Therefore, it is important to determine the sample size that provides good outcomes in a timely manner while eliminating the need to conduct an exhaustive examination of the drive.

Sampling theory is engaged to determine the number of sector samples for random extraction. Specifically, random sector sampling without replacement is employed.

Sample Size Determination. Bharadwaj and Singh [3] have specified the numbers of random samples that need to be evaluated to identify sectors that are identical to target file fragments on storage media with different probabilities, regardless of the presence or absence of filesystem metadata. However, when the target data of interest is not known, it is difficult to determine the number of random samples that need to be examined.

The determination of the number of sector samples that can provide adequate insights about the resident sectors on a drive resembles the problem of determining the adequate sample size for a finite population [9]. Four parameters are needed to determine the sample size: (i) population size (total number of accessible sectors on a drive); (ii) precision (user-specified); (iii) confidence level (user-specified); and (iv) degree of variability (user-specified).

Precision (sampling error) is the range in which the true value of a population is estimated to reside [9]. Precision has an inverse relationship with the number of samples – the lower the specified precision, the greater the number of samples required [20]. In general, precision is expressed as a percentage (e.g., ±3%, ±5%, ±10%). For example, if 60% of the sector samples were determined to be unallocated with a precision of ±3%, then between 57% and 63% of the sectors on the drive are actually unallocated.

The confidence level, which originates from the central limit theorem, provides the probability that the sample contains the value being estimated. It is expressed as a percentage (e.g., 90%, 95%, 99%). The confidence level generally corresponds to the standard (constant) z-score value [20]. Different z-scores based on different confidence levels must be employed when deriving the sample size.

Finally, the degree of variability expresses the distribution of attributes in a population. A more heterogeneous population requires a larger number of samples whereas a more homogeneous population requires fewer samples. A safe decision is to use 0.5 (50%) as the degree of variability because it balances a large sample size against maximal population variability.

According to Cochran [6], the following equation can be used to obtain a representative sample size n_0 as a proportion of a population:

$$n_0 = \frac{Z^2 pq}{e^2} \tag{3}$$

where Z^2 is the abscissa of the normal curve that cuts off the area of the desired confidence level, e is the desired precision or sampling error, p is the estimated proportion of attributes present in the population and $q = (1 - p)$.

When the population is finite, the desired sample size n is given by:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}} \tag{4}$$

where N is the population size.

Equations 3 and (4) are used to determine the number of sector samples that need to be processed in order to estimate the characteristics of the sectors residing on storage media.

4. Experiments and Analysis

Experiments were conducted to evaluate the efficacy of the proposed significant region identification methodology for drive triage. A generic

Table 2. Minimum sample sizes for various drive capacities.

Confidence	Precision	Drive Capacity				Sample Size
		4 GB	8 GB	16 GB	1 TB	
99%	±1%	16,558	16,574	16,582	16,590	**17,000**
	±2%	4,146	4,147	4,148	4,148	**4,500**
	±3%	1,844	1,844	1,844	1,844	**2,000**
	±5%	664	664	664	664	**700**
	±10%	166	166	166	166	**200**
95%	±1%	9,594	9,599	9,602	9,604	**10,000**
	±2%	2,401	2,401	2,401	2,401	**2,500**
	±3%	1,068	1,068	1,068	1,068	**1,100**
	±5%	385	385	385	385	**400**
	±10%	97	97	97	97	**100**
90%	±1%	6,761	6,764	6,765	6,766	**7,000**
	±2%	1,692	1,692	1,692	1,692	**1,800**
	±3%	752	752	752	752	**800**
	±5%	271	271	271	271	**300**
	±10%	68	68	68	68	**80**

eight-core computing system with 4 GB RAM running Kali Linux 2.0 was employed in the experiments. The implementation is available at GitHub (`github.com/niteshdiat2014/Resident_Data_Pattern_Analysis`).

The experiments were conducted on four storage drives, D_1, D_2, D_3 and D_4, with capacities, 4 GB, 8 GB, 16 GB and 1 TB, respectively. Drive D_1 was completely filled with data whereas D_2, D_3 and D_4 were partially filled with data. The analysis was performed using a custom Python 2.7 script. Clustering was implemented as described in the scikit-learn documentation [18].

