


Holistic Analytics of Digital Artifacts: Unique Metadata Association Model

Ashok Kumar Mohan, TIFAC-CORE in Cyber Security, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham University, India

 <https://orcid.org/0000-0003-3519-2228>

Sethumadhavan Madathil, TIFAC-CORE in Cyber Security, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham University, India

Lakshmy K. V., TIFAC-CORE in Cyber Security, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham University, India

ABSTRACT

Investigation of every crime scene with digital evidence is predominantly required in identifying almost all atomic files behind the scenes that have been intentionally scrubbed out. Apart from the data generated across digital devices and the use of diverse technology that slows down the traditional digital forensic investigation strategies. Dynamically scrutinizing the concealed or sparse metadata matches from the less frequent archives of evidence spread across heterogeneous sources and finding their association with other artifacts across the collection is still a horrendous task for the investigators. The effort of this article via unique pockets (UP), unique groups (UG), and unique association (UA) model is to address the exclusive challenges mixed up in identifying incoherent associations that are buried well within the meager metadata field-value pairs. Both the existing similarity models and proposed unique mapping models are verified by the unique metadata association model.

KEYWORDS

Association Grouping, Digital Forensic Automation Dataset, Metadata Analytics, Metadata Association Model, Metadata Dataset, Metadata Forensics, Similarity Mapping, Unique Association(s)

INTRODUCTION

Metadata, in general, is “data about data” and in principle, it is a unique set of attributes (data) that describes the inconsistent possessions about the object (data) it’s tailgating at all times. A digital forensic investigation visualizes the same definition as “evidence about evidence”, resembling a set of clues (evidence) about an object of digital archaeological interest (evidence) as quoted in the digital forensic research works of Raghavan, S. (2013). Having the capability to pass through a filter over metadata that puts together the missing dots to locate a precise suspect document and prove its origin via reconstructing the timeline in a forensically sound manner. Most metadata are piggybacked to the context file displaying information such as file name, file size, file extension, modified, accessed, and created (MAC) timings. Metadata for a digital forensic investigator is a unique way to know something or everything that is fused around the actual data. It can be visualized as a cover layer closely surrounding a piece of evidence completely or partially at all times. So that the forensic analyst will have a better idea of what that evidence is all about and the potential clue it reveals to

DOI: 10.4018/IJDCF.20210901.0a5

This article, published as an Open Access article on July 2nd, 2021 in the gold Open Access journal, the International Journal of Digital Crime and Forensics (converted to gold Open Access January 1st, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

support the hypothesis of the investigator. Everything from the unique name, information on how data combines together, when and by whom the data was created, by whom the data was reproduced and lists of web pages visited by people, and even network packets and system logs can be classified as metadata. Balasubramanian, V., Doraisamy, S. G., et al., (2016) explains the ever-evolving lecture videos and proposes a multimodal metadata extraction system based on Naive Bayes and rule-based classification on keyphrases and topic-based segments of the video files.

The primary purpose of metadata is meant for sorting out the huge library, indexing them for easy access, fixing bugs, and versioning for tracking objects. The supplementary task of any standard library model in particular to metadata is helping the investigator to find the actual information they are looking for. It would make better sense for evidential data to be associated using a compelling relationship with each other via unique metadata matches. This classification of metadata not only makes their job easier but also promising to give a good reason for their algorithm proven right away. The traditional file system based metadata as portrayed by Daniel, L., & Daniel, L. (2012) covers the broad categories of more common types of metadata. It holds the time-stamp for their associated time zones accumulated by the operating system and chronologically rendered when an artifact/file is produced, accessed, or modified. The current day NTFS file system as explained by Casey, E. (2009) depicts the metadata created by the file system resides well within its traditional indexing data structure called Master File Table (\$MFT). When compared with the traditional FAT based file systems, this NTFS metadata comprises several complementary metadata information like the origin, the current active status (disk or trash), and the access control permissions of the file. The present-day advancement in big data technologies via Hadoop and Cassandra by now has an inbuilt feature called a backup node by Krishnan, K. (2013) which contains the exact copy of the majority of the file system metadata. About one-third of the population of the files collected from the annual snapshots of windows computers by Agrawal, N., Bolosky, W. J., et al., (2007) were from the most commonly used top ten windows file formats namely exe, gif, jpg, mp3, wma, dll, htm, cpp, lib, and h. Rajendiran, K., Kannan, K., et al., (2020) emphasized the application of machine learning in cyber forensics to automate and enhance the investigation strategies.

Metadata with respect to the analytics made by the popular U.S tech stocks GAFAM's metadata exploration services targeting the millennial population to convert raw data into contextualized records gave an insight on the third dimension of exploring semantics between objects via metadata. A classic exemplar of the modern advancement in metadata is Uber's Databook: Turning Big Data into Knowledge with Metadata at Uber (n.d) lineage via big data-based architecture backed by vertica and hive for collecting metadata, while back-end storage is facilitated with elastic search and finally visualize a wide variety of frequently refreshed metadata via RESTful APIs in a dashboard. The grid exposed in Table 1 represents the very abstract classification of artifacts with their relationship and priorities calculated via the occurrence and frequency of certain unique metadata filed-value pairs. The hybrid access control matrix for artifact mobility categories A11 and B21 is the prime focus for an investigator as they may or may not transfer artifacts at regular intervals, but in most cases they will carry rich metadata corresponding to their notable actions.

The main points addressed by Raghavan, S., & Raghavan, S. V. (2014) shows ground-breaking approaches that have been investigated to get hands-on and analyze digital evidence from divergent resources. This article concise the key process involved in digital forensics, research challenges, time zone issues and time interpretation based on skew and drift, standards and guidelines for NIJ and DoJ, AFF- an open-source forensic image format, digital evidence bags, forensic acquisition tools, and frameworks, file carving for data hiding and steganography, metadata, timestamps and time-lining, correlation and corroboration using visual similarity. The whole work summarizes in making a model that automatically associates evidence by recognizing the correspondence over multiple digital pieces of evidence for a holistic view of every single action across all digital evidence sources.

SIFT based forensic analysis for the copy-move forgery detection is well established by Amerini, I., Ballan, L., et al., (2011) in their attack detection article and the same applies to the generic image-

Table 1. Hybrid Access Control Matrix for Artifact Mobility

Artifact Access Control Matrix	Artifacts regularly created/sent by the sender A (A11)	Artifacts never created/sent by the sender A (A12)
Artifacts regularly modified/resent by the receiver B (B21)	jpg, mp3, avi, doc, rar	icns, cfg, drv, ini, sys
Artifacts never modified/reset by the receiver B (B22)	pst, ost, vcf, exe, tmp	e01, bin, dd, raw ad1

based artifacts. Raghavan, S., & Raghavan, S. V. (2013a, November) formulated the design of the conventional tools for analyzing each source of digital evidence as a BLOB (Binary Large Object) and it is up to the investigator to find similar items from evidence. Since there is a quick augmentation in technology, the crucial task for digital evidence is to recognize a precise tool to carry out the analysis. In this paper, a systematic study of existing forensic tools utilizing an algorithm based assessment to stumble on the singular functionalities supported by these tools has been carried out. So the restrictions of the forensic tools based on evidence verification and a sample case for building evidence correlation functionalities into these tools were also discussed to substantiate the same. The future work of this assessment was to broaden their AssocGEN architecture in detecting and grouping similar artifacts for digital investigation.

The authors illustrated the fact that any cyber forensic tools need manual interaction to analyze and report an incident. They proposed an automatic method to reduce the manual interception for evidence analysis. To achieve this, they have combined the working model of an analysis tool and machine learning. Machine learning will help in formulating the analysis of evidence and to identify the hidden features for investigating the evidence and to make decisions accordingly. The author also suggests considering the time and space to make their calculations more accurate. Prem, T., Selwin, V. P., et al., (2017, April) emphasizes the metadata related to disk memory via a forensic framework that analyzes the memory-related metadata identifiers for forensic investigation. All the above-mentioned metadata preface portrays the significance and the role of metadata in supporting digital forensic investigations. The author's contribution of the proposed unique models will address the research gap in metadata forensics for identifying and mapping the sparse relationship between heterogeneous metadata field-value pairs. The authors unique association model can be applied analogous with Bhattacharya, S., Kaluri, R., et al., (2020) recent machine learning models that employs technique like PCA-firefly for adding metadata related features to improve the competence of native IDS. A similar work was carried out by RM, S. P., Maddikunta, et al., (2020) on IoMT attacks possesses an extended applicability of the authors metadata model on network packets for IDS.

Metadata Standards

Exchangeable Image File Format (Exif) is a de facto standard that sets the formats for images, audio, video, documents, and other supplementary identifiers used by imaging devices. A distinctive XML-based metadata format that depicts the contents of the artifact is termed as Extensible Metadata Platform (XMP) standard has the advantage of permitting the user to append custom parameters to suit your needs. XMP metadata has engagements with various other metadata standards like TIFF, JPEG, and PDF while handling the artifacts. Accurate and consistent data about artifacts can be facilitated by the IPTC Photo Metadata Standard via the most common sub-standards namely IPTC Core and IPTC Extension which are so keen on addressing the digital rights and access control related metadata properties of an image. Error Level Analysis in short identified as ELA is the supportive algorithm employed with these standards for identifying anomalies at different compression levels to verify the integrity of the artifact. The scope of this article is achieved by the following three phases

that serve as the route for accomplishing the anticipated unique association mapping via the present unique grouping model as shown in Figure 4.

Phase One: Artifact Collection

Phase Two: Metadata Classification

Phase Three: Unique Mapping

METADATA FOR DIGITAL FORENSICS

Raghavan, S., & Raghavan, S. V. (2016, January) proposed 'Rachna' algorithm to recreate numerous instantaneous browser activities based on coherency and concurrency of metadata association across the evidence from browser history logs, number of GET resource requests, and network packet archive. The algorithm was demonstrated by analyzing five-tabbed Mozilla firefox browser sessions to derive the relationship to sort out the leading set of synchronized artifacts and make out the number of instantaneous active sessions by tracking the metadata from the dynamic browser elements for the user activities. Mendelman, K. (n.d.) demonstrated an authentic case in point for the apt use of extracting and mapping metadata that is publicly available on the websites related to Estonian government organizations. This materialization is accomplished by developing a fingerprinting model to give us an idea about the fact that only a small number of documents metadata were properly shredded and the metadata of other documents remain safe and sound within the source documents.

The significant works of Raghavan, S., & Raghavan, S. V. (2014) on sorting out the derivation of the evidential facts using metadata that represents the state of a file and it helps in identifying sections of the evidence that has relevant information. This phenomenon is applied in solving real-world cases in categorizing doctored photographs and imitative documents with the help of inter-file relationships. It is accomplished using a Markov model-based approach for identifying metadata associations sorting out the semantic relationships to files, folders, and their corresponding logs as in the work of Amato, F., Castiglione, A., et al., (2020). Raghavan, S., & Raghavan, S. V. (2009, September) is a quantitative time-based approach that attempts in correlating the incident in particular with multiple evidence sources. The evidence composition model classifies the pairs of correlated events that occur within a stipulated time frame for a suspected crime with general search complexity of $O(N \log N)$ across the evidence pool. To be precise, the model takes into consideration the three most commonly used families of network protocols namely, DNS, UDP/TCP/HTTP, and IRC within a specific time frame. Later it is proposed to extend the same replica to larger time slots with all protocols that are available in a given network packet (evidence).

A Desire for Metadata Extraction

Metadata appears to be some variety of a spin-off from an evidence file, nevertheless, it can be exploited to examine authoritative actions of a subject (user) over an object (file) similar to the author's rationalization headed for the need for accessing, sorting out and analyzing all available atomic metadata and their inherent relationship as shown in Table 2. The example artifacts are taken into account from the user-generated UMAM-DF dataset as shown in Figure 2. The metadata match between the evidence shown below in indexes X1-X5 is common heterogeneous metadata matches via the generic field-value pairs of the metadata performed in the existing models. The need for mapping the matches between the artifacts from indexes X6-X8 where a random file of any universally accepted file type (extension) has a sparse metadata match between them or their known predecessors from X1-X5 is on need of this hour. As shown in the table, X6 carries a random file with a '.vbe' extension that runs encrypted scripts upon execution, X7 has an unusual file with 'xlsm' macro enabled excel sheet that is not a common artifact but can run malicious scripts if enabled and finally, X8 depicts corrupt files like '.raw' extension that might carry broken application metadata

Table 2. Demonstration of Assorted and Sparse Metadata (Filed-Value) Combinations

Index	Artifact (Evidence)	Source	Field: subject	Field: tags	Field: category	Field: copyright	Field: title	Field: <sparse field>
X1	pinkie.jpg	Ex1:C2M	pirated	stolen	<null>	<null>	<null>	<null>
X2	birds.jpg	Ex1:C2M	<null>	pirated	<null>	<null>	<null>	<null>
X3	DOC-S1As1.docx	Ex1:P2D	<null>	stolen	pirated	<null>	<null>	<null>
X4	pinkie.jpg	Ex1:L2P	<null>	<null>	<null>	stolen	<null>	<null>
X5	pinkie.jpg	Ex1:D2C	stolen	<null>	<null>	<null>	pirated	<null>
X6	Filename.vbe <random file type>	Ex2:*	<null>	stolen	<null>	<null>	<null>	amazon
X7	Filename.xlsm < unusual file type>	Ex3:*	<null>	<null>	<null>	pirated	<null>	fighter
X8	Filename.raw <corrupt file type>	Ex4:*	<null>	<null>	<null>	<null>	<null>	rao

in place. In existing similarity metadata matches, these sparse occurring file types are ignored totally and are addressed in the proposed unique association models. In the course of this article, the authors explain the unique mapping methodology to achieve the same. As a proof of concept the metadata field values namely amazon, fighter, pirated, rao, and stolen are embedded into the artifact metadata fields for demonstration.

The authors make use of Exiftool(a platform-agnostic CLI application) created and managed by Phil Harvey (2005) for interpretation, marking, and even restricting metadata over a variety of file types. It is powerful, speedy, customizable, and also provisionally processes files based on the value of any metadata taking numerous output formatting options. It also notes down every change in the file to creation, modification, and access date. Also, it's straightforward to create a text output file for each image file and the same can be extended to be stored in json, csv, and xls file formats.

With reference to the standard digital evidence analysis models by Agrawal, N., Bolosky, et al., (2007), the authors have categorized every digital artifacts (Origin O) into six major variety of families namely image (Family 1), file archiver (Family 2), executable (Family 3), document (Family 4), multimedia (Family 5) and forensic image (Family 6) as in Figure 1. The authors demonstrate the raw headers of one of the sample artifacts from the recently generated Amrita-TIFAC-Cyber/ Digital-Forensics/UMAM-DF (Unique Metadata Association Model - Digital Forensics) datasets (2020). It shows the shift of metadata identifiers from the source (z) and the same artifact copied to

Figure 1. Families and Groups of Digital Artifacts (Author's Perception)

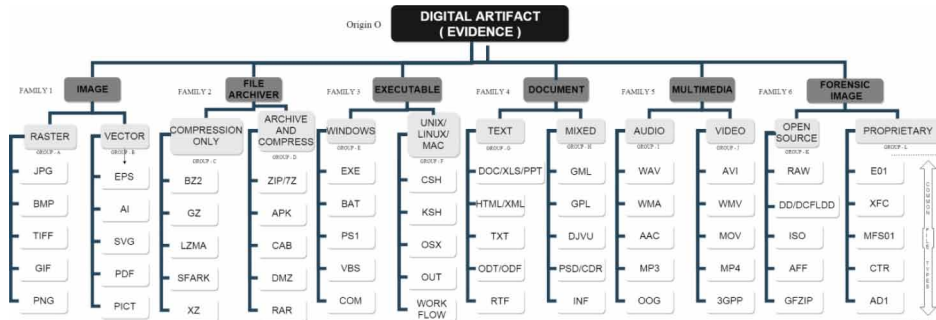


Table 3. Homogeneous and Heterogeneous Artifact Mapping

Artifact Mapping	Same Family-Same Type	Same Family-Different Type	Different Family - Same Type	Different Family -Different Type
File (pair) Nature	Purely Homogeneous	Habitually Homogeneous	Habitually Heterogeneous	Purely Heterogeneous
Example 1	G1: JPG - GIF	G1: JPG - EPS	G1: TIFF - PS1	G1: JPG - MP3
Example 2	G2: PNG - JPG	G2: TIFF - SVG	G2: BZ2 - 3GPP	G2: EXE - ISO
Example 3	G3: JPG - PNG	G3: PNG - PICT	G3: CAB - GFZIP	G3: TXT - E01

social media platform Facebook (z') illustrated around 90% of the actual metadata is modified or removed by the social media platform that possesses a nightmare for digital forensic investigators while proving their hypothesis before the jurisdiction.