4.1 Sector Sample Size

Equations (3) and (4) were used to estimate the numbers of samples necessary for drive analyses. Table 2 shows the numbers of sector samples for various drive capacities at precision (sampling error) values of ±1%, ±2%, ±3%, ±5% and ±10%, where the estimated proportion of attributes present in the population $p = 0.5$. The sample sizes in the last column of the table are the upper bounds on the sample sizes used to analyze evidence.

The computed sample sizes are valid for the considered scenario; however, the sample sizes would vary when Equations (3) and (4) are computed with different parameter values depending on the scenario require-

ments. The sample size does not change much for populations larger than 20,000, which implies that the total number of samples should be considered at least during data pattern analysis. Similarly, the computed sample sizes are not very different for different storage media with different numbers of sectors for a particular precision and confidence level.

The computed sample size should guarantee well-distributed sectors from a drive. However, forensic practitioners may use arbitrarily large numbers of samples according to their investigative needs. Obviously, the larger the sample size, the better the ability to make precise decisions about a drive, but this comes with increased analysis effort.

4.2 Significant Region Analysis

In order to identify the important regions on the drives, features were recorded for every randomly-selected sector in the retrieved sample set. The features, ASCII score and entropy value, were clustered separately using k-means and DBSCAN. This enabled the sectors with similar feature values to be segregated from sector groups with completely distinct feature values. Finally, the sector samples were mapped based on the computed cluster labels to make inferences about the important regions on the drives.

It was observed that the resident data patterns obtained using the two clustering approaches were very similar. In general, k-means provided better results when drives had fewer null sectors. However, k-means sometimes misclassified sectors because it produces clusters with linear structures; this was mitigated by DBSCAN clustering that handles clusters of arbitrary (non-linear) shapes.

In order to measure the efficacy of the proposed methodology, analysis was performed using the computed number of sector samples (e.g., 17,000 with 99% confidence and $\pm 1\%$ precision).

The k-means clustering technique was first used to segregate the sample set into three clusters. Figure 2 shows the clusters obtained by k-means on the 16 GB drive based on the ASCII scores and entropy values. The three clusters correspond to the different types of data on the drive: (i) null sectors with ASCII scores and entropy values of zero; (ii) sectors with moderate ASCII scores and high entropy values; and (iii) sectors with high ASCII scores and medium entropy values.

The cluster labels were utilized to map the sector numbers with their corresponding feature values from the sample set. Figures 3(a) and 3(b) show the resident data pattern analysis using k-means clustering on the 16 GB drive. Figure 3(a) shows the feature maps based on sector samples and ASCII scores. Figure 3(b) shows the feature map based on

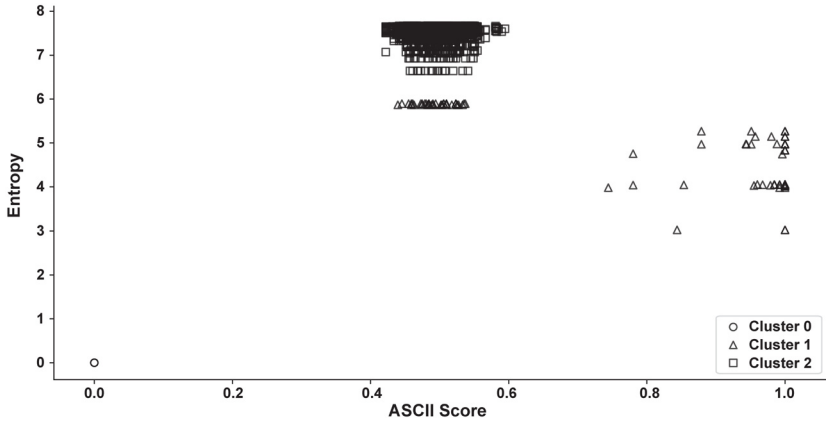
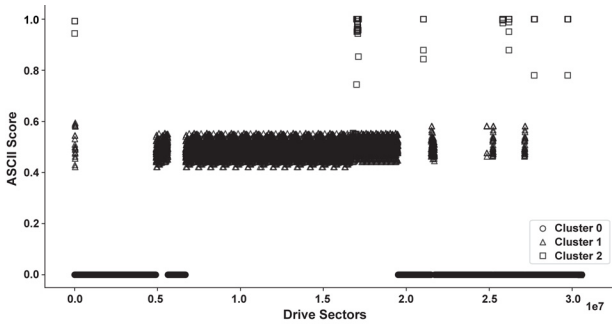
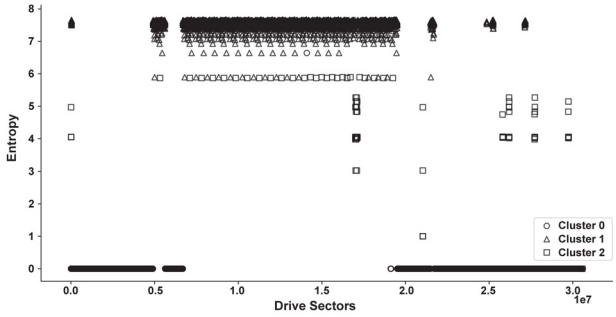