(z) pinkie.jpg (S1As1-Mobile)

FF D8 FF E0 00 10 4A 46 49 46 00 01 01 0048 00 48 00 00 FF E1 13 EA 45 78 69 66 00 00 4D 4D 00 2A 00 00 00 08 00 0E 01 28 00 03

(z') pinkie.jpg (S1As6-Facebook)

FF D8 FF E0 00 10 4A 46 49 46 00 01 01 00 0001 00 01 00 00 FF ED 00 84 50 68 6F 74 6F 73 68 6F 70 20 33 2E 30 00 38 42 49 33 30 30

Metadata Association Models

The lemma based theorems on metadata similarity by Raghavan, S., & Raghavan, S. V. (2017) to identify the cause and effect of the relationship between metadata values to derive a grouping artifact on reducing the volume of metadata to be examined is a remarkable work. They gave details about the similarity between metadata in two hierarchies as similarity pockets and similarity groups. Afterward from these two association group is derived to find out the reduction factor and grouping efficiency by performing a lemma based analytics on metadata. Their future works were comprehensible on applying the theoretical proofs to existing datasets and to evaluate the difference between the forthcoming practical results of lemma implementation of their models. They also put forward to broaden the operational metadata association model to heterogeneous data sources and automating the same to be valid for digital evidence stored and processed during big data. This metadata association model is pretty good while handling any evidence with a distinct number of digital artifacts where a set of distinct extensions from a selected source is considered. The authors categorize artifacts into evidence types in various families and distinct file types with the example grouping shown in the following Table 3 with respect to Figure 1.

Determining Sparse Associations Between Metadata

With respect to the demonstration of assorted and sparse metadata (filed-value) combinations from Table 2, being motivated to generate and share the unique metadata-based dataset to the digital forensic research community. After comprehensive literature, on existing digital forensic datasets the authors have taken the following ten unique JPG images from dataset mobile source S1 and these acts as the reference (genesis) artifacts for the proposed unique mapping algorithm. The same set is synthetically recreated across all other sources as shown in Figure 2 keeping in mind each file holds the metadata created from their corresponding source file system and application for the visually similar images as stated by Buchholz, F., & Spafford, E. (2004). The ultimate purpose of this dataset is to recreate

Figure 2. Real-World Images Obtained from S1: Mobile (S1As1) with Complete Metadata



visually similar evidence (images in this case) at all sources and monitor the change or degradation of metadata on each iteration as shown in Figures 5 and 6.

WIDESPREAD SIMILARITY ASSOCIATION(S)

Metadata associations have been discussed in handling the digital forensic investigation for a while and there exist a plethora of syntactical models that roughly match the metadata composition and are not as much of predominant in addressing the explicit semantic behavior of the metadata attributes and their corresponding parameters. Raghavan, S., Clark, A., et al., (2009, January) hypothetically explicate the handling of multiple sources of evidence in a single framework (FIA) classified based upon source, data semantics, and storage file formats with the help of Malcolm Corney case on car theft investigation at Queensland University of Technology. They also emphasize extending this framework to design a suitable contrivance for validating their prototype amid real-world digital forensic datasets.

Raghavan, S., & Raghavan, S. V. (2013b, November) plotted metadata associations to establish a relationship between the artifacts and group the associated artifacts. AssocGEN analysis engine determines the relationship stuck between artifacts from files, logs, and network packet source to group the interrelated artifacts with respect to the circumstance of a digital investigation. Raghavan, S., & Saran, H. (2013, November) put forward the Provenance Information Model (PIM) to deal with the challenges related to timestamp analysis transversely for manifold time zones to precisely take into custody, the time zone in sequence and authenticate time-related affirmation during metadata analysis named after UniTIME timelining tool. Raghavan, S. (2014) thesis on Metadata Association Model

(MAM) demonstrates the well-designed MAM algorithm for identifying the image file associations and intentionally modified image files via metadata association. The source file, ownership, timestamps, and application-related metadata are extracted and analyzed the evidence. It has a variety of carved images, images downloaded directly from the internet, digital photographs, and images generated using regular designing software.

The contribution in this article from the researchers Raghavan, S., & Raghavan, S. V. (2017) in correlating metadata associations by formulating the similarity pocket (Sp), similarity group (Sg) and (Ug) was established using a sequence of lemma(s) and corresponding proof of concept to defend the functionality of the mathematical model. This metadata analytic association model serves as a foundation for the authors to formulate the series of supportive and unique metadata association models (Ua) rendered in the later part of this article. The minimal understanding from the above article's existing system that is mandatory to proceed with formulating the author's proposed system is quoted below as the prerequisite for this work. Native notations from the source may vary with respect to synchronizing it with the flow of the author's contribution substituted by the fresh notations to avoid objectionable confusion throughout the course of the proposed article. The algorithms mentioned below are precisely established by Raghavan, S., & Raghavan, S. V. (2017) with a stronghold of lemma and their corresponding proofs for volume reduction and grouping efficiency. This acts as the foundation for the author's algorithm transformation from existing association grouping to proposed unique associations.

Similarity Pocket (Sp)

Similarity pocket is the collection of all artifacts (An) with field-value pair match for the same metadata field ($MF-IDn$). Every artifact may have two or more similar pockets within the source. Each similarity pocket follows a ruling of a minimum of two artifacts having the same metadata match and a maximum of all metadata value matches for a specific field. So the largest set of similarity pockets that can be formed by the algorithm will be half the size of the total artifacts count, but it is only the theoretical bound and most often not accomplished in the real world.

Similarity Group (Sg)

After the creation of all available similarity pocket (Sp) shown above, Raghavan, S., & Raghavan, S. V. (2017) clusters the largest union of Sp by transitive closure property to cluster them into a scaled-down version named as similarity groups (Sg). Same as the predecessor following transitive closure property, the similarity groups created here should be mutually exclusive to each other.

Association Group (Ag)

The same principle applied to similarity grouping (Sg) is applied to the groups across all the sources (Sn) is termed as association group and they further tend to exponentially trim down the size of data that have to be considered for any digital investigation. The purpose of these normalizing algorithms is much needed in the current day IT infrastructure relying much on big data based data storage and classification system where the volume of data generated at an unimaginable speed.

Influential Unique Association(s)

The authors address the sparse set of left-out artifacts namely a single artifact without any metadata match will not be a part of the existing metadata association model following the similarity grouping principles of the alliance. But also the work of Raghavan, S., & Raghavan, S. V. (2017) may possess a greater impact if subsided by this work of finding unique associations and unnoticed metadata matches.

Need for Unique Associations

The existing system addresses the common and the most expected similarity that is unseen, so the authors focus on sorting out and grouping the required capability needed to detect a rare match from the ignored unique pockets. The similar pockets may provide evidence for the security analyst to track the connection between artifacts and narrow the scope for troubleshooting. But, the unique similarity between two pockets and the exact expected clue (like copyright='stolen') on metadata from evidence in one source and the corresponding sparse match (like tag='stolen') as shown in Table 2 from evidence in another source does not fall well with the scope of similarity pocket category. As it is mainly proven by a series of mathematical equations with the theoretical proof for the existing similarity, the inferences by the forerunners will be void in this special case when subjected to unpredictable real-world conditions that govern the nature of ever-changing metadata. To address the sparse similarity, models similar to Liu, Z., Lai, Z., et al., (2020) on the extraction of unique features by semi-supervised learning via discriminated projection are to be integrated into the existing similarity grouping model for effective grouping efficiency and achieving the expected trends in volume reduction.

During the course of discovering metadata matches, it is essential to understand the need for including the sparse features and rare combinations of metadata filed-value matches across all heterogeneous sources. The authors establish the three parallel models as shown in Figure 3 working hand-to-hand with the existing similarity grouping for enabling the procedure of shaping the sparse or rare metadata relations and resulting combination. It aims at not only accomplishing the volume reduction factor as expected but also sorting out all rare combinations of the ignored metadata features via unique grouping.

The proposed unique metadata mapping algorithm to support the existing similarity models are as follows.

Recognition of Overlooked Unique Pockets (UP) Algorithm

Unique Pockets account for all those distinct and atomic metadata field-value pairs that failed the condition of similarity pocket (an enforced set to have a minimum of two artifacts in common) that have the unique metadata value match for any metadata. This is also restricted to single metadata for a particular value (*MV-IDn*) and single source as in similarity pockets.

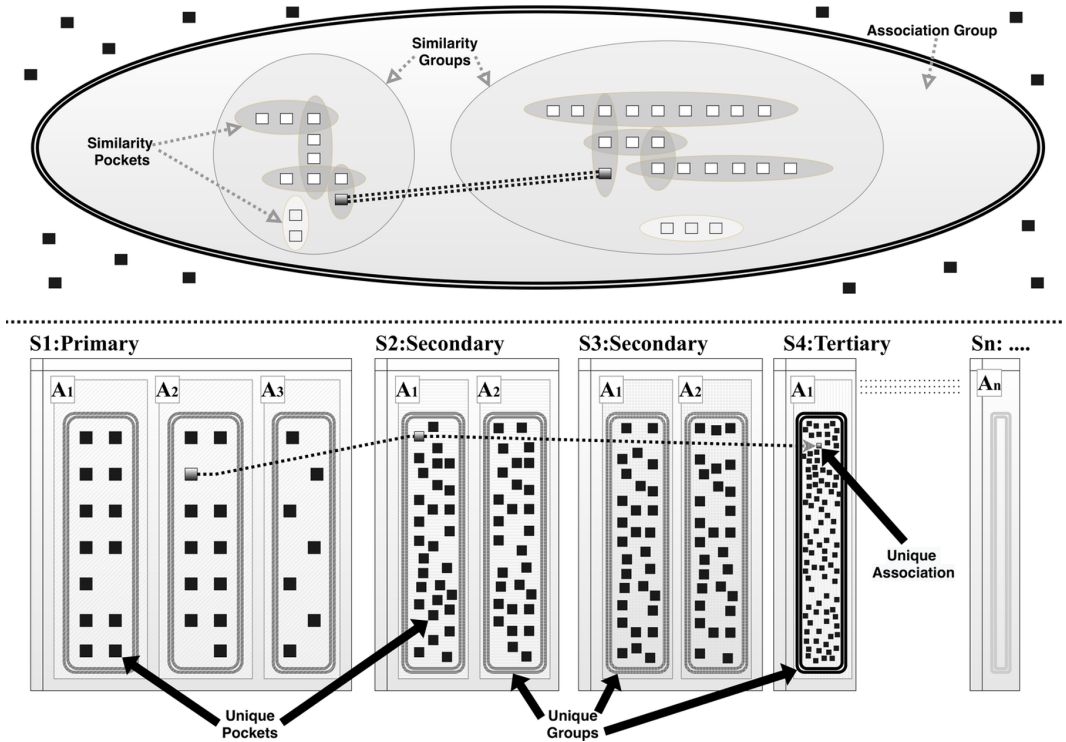
Revamping the Pockets as Unique Groups (UG) Algorithm

Unique Groups are assembled by the largest set of Unique Packets that belongs to the same artifact ID (*An*). The construction of the Unique Group is executed on a single artifact across all of its metadata and values solitary within the source. In no doubt, it also excludes the filed-value pair that are already labeled as Similarity Pockets and grouped as Similarity Groups.

Discovery of Unique Associations (UA) Algorithm

Unique Associations are produced by the largest set of Unique Groups that exhibits a minimum field-value match for one artifact between them irrespective of the metadata field (*MF-IDn*) Identifiers. The construction of the Unique Association is widespread similar to Association Group considering every metadata and transversely to the entire sources poll. Based on the three algorithms formulated above, the digital sources are conceived to be in four hierarchies of extension namely H1 the source itself, H2 when Unique Packets are formed, H3 when Unique Groups are formed, and finally H4 when Unique Associations are formed. The mapping of sparse field-value pair from H1 to H2, H2 to H3, and H3 to H4 heads for the unique metadata mapping of the effort as UM_{op} , UM_{pg} and UM_{gu} . Each attempt is exclusively executed in parallel straight away after the calculation of the readily available similarity grouping models.

Figure 3. Identifying Research Gap (Unique Association Ignored by Association Group)



After the successful collection of artifacts and classification of metadata, at this juncture of initiating phase three as shown in Figure 4 where the authors start mapping from H1. Set 'D' symbolizes the complete data pool with all sources of artifacts that are to be holistically scrutinized among the heterogeneous pool of evidence. The metadata field-value pair namely $S_n:A_n-MF-ID_n:MV-ID_n$ uniquely tagged to the artifact is the key vector in mapping effort as UM_{ap} , UM_{pg} , and UM_{gu} . H1 starts with a mapping effort of a fixed filtering mapping UM_{ap} using metadata matches to sort out the unattended (unique) artifacts into a category with only one $MF-ID_n:MV-ID_n$ named as unique pockets that be in contact with an artifact. The set of all unique pockets assembled based on metadata matches is characterized as UP shown in Figure 4.

As the collection of all atomic missed out metadata features in unique similarity pocket might be available in many Unique Pockets. Each artifact (A_n) tends to include numerous such pockets, the authors aim at eliminating this redundant state by the mapping effort UM_{pg} to group all the unique pockets that belong to the same artifact (A_n) to form unique groups that are mutually exclusive with similar pockets. The set of all unique groups are represented by UG in Figure 4. Once the above mapping practice is accomplished, each unique group will have a collection of unique pockets only mapped from a single artifact on a single source and are mutually exclusive. Taking into account all the metadata values ($MV-ID_n$) produced by the artifacts belonging to unique groups across all sources, the authors further plots a makeover to identify the sparse matches. It is achieved by mapping artifacts concerning the metadata equivalence giving out a semantic relationship on the metadata field-value ($MF-ID_n:MV-ID_n$) pair in each of the unique groups and combinations tried across all the sources from the evidence data pool.

Artifact collection is the initial phase to identify and collect artifacts from different sources and copying the same to the local storage in a forensically evident manner preserving the file integrity and mainly the MAC timings of the evidence. Before the triage, the basic footprinting of the evidence files is completed by indexing their basic file system metadata namely file name, type, size, last accessed time, file created time, and modified time. The authors also store two or more standard hashes like MD5 and SHA variants to maintain the integrity of the file while performing structured metadata analysis. While discussing about hashing based integrity checks with respect to metadata, the file system metadata is like an envelope to the data. So the majority of the hashing algorithm does not include them in the act of hashing, which emphasizes this file system metadata that is located outside the file and can be altered without impacting the actual data. In contrast, application metadata is a part of the file and moves with the file's actual data, so eventually, any change in the application metadata will affect the integrity of the file.

Phase two shown in Figure 4 is a triage model to categorize and prioritize the evidence. The artifacts are categorized into three categories based on the familiarity and frequency of the occurrence of specific types of files in an evidence pool. The most common file types belonging to images, audio, video, and documents are categorized as primary type. This investigator ought to collect all available metadata and verify the same with publicly available metadata standards to sort out missing values if any. Two third of the files in any typical digital evidence falls under the primary category. Supplementary file types like executables, archives, e-mail files, web documents, backup files, and temporary files are grouped under secondary type and need to collect all the publicly available metadata from these file types. Starting from a simple heterogeneous match like file type, file size, and timestamps across all the primary, secondary and tertiary evidences as shown in Figure 4, Unique Association (UA) mapping can be established to map the artifacts from unique groups taken transversely between sources using mapping UM_{gu} in the destination mapping at H4. The set of all association groups are represented by AG in Figure 4.

CONSTRUCTIVE STEPS/PHASES MAPPING FOR UNIQUE ASSOCIATION

The unique association is the cumulative representation of all the three Algorithms (UP, UG, and UA) proposed by the authors to assemble the bits and pieces of potentially unique metadata associations overlooked by the predecessor's similarity association models (Table 4).