Figure 2. Clusters obtained using k-means clustering (16 GB drive).



(a) Feature map based on sector numbers and ASCII scores.



(b) Feature map based on sector samples and entropy values.

Figure 3. Resident data pattern analysis using k-means clustering (16 GB drive).

sector samples and entropy values. The two figures clearly illustrate the resident data pattern, revealing the regions on the drive that contain data for a forensic practitioner to prioritize for further analysis. The figures also reveal that the drive contains a reasonable amount of null sectors at the beginning and end whereas a large proportion of implicit information resides between sectors 0.4×10^7 and 2×10^7.

The implicit sectors may contain images, videos and other encoded information that are directly understood by a forensic practitioner. Targeting these high entropy regions to analyze multimedia and other files that are usually encoded is more effective than examining the entire drive. Figures 3(a) and 3(b) also indicate that considerable amounts of human-readable plaintext information (with high ASCII scores and low entropy values) exist in the sector range 1.5×10^7 to 2×10^7; these regions can be directly interpreted by a forensic practitioner. Selective file carving and recovery approaches can be used to improve the overall efficiency. On the other hand, regions with large amounts of null data, such as those in the sector range 1.5×10^7 to 2×10^7 in Figures 3(a) and 3(b), should be excluded from further analysis to enhance performance.

As discussed above, the centroid-based k-means clustering technique causes some misclassifications. Figure 3(a) shows that fewer sectors are labeled incorrectly (e.g., sectors with high ASCII scores labeled as null data and sectors with low ASCII scores labeled as human-readable). However, despite the misclassifications, the extracted data pattern can still provide a digital forensic practitioner with valuable insights that would enhance the efficiency and effectiveness of the analysis.

Figure 4 shows the clusters obtained by DBSCAN on the 16 GB drive based on the ASCII scores and entropy values. The clustering is based on the densities (closeness) of features regardless of the mean values of the clusters. Note that the sectors with similar features are in the same clusters. Although the number of clusters is not required for DBSCAN clustering, the technique still yielded three clusters based on the feature values and the related parameters ($minimum\_samples = 10$ and $eps = 0.5$).

Figures 5(a) and 5(b) show the resident data pattern analysis using DBSCAN clustering on the 16 GB drive.

Figures 6(a) through 6(d) show the resident data pattern analyses using k-means and DBSCAN clustering on the partially-filled 8 GB drive. A total of 8,000 random samples were selected. Note that different DBSCAN parameter values $minimum\_samples = 8$ and $eps = 0.18$ were employed for the 8 GB drive.