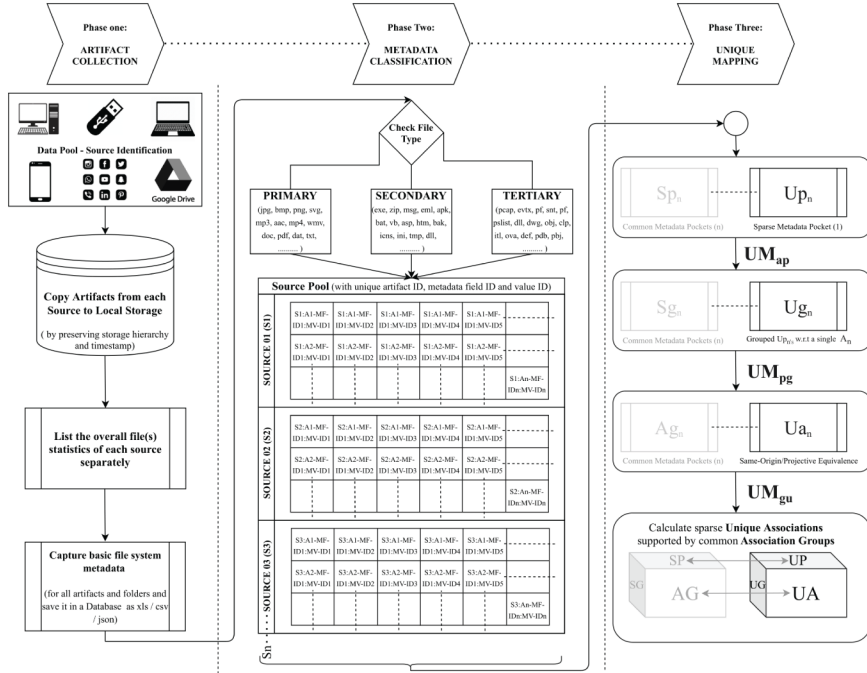
After the collection of artifacts and classification of metadata in phase1 and phase2 respectively, the authors take the root folder (or) the source pool to be denoted by 'D' formulated in (1). It has the categorized pool of all the sources in separate folders which are imaged/copied in a forensically sound manner for the proposed phase3 of Unique Mapping. It is achieved via implementing the following modus operandi hand-in-hand in a hierarchical way using the existing similarity grouping algorithms at the foundational stage, in other words, the readily available algorithm quoted above serves as a precondition for the anticipated model to accomplish its objective at the same time as expected.

$$D = \{S_n \mid S_n \text{ is the } n^{th} \text{ source from the data pool} \in [1, N] \forall \text{ values of } n\} \quad (1)$$

The individual sources S_n may contain the files and folders commonly grouped via Equation (2) as artifacts (A_n) in this work as shown below. Each artifact has its own well-defined set of unique identifiers or headers called metadata (M) based on the file type they belong to.

$$A_n = \{a_1, a_2, a_3 \dots a_n\} \in \text{their corresponding source } S_n \quad (2)$$

Figure 4. Phase-wise Implementation and Data Flow of Unique Associations



The artifacts are categorized into three distinct classifications namely primary, secondary, and tertiary as shown in Figure 4 for a convincing artifact triaging. The author's scope on this cataloging is to collect each and every metadata from a primary category like images, documents, and multimedia files in a forensically sound manner. Then the necessary metadata is collected in a secondary category based on the combination of EXIF, ICC, IPTC, and XMP metadata standards and lastly, the universally obtainable file system metadata is collected in the tertiary category. Unique metadata mapping aims at collecting all metadata even from tertiary evidence like pcap or evtx that might have a sparse association with any of the primary or secondary evidences.

The building blocks for the metadata element for any artifact is represented by a regular 2-tuples notation by the authors throughout the article as $\langle field: value \rangle$ pair as in (3,4) for the publicly available metadata standards.

$$M_f - ID_n \text{ be the identifier for the 1st tuple } \langle field : \rangle$$

$$\forall f \text{ identifi cal notation } \exists \text{ an fixed } n \in [ASCII (num | char)] \quad (3)$$

$$M_v - ID_n \text{ be the identifier for the 2nd tuple } \langle : value \rangle$$

$$\forall v \text{ identifi cal notation } \exists \text{ an viable } n \in [ASCII (num | char)] \quad (4)$$

The combine notation of any metadata value corresponding to a metadata field that belongs to a unique artifact from a selected source is represented via (5) the below distinctive notation.

Table 4. Steps in Scrapping and Analyzing Metadata Filed-Value Similarity for Unique Associations

Steps/ Phase	Notation	Authors Action / Mapping Rules	Illustration and Statistics
1/1	D	<ul style="list-style-type: none"> Select the artifact pool (root folder) Fix the number of sources 	<ul style="list-style-type: none"> /var/tmp/ C:\Users\MAM\Documents\...
2/1	Sn	<ul style="list-style-type: none"> Select the artifacts from different sources Local Disks, Network Shares, Removable Storage Devices, Mobile Devices, Forensic Images... 	<ul style="list-style-type: none"> S1: Mobile, S2: Laptop, S3: Pendrive, S4: Desktop, S5: Cloud...
3/1	Sn	<ul style="list-style-type: none"> Copy the artifacts from 'Sn' to 'D' Preserve MAC timings Maintain file structures, folder hierarchy, and integrity of all the artifacts 	<ul style="list-style-type: none"> Using safe copy scripts to preserve MAC timings (to the best of the author's ability)
4/1	Sn:An	<ul style="list-style-type: none"> List the total information of Sn Tag each artifact with a unique ID (An) 	<ul style="list-style-type: none"> S1:A1, S1:A2, S1:A3, Sn:An
5/1	Fp	<ul style="list-style-type: none"> Capture the basic file system metadata for all artifacts and folders Save it in a Database (xls, csv, json) <p>[Exclusively as a part of the preservation of distinctive metadata for artifact verification in later phases]</p>	<ul style="list-style-type: none"> File/Folder Name File Type File Size Location Modified Time Accessed Time Created Time Generate CRC, MD5, SHA5 Hashes
6/2	Sn:An-Pr Sn:An-Se Sn:An-Tr	<ul style="list-style-type: none"> Categorize artifacts based on artifact type [with appropriate Prefix artifact ID created in Step 4] 	<ul style="list-style-type: none"> PRIMARY (most commonly used artifact types) SECONDARY (less frequently used artifact types) TERTIARY (rarely created/modified artifact types)
7/2	MF-IDn:MV-IDn	<ul style="list-style-type: none"> Extract all Metadata FILED-VALUE pairs for each and every artifact Save it in a Database (xls, csv, json) GIF, IPTC, XMP, ICC and other available standard metadata from artifacts ought to be extracted 	<ul style="list-style-type: none"> Example (artifact.jpg) Aperture, Camera Model Name, Color Space, Color Tone, Contrast, Date/Time Original, Drive Mode, Exposure Compensation, File Name, File Number and File Size. (but not limited to...)
8/2	Sn:An-MF-IDn	<ul style="list-style-type: none"> Generate Metadata Field ID (with appropriate complete Prefix from Step 6) Categorize to groups namely <ul style="list-style-type: none"> MANDATORY OPTIONAL CONDITIONAL 	<ul style="list-style-type: none"> Index and Number all unique Metadata Field as <ul style="list-style-type: none"> SN:AN-MF-ID1 SN:AN-MF-ID2 SN:AN-MF-IDn
9/2	Sn:An-MF-IDn:MV-IDn	<ul style="list-style-type: none"> Generate Metadata Value ID (with appropriate complete Prefix from Step 8) Categorize to groups namely <ul style="list-style-type: none"> VALID INVALID CORRUPT 	<ul style="list-style-type: none"> Index and Number all unique Metadata Values as <ul style="list-style-type: none"> SN:AN-MV-ID1 SN:AN-MV-ID2 SN:AN-MV-IDn
10/3	SP,SG & AG	<ul style="list-style-type: none"> Authors have practically implemented the three existing similarity algorithms 	<ul style="list-style-type: none"> Raghavan, S., & Raghavan, S. V. (2017)
11/3	Unique Pockets (UP)	<ul style="list-style-type: none"> All the metadata MF-IDn:MV-IDn pair sorted under Similarity Pockets have to be ignored/excluded for this match. Each pocket should have only one unique metadata MF-IDn:MV-IDn element. All artifacts (An) can occur in more than one Unique Pocket. <ul style="list-style-type: none"> Empty filed-value IDs have to be removed. Meaningless filed-value IDs have to be decoded or converted to a generic format. 	<ul style="list-style-type: none"> <example> "Unique Pockets": "UP1": "S1A2", "UP2": "S1A3", "UP6": "S1A11", "UP7": "S1A12", "UP8": "S1A13", "UPx": "SxAx",
12/3	Unique Groups (UG)	<ul style="list-style-type: none"> All the distinct field-value pairs grouped in Similarity Groups have to be ignored/ excluded for this match. Unlike Unique pockets, at this point the Unique Groups can have two or more elements. Unique Groups are Mutually Exclusive with respect to Artifact ID (An). 	<ul style="list-style-type: none"> <example> "Unique Groups ": "S1UG1", "S1UP127","S1UP131", "S1UG2": S1UP14","S1UP141","S1UP98" "UGx": "UPx"
13/3	Unique Association (AG)	<ul style="list-style-type: none"> All the distinct field-value pairs grouped in Association Groups have to be ignored/ excluded for this match. MF-IDn:MV-IDn matches in AG can be mapped irrespective of the existing field-value pair. MF-IDn:MV-IDn matches in AG can be mapped to any other artifact across the common pool having no repetitive MF-IDn:MV-IDn from all sources. Unique MF-IDn:MV-IDn matches for duplicate copies of files have to be categorized separately for further investigation. 	<ul style="list-style-type: none"> <example> "Unique Associations ": "UA1": "S1UG2","S2UG2", "S2UG4", "UA2": "S1UG2","S4UG6","S6UG7" "UAX": "UAX"

$$S_n : A_n - M_f - ID_n : M_v - ID_n \quad (5)$$

To explain the notation with a real-world example(6), the authors consider the artifact tagged as 121 from source (Laptop) S2 and the file type be jpg with metadata field ID as ‘owner’ and the corresponding metadata value ID ‘ABC’.

$$S_2 : A_{121} - M_f - ID_{owner} : M_v - ID_{ABC} \quad (6)$$

The proposed unique metadata mapping can be established via a couple of matching principles described below.