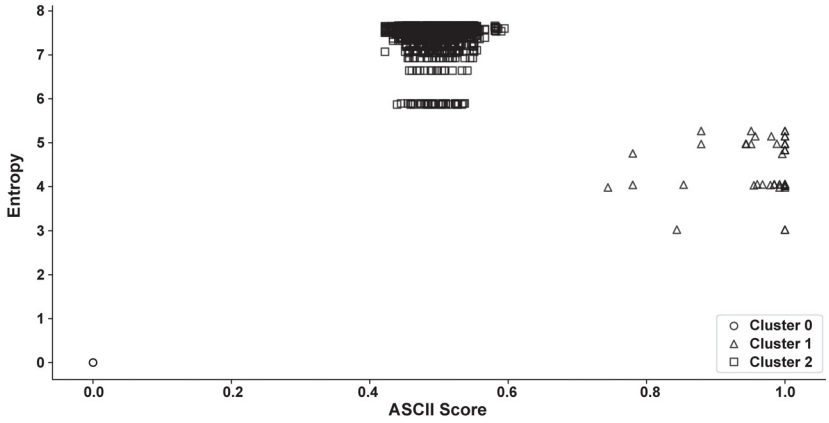
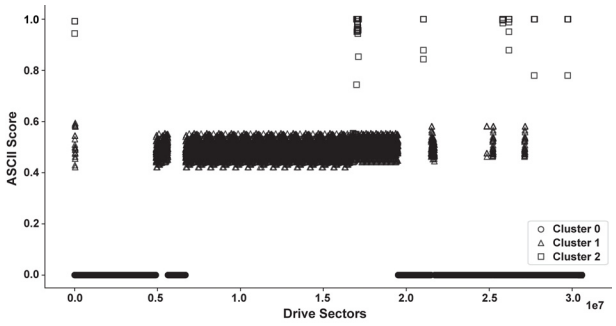
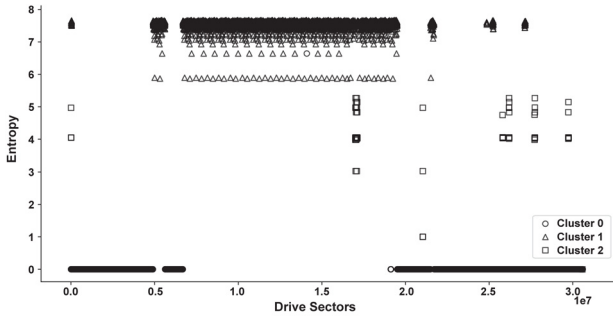


Figure 4. Clusters obtained using DBSCAN clustering (16 GB drive).

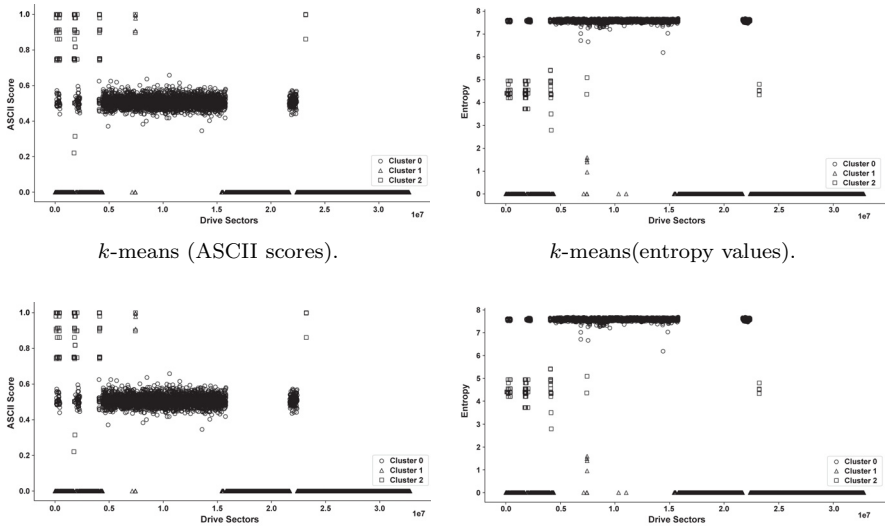


(a) Feature map based on sector numbers and ASCII scores.



(b) Feature map based on sector samples and entropy values.

Figure 5. Resident data pattern analysis using DBSCAN clustering (16 GB drive).



k-means (ASCII scores). k-means(entropy values).

DBSCAN (ASCII scores). DBSCAN (entropy values).

Figure 6. Resident data pattern analyses using k-means and DBSCAN (8 GB drive).