Unique Metadata Mapping Principles

• Unique Associations

- Same-Origin Equivalence:
 - Unique metadata matches between the same source, but the metadata matches will be mutually exclusive with the existing similarity model
- Projective Equivalence:
 - Unique metadata matches projected across different metadata field-value pairs $M_f - ID_n : M_v - ID_n$ to $M_f - ID'_n : M_v - ID'_n$ across all sources as shown in Figure 3.

At the foundation mapping stage, the authors establish the notion named as a unique pocket (Up) as the set of atomic metadata field that holds a distinctive value for a specific metadata corresponding to an artifact.

Unique Pocket (UP)

A unique pocket (up_n) as in (7) is the unique subset of individually collecting all the metadata (only one of its kind) from source Sn . Unique pockets are mutually exclusive for the existing similarity pockets with a distinctive value n for $M_f - ID_n : M_v - ID_n$. In other words, all the metadata field-value pairs identified/grouped under similarity pockets are ignored and marked as out of scope while creating a unique pocket.

$$up_n = \{a_n \mid up_n \not\subset Sp, \forall n \in a_n [1, N] \text{ } \delta another \equiv n \text{ with } dM_v - ID_n \text{ for the same } M_f - ID_n\} \quad (7)$$

The list of unique pockets for metadata $M_f - ID_n$ for $n \in [1, N]$ is represented as up_n formulated in (8) be the collection of all such unique atomic pockets across all metadata field-value pairs for a source Sn .

$$up_n = \bigcup_{M_f - ID_n=1}^N up_n \quad (8)$$

up_n = union overall set of unique pockets within a single source $S_n \forall$ metadata field-value pair \neq any of the existing similarity pocket metadata matches.

$$UP = \bigcup_{S_n=1}^N \bigcup_{M_f-ID_n=1}^N \{up_n; M_f - ID_n | n \in N\} \quad (9)$$

UP = summarized in (9) is the mapping of every single unique pocket created for the entire metadata transversely for all elements from the data pool 'D'.

Unique Group (UG)

A unique group (ug_n) represented in (10,11) is the largest combination of all the unique pockets from an artifact (a_n) within the given source that is grouped with the artifact ID. So that each unique group will be the sparse collection of all the left out metadata field-value pair corresponding to a specific artifact a_n that is not grouped under any of the Existing similarity groups (Sg). The condition to be followed for each unique group is that for a selected unique pocket $up_n \subset ug_n, up_n \not\subset Sp$ there exists another unique pocket $up'_n \subset ug_n, up'_n \not\subset Sp$ such that $up_n \cap up'_n = \emptyset$.

$$ug_n = \{\cup up_n | ug_n \not\subset Sg, \forall n \in a_n [1, N] \text{ group all } up_n \in \text{same } a_n\} \quad (10)$$

$$UG = \bigcup_{S_n=1}^N \bigcup_{A_n=1}^N \{ug_n; A_n | n \in N\} \quad (11)$$

Unique Association (UA)

A Unique Association (UA) (12,13) is mapped by the largest union of unique groups between all the sources (S_n) from the data pool 'D' where (i) same-origin equivalence and (ii) projective equivalence reveals the sparse and concealed existence of missing metadata matches that were ignored in existing Association Group. The sparse artifact match for a unique association has a unique group $ug_n \forall n \in a_n [1, N]$ that is directly (i) or indirectly (ii) mapped with another unique group $ug'_n \forall n \in a'_n [1, N]$ such that there is a metadata mapping between artifact a_n and a'_n for a metadata field-value pair match namely $M_f - ID_n : M_v - ID_n = M_f - ID_n : M_v - ID'_n \forall (i)$ and $M_f - ID_n : M_v - ID_n = M_f - ID'_n : M_v - ID'_n \forall (ii)$ respectively.

$$ua_n = \{a_n | ua_n \not\subset Ag, \forall n \in a_n [1, N] \text{ group all } ug_n \forall (i) \text{ and } (ii)\} \quad (12)$$

$$UA = \bigcup_{S_n=1}^N \bigcup_{A_n=1}^N \{ua_n; A_n | n \in N\} \quad (13)$$

COMMISSIONED DATASET CONSTRUCTION (UMAM-DF 2020)

Unique Metadata Association Model - Digital Forensics Dataset (UMAM-DF 2020)

The demand for exclusive datasets as in the author's case dealing with specific digital forensic specializations like metadata analysis is the need of the hour for constructing fast and effective solutions for solving real-world cybercrime incidents by Grajeda, C., Breitingner, F., et al., (2017). Even though the credibility of the very few existing data sets from standard digital forensic research interest groups is satisfied, most of them are the datasets that spawned around decades ago. The origin of the real world or machine-generated (synthetic) datasets, the proficiency of the researcher created it, and maintaining the integrity of the dataset in a forensically sound manner plays a vital role in usability and grading a digital forensic dataset.

Researchers who are badly in need of appropriate datasets currently faces several challenges as genuine or verified datasets are rarely put in public for the research community. The above fact encourages the authors to create and contribute a metadata-specific dataset for digital forensic examiners and thanks to Phil Harvey (2005) and Jeffrey (n.d). for a solid foundation laid for extracting file metadata. This dataset is not destined to be exhaustive, but it will serve as a benchmark for understanding basic metadata composition and their analysis in the real world. The dataset has the standard metadata features like IPTC, composite, and ICC being extracted from real-world file creation and transfer activities.

The first set of experiments (1-5) are designed to monitor the addition and removal of metadata features between source devices S1,S2,S3,S4, and S5 while they are copied from primary source to supplementary source and metadata is extracted at the destination in a forensically sound manner. The mobility of these real-world evidences created as shown in Figure 2 is performed via five sequential experiments namely Ex1, Ex2, Ex3, Ex4, and Ex5 as shown in Figure 5. Experiment 1 has the genesis of artifacts i.e., the root source of the UMAM-DF dataset evidences, and it alone will have six sets of evidence to demonstrate a complete cycle from source to destination. The file transfer follows the thick arrow for the primary copy sequence and the dotted arrows denote the secondary copy (recopy) of evidences as depicted in Figure 5. It follows a sequential flow via iterations (in) performed as in (14) to cover the expected permutations of the flow of evidence from source to destination.

$$i_n = \{i_n \mid \forall n \in [1, 5]\} \quad (14)$$

Consequently, at the end of the course, the authors were able to collect five original evidences namely S1As1, S2As2, S3As3, S4As4, and S5As5 from Mobile (M), Laptop (L), Pendrive (P), Desktop (D), and Cloud(C) sources respectively. Also, a set of four supplementary evidences per experiment from source to destination (SRC2DST) is collectively created to gather twenty other sets of evidence as shown below:

Ex1: **S1As1**→M2L→L2P→P2D→D2C→C2M

Ex2: L2M←**S2As2**→L2P→P2D→D2C

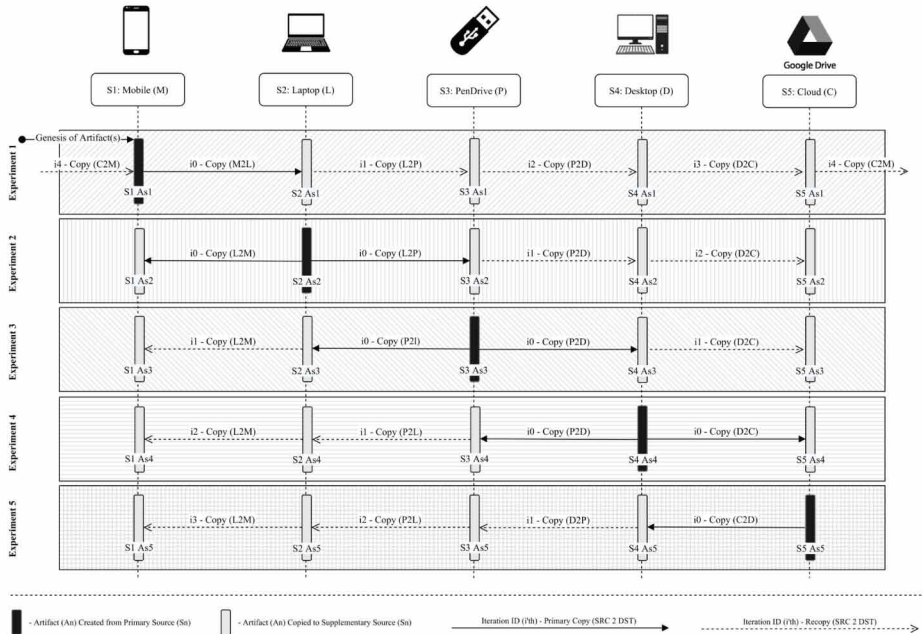
Ex2: L2M←P2L←**S3As3**→P2D→D2C

Ex4: L2M← P2L ← D2P ← **S4As4**→D2C

Ex5: L2M← P2L ←D2P←C2D← **S5As5**

The second set of experiments (6-10) are designed to monitor the extreme modification or deletion of metadata features across the standard primary source evidences (Ex 1-5) and unique destination social media platforms namely D1: Telegram(TG) as S1As6, D2: Whatsapp(WA) as S2As6, D3: Instagram(IG) as S3As6, D4: Twitter(TW) as S4As6 and D5: Facebook(FB) as S5As6. The other

Figure 5. Iterative and Sequential Mobility (of Artifacts) in UMAM-DF Dataset



considerations for dataset collection are unchanged as the first set of experiments and it results in ten unique datasets. It collects the metadata of the file before and after sharing them between source devices and social media to calculate the final Association Group (AG) and Unique Association(UA) matches are shown in Figure 6.

The authors labeled the following metadata archive as “UMAM-DF” (Unique Metadata Association Model - Digital Forensics) dataset and are made publicly available at Amrita-TIFAC-Cyber/Digital-Forensics/UMAM-DF (Unique Metadata Association Model - Digital Forensics) datasets (2020) for suggestions and recommendations to enhance the same in near future for upcoming research works.

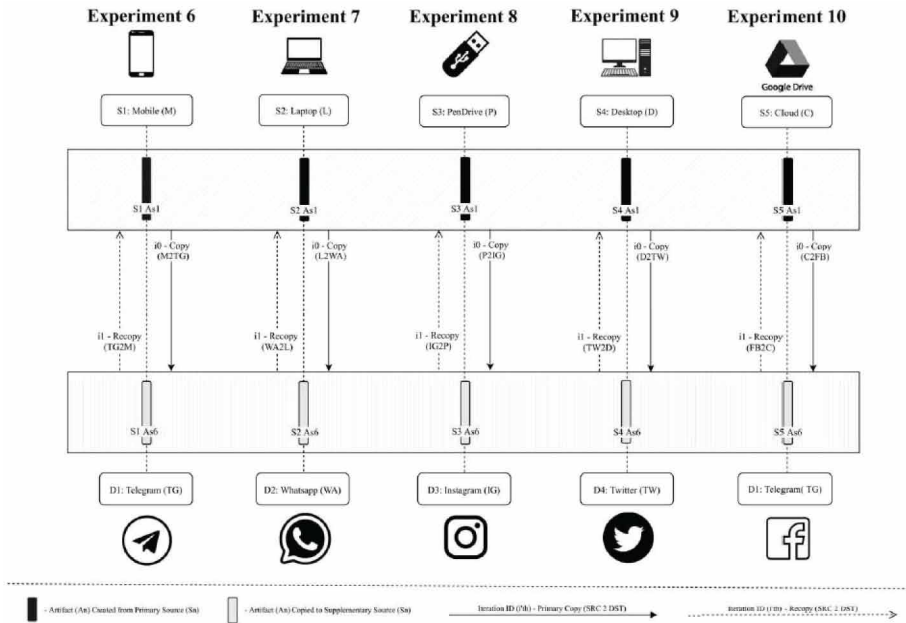
PROTOTYPE IMPLEMENTATION ON UMAM-DF DATASET

The series of sequential experiments collected with UMAM-DF dataset is engaged in testing the availability of metadata field-value pair matches across the sources with the collected set of 36 evidence sets as shared in Amrita-TIFAC-Cyber/Digital-Forensics/UMAM-DF (Unique Metadata Association Model - Digital Forensics) datasets (2020). The statistics of the similarity model and unique model of unaltered datasets are depicted in Table 5 resulting in linear Unique Group (UG) matches and variable Unique Association (UA) matches to adhere with their mathematical proof and algorithmic sequences.

The authors post a disclaimer for the repetitive values in SG produced during the experiment, as it is purely caused due to the availability of multiple identical metadata $S_n : A_n - M_f - ID_n : M_v - ID_n$ field-value pairs. This coherence can be ignored to maintain the integrity of the dataset as it is shared across the forensic community for reproducing the results as expected to verify the proposed model. The extended version of the same with normalized features is tabulated in Table 6.

Experiment 1 as shown in Figure 5 reveals the metadata matches of SP increases from 23(S1AS1) to 26(C2M) concluding that the additional metadata field-value pairs to be 22.5 and shows for every

Figure 6. Social Media Mobility (of Artifacts) in UMAM-DF Dataset



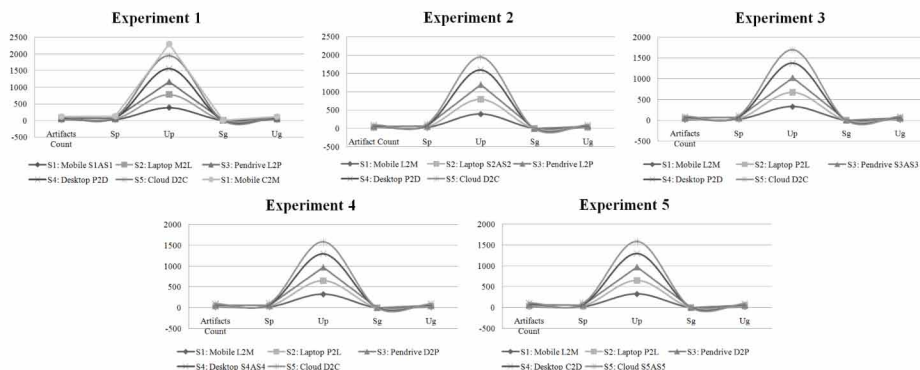
copy/paste at an average ± 2 SP is achieved. The UP count reducing from 388 at S1As1 in step 1 to 341 in step 6 reveals that around 47 unique pockets went missing when the files (namely 01.betta-left.jpg to 10.sunset.jpg) went on to a complete round from mobile, back to mobile passing all other four sources as plotted in Figure 7. The experiment 2,3,4&5 expresses a similar shift over 47,21,62&62 unique pockets respectively in UP. The UG for all the experiments varies by \pm SP across all experiments.

Unique pockets count of 380, 396, 339, 319 & 319 from source S1As6 drastically got reduced to 95,210, 96, 66 & 66 after passing via Telegram, Whatsapp, Instagram, Twitter, and Facebook

Table 5. Results for UP, UG, UA with respect to SP, SG, AG. (Unaltered UMAM-DF dataset)

UMAM-DF Dataset	Source	SP	UP	SG	UG	AG	UA
Experiment 1	S1As1	23	388	01	20	02	31
Experiment 2	S2As2	23	400	01	20	04	33
Experiment 3	S3As3	23	347	01	20	01	26
Experiment 4	S4As4	21	327	01	20	03	25
Experiment 5	S5As5	23	183	01	24	02	25
Experiment 6	S1As6	22	380	11	20	07	07
Experiment 7	S2As6	24	396	01	20	03	27
Experiment 8	S3As6	23	339	01	20	03	08
Experiment 9	S4As6	21	319	03	20	06	13
Experiment 10	S5As6	22	181	01	24	01	01
Overall Matches in SnAsn		225	3260	22	208	32	196

Figure 7. Standard Source-based metadata ratio for UMAM-DF Dataset (Ex 1-5)



respectively as plotted in Figure 8. The UG count of social media metadata shows an X/2 reduction of metadata field-value pairs resulting in the reduction of UG.