4.3 Performance Metrics

It is important for forensic practitioners to assess the performance of the proposed methodology to satisfy the scientific testing criterion [7]. The performance measures employed are the true positive rate (TPR) and false positive rate (FPR), along with the receiver operating characteristic (ROC) curve.

The TPR and FPR values associated with each clustering technique were computed by comparing the actual labels against the observed outcomes for arbitrary numbers of samples (e.g., 1,000, 5,000, 10,000, 50,000 and 100,000). Arbitrary sample sizes were chosen to evaluate the efficacy of the proposed methodology in situations where it is needed to evaluate a range of sector samples (few samples to a considerably large number of samples). The sample sizes cover the minimum number of samples (up to 17,000) required to provide a general pattern of the contents of an entire drive.

The computed TPR and FPR values obtained with k-means and DBSCAN clustering on drives D_1, D_2, D_3 and D_4 are plotted as ROC curves in Figure 7. Since drive D_1 (4 GB) was completely filled, the assumed number of clusters and the *minimum_samples* and *eps* values do not provide satisfactory outcomes (low TPR and high FPR values). This is due to the very small number of unallocated or null sectors on

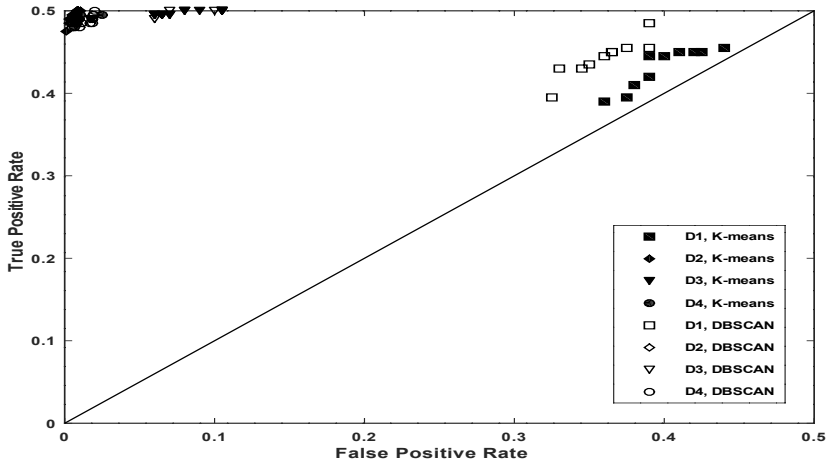


Figure 7. ROC plots for k-means and DBSCAN based pattern analyses.

the drive. However, acceptable results – high TPR and low FPR values – are obtained for drives D_2 (8 GB), D_3 (16 GB) and D_4 (1 TB) because they were partially filled with data.

4.4 Evaluation Delay

Increasing the sample size increases the evaluation delay. The evaluation delay is also affected by the input/output performance of the storage media and computing system, efficiency of feature value derivation and computational effort associated with the clustering techniques. Increasing the number of features also increases the computational effort.

Figure 8 shows the evaluation delays for various proportions of random sector samples. Increasing the number of random samples increases the evaluation delay. In contrast, the input/output rate is platform centric, implying that different outcomes are expected for different scenarios and computing environments. A small number of samples can be examined at a high input/output rate whereas a large number of samples significantly reduces the input/output rate.

4.5 Error Rate

Although random sector sampling is effective for rapid drive analysis, it is difficult to ignore its associated error rate (i.e., evaluation of repeated sectors) in the absence of a perfect random number generator. In order to assess the error rate related to significant region determination and resident pattern analysis, different proportions (0.1% to 50%) of random

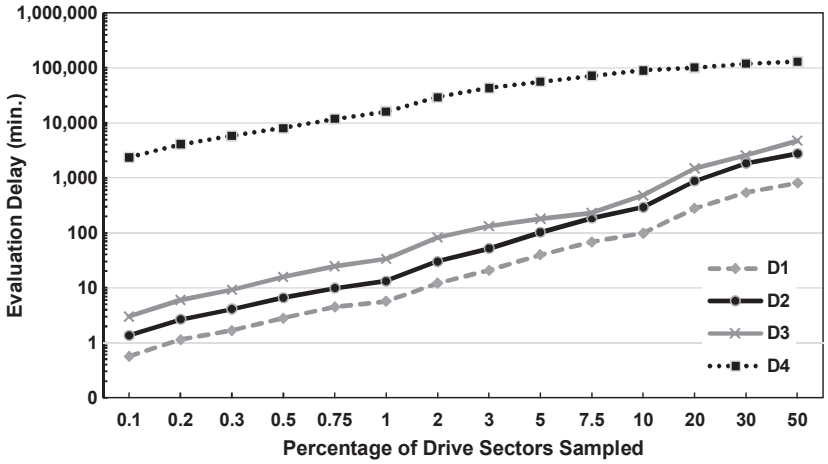


Figure 8. Evaluation delays for various proportions of random sector samples.

samples from the four drives were analyzed to measure the extent of repeated sector evaluation.

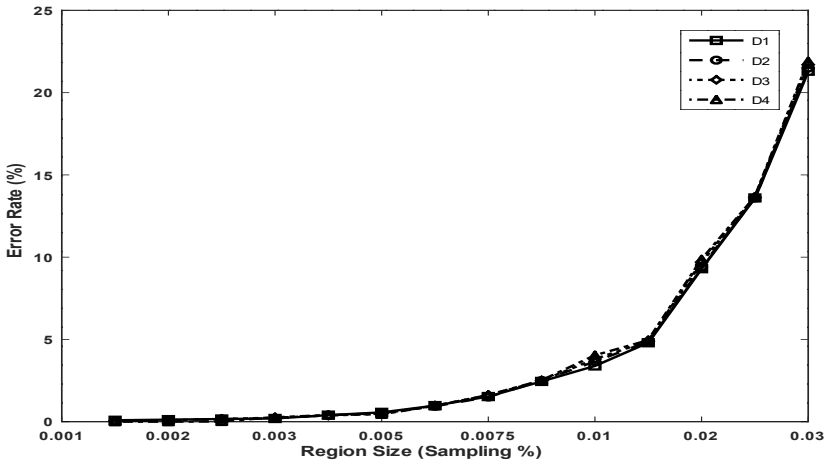


Figure 9. Error rates for various proportions of random sector samples.

Figure 9 presents the error rates for various proportions of random sector samples. The error rate increases at an average rate of approximately 20% as the proportion of random sectors increases. The error rate can be managed by keeping track of previously-generated sector samples.

If a previously-generated sector sample is selected, it is dropped from consideration and a new random sector sample is selected in its place.

5. Conclusions

The proposed triage methodology assists digital forensic practitioners in rapidly evaluating resident data patterns on storage media to narrow the scope of evidence acquisition and examination to forensically-relevant data. It leverages random sector sampling and unsupervised clustering to provide insights about the regions of interest on storage media. The proposed methodology is applicable when metadata information or resident data content are not readily available, for example, when filesystem metadata is corrupted, altered, deleted or overwritten, or when drives are formatted, deleted or overwritten. Without the methodology, the only alternative in these situations would be to exhaustively examine every sector for evidentiary artifacts. The methodology is not intended to replace full evidence examination. Instead, it is most effective for conducting visual examinations of drive content layout, intelligence analyses, resident data pattern analyses, rapid reviews, quick forensic surveys, pre-seizure media analyses, drive triage, and partial or selective evidence processing.

Experiments involving storage drives of various capacities illustrate the effectiveness and usability of the extracted patterns for rapid drive triage. However, the performance degrades when large numbers of random sector samples have to be evaluated when processing large-capacity storage media. The methodology is designed to handle three types of clusters corresponding to null, plaintext and compressed/encrypted data; however, the results are negatively impacted when insufficient data is associated with the clusters. The methodology is unable to handle completely encrypted drives where sectors have low ASCII scores and high entropy values; in such cases, it is necessary to decrypt the storage media before applying the methodology. Additionally, the methodology cannot handle advanced and compressed file formats such as the Advanced Forensic Format (AFF) and Encase image file format (E01).

Future research will focus on extending the types of data that can be handled. Also, it will focus on enhancing the efficiency of the methodology and reducing error rates.

References

[1] N. Beebe, Digital forensic research: The good, the bad and the unaddressed, in *Advances in Digital Forensics V*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 17–36, 2009.

[2] N. Beebe and J. Clark, Dealing with terabyte data sets in digital investigations, in *Advances in Digital Forensics*, M. Pollitt and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 3–16, 2006.

[3] N. Bharadwaj and U. Singh, Efficiently searching for target data traces in storage devices with region-based random sector sampling, *Digital Investigation*, vol. 24, pp. 128–141, 2018.

[4] N. Bharadwaj and U. Singh, Significant data region identification and analysis using k-means in large storage drive forensics, *Security and Privacy*, vol. 1(4), paper no. e40, 2018.

[5] N. Canceill, Random Sampling Applied to Rapid Disk Analysis, Master's Research Project Report, Department of System and Network Engineering, University of Amsterdam, Amsterdam, The Netherlands, 2013.

[6] W. Cochran, *Sampling Techniques*, John Wiley and Sons, New York, 1977.

[7] S. Garfinkel, Digital forensics research: The next 10 years, *Digital Investigation*, vol. 7(S), pp. S64–S73, 2010.

[8] S. Garfinkel, Fast disk analysis with random sampling, presented at the *Annual CENIC Conference*, 2010.

[9] G. Israel, Determining Sample Size, Fact Sheet PEOD-6, Florida Cooperative Extension Service, University of Florida, Gainesville, Florida, 1992.

[10] B. Jones, S. Pleno and M. Wilkinson, The use of random sampling in investigations involving child abuse material, *Digital Investigation*, vol. 9(S), pp. S99–S107, 2012.

[11] R. Lyda and J. Hamrock, Using entropy analysis to find encrypted and packed malware, *IEEE Security and Privacy*, vol. 5(2), pp. 40–45, 2007.

[12] D. Quick and K. Choo, Data reduction and data mining framework for digital forensic evidence: Storage, intelligence, review and archival, *Trends and Issues in Crime and Criminal Justice*, no. 480, 2014.

[13] D. Quick and K. Choo, Impacts of the increasing volume of digital forensic data: A survey and future research challenges, *Digital Investigation*, vol. 11(4), pp. 273–294, 2014.

[14] D. Quick and K. Choo, Big forensic data reduction: Digital forensic images and electronic evidence, *Cluster Computing*, vol. 19(2), pp. 723–740, 2016.

[15] G. Richard and V. Roussev, Next-generation digital forensics, *Communications of the ACM*, vol. 49(2), pp. 76–80, 2006.

[16] V. Roussev and C. Quates, Content triage with similarity digests: The M57 case study, *Digital Investigation*, vol. 9(S), pp. S60–S68, 2012.

[17] V. Roussev, C. Quates and R. Martell, Real-time digital forensics and triage, *Digital Investigation*, vol. 10(2), pp. 158–167, 2013.

[18] scikit-learn, Machine learning in Python (`scikit-learn.org`), 2019.

[19] M. Shannon, Forensic relative strength scoring: ASCII and entropy scoring, *International Journal of Digital Evidence*, vol. 2(4), 2004.

[20] A. Singh and M Masuku, Sampling techniques and determination of sample size in applied statistics research: An overview, *International Journal of Economics, Commerce and Management*, vol. II(11), 2014.

[21] J. Taguchi, Optimal Sector Sampling for Drive Triage, M.S. Thesis, Department of Computer Science, Naval Postgraduate School, Monterey, California, 2013.

[22] R. Verma, J. Govindaraj and G. Gupta, Data privacy perceptions about digital forensic investigations in India, in *Advances in Digital Forensics XII*, G. Peterson and S. Shenoi (Eds.), Springer, Cham, Switzerland, pp. 25–45, 2016.

[23] R. Verma, J. Govindaraj and G. Gupta, DF 2.0: Designing an automated, privacy preserving and efficient digital forensic framework, *Proceedings of the Annual ADFSL Conference on Digital Forensics, Security and Law*, pp. 127–150, 2018.

[24] J. Young, K. Foster, S. Garfinkel and K. Fairbanks, Distinct sector hashes for target file detection, *IEEE Computer*, vol. 45(12), pp. 28–35, 2012.