The expected outcome of metadata mapping from Association Group (Ag) of the existing system to the increased count of Unique Associations (Ua) as plotted in Figure 9 is significant. As the existing model aims at volume reduction with grouping efficiency, the proposed model at an average extended sum of all experiments performed via UMAM-DF dataset displays a 1:6 ratio of Unique Associations newly discovered for the benefit of digital forensic investigator including the left out artifacts and the sparse metadata matches between them.

The existing unaltered results and graphs as shown in Table 4 and Figures 7, 8 and 9 are maintained for reference statistics to demonstrate custom-made UMAM-DF dataset without any adulteration. Further a extended version of normalized test results are illustrated in Table 6 to show the difference mainly in SG after ignoring the completely repetitive metadata features namely 'source_id', 'File:Directory', 'File:File Access Date', 'File:File Modify Date', 'File:File Permissions', 'File:File Inode Change Date', and 'ExifTool:Exif Tool Version'.

Figure 8. Social Media metadata ratio for UMAM-DF Dataset (Ex 6-10)

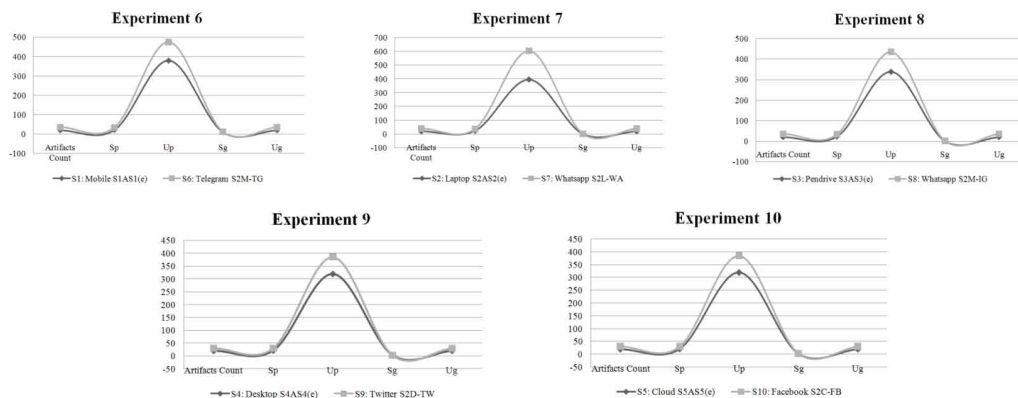
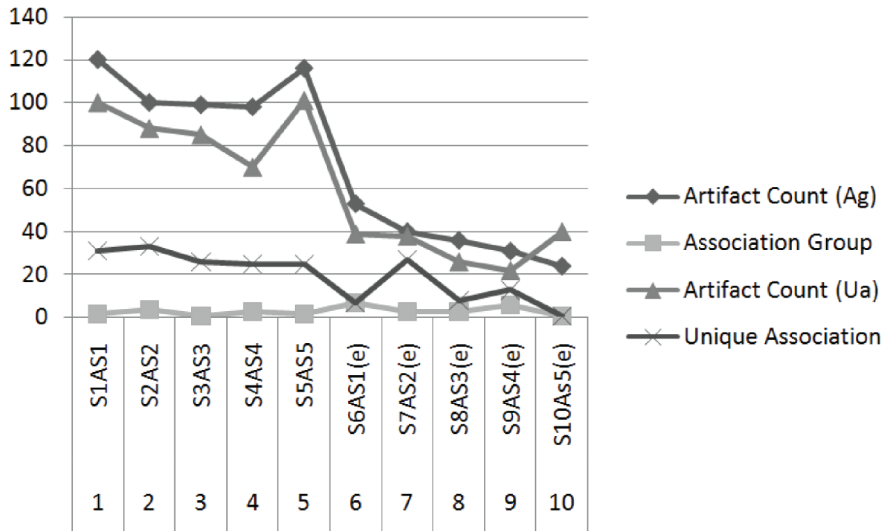


Figure 9. Overall count up from AG to UA for UMAM-DF Dataset (Ex 1-10)



CONCLUSION AND FUTURE WORKS

From the existing similarity grouping, mapping the artifacts into an average maximum of 1-7 possible AG evidence groups to an extended count of 1-33 possible combinations of the reduced groups including the sparse metadata matches is achieved. This combination of similarity and unique models might reveal all possible combinations of the potential clue that was anticipated to support the investigator's hypothesis before the court of law. The future work aims at removing or normalizing the duplicate entries that created anomalies in most of the Similarity Pocket (SP) creation. It also finds all the clues that are intentionally hidden inside sparse or unused metadata field-value pairs that might be a new trend in obfuscating the secret message between malicious threat actors, bypassing the standard security measures performed while scanning the artifacts in and out of the organizations IT

Table 6. Results for UP, UG, UA with respect to SP, SG, AG. (normalized UMAM-DF dataset)

UMAM-DF Dataset	Source	SP	UP	SG	UG	AG	UA
Experiment 1	S1As1	18	378	3	17	7	26
Experiment 2	S2As2	22	370	6	15	7	28
Experiment 3	S3As3	22	326	3	17	3	22
Experiment 4	S4As4	20	302	3	14	5	22
Experiment 5	S5As5	20	169	10	21	12	23
Experiment 6	S1As6	15	372	3	17	4	7
Experiment 7	S2As6	23	368	6	15	8	21
Experiment 8	S3As6	20	320	3	17	3	8
Experiment 9	S4As6	19	294	3	14	3	12
Experiment 10	S5As6	19	168	10	21	5	26
Overall Matches in S _n As _n		198	3067	50	168	57	195

infrastructure. It can also be extended to all heterogeneous families, groups, and types of evidences irrespective of the source, form, file state, and occurrence of the file. The authors feel that some additions in the metadata parameters of the source file would result in precise extraction and analysis of authentic digital evidences. Gopalakrishnan, A., Vineti, E., Mohan, et al., (2018) proposed a model for maintaining the integrity of evidence by heterogeneous piecewise hashing and also issuing a digital certificate named as Digital Evidence Integrity Certificate (DEIC) for evidence files. The authors wish to extend the DEIC model to add the attributes like owners public key, issuer public key, the hash algorithm used with precise version and hash of the whole file including DEIC be a part of the information that can be extracted from the metadata for verification and analysis of digital artifacts. Chhabra, G. S., Singh, V. P., et al., (2018) demonstrated a generic forensic framework has been put forwarded which uses MapReduce in Big Data that can be extended to metadata classification and analysis

Thanekar, S., Subrahmanyam, K., & Bagwan, A. (2016) explained the problems faced by the investigator when they try to analyze the crime scene, the investigator has to deal with a huge volume of data such as when retrieving the data from the suspicious system, metadata analysis, etc., where these data gives more information for evidence. In such cases, the investigation requires storage to store high volumes of data and also requires processing speed. Hadoop is a technology that stores data on disk and in memory. To identify the evidences on big data, initially, the authors need to understand the structure of the big data, through which the authors can find the evidences more efficiently. By using the different tools and technology the authors can do the forensic of Big Data. As in Big Data, the volume of data to be considered is very big, an automated tool can help us to do it efficiently. Mohammed, H., Clarke, N., et al., (2016) proposed a new novel framework for digital forensic analysis of heterogeneous Big Data has been introduced. This mainly focuses on the investigations of three core issues such a data volume, heterogeneous data, and the investigator's cognitive load in studying the relationships between artifacts. This framework is used to assess the possibility of using metadata and semantic web-based ontologies by Mohan, A. K., & Venkataraman, D. (2017, January) to solve the problems of big data and heterogeneous data sources correspondingly. The advantage of extending these two models to an ontology-based forensics analysis will help in a greater way for evidence examination. The author's model can be extended to Akremi, A., Sriti, M. F., et al., (2020) ontology-based forensic analysis for fast and effective metadata association models.

The authors extend a note on public interest over the applicability of the proposed unique models on multimedia contents shared by pedophiles on social media to be monitored by law enforcement agencies similar to the work of Amuchi, F., Al-Nemrat, A., et al., (2012, October) done on chat contents. The near future of cyber-crime investigation will witness many such similarity models to effectively analyze and identify the sparse associations between digital artifacts. The reliability of the evidence is based on the efficiency of feature extraction, evidence collection, and their transformation with the similarity and unique models. The authors work can also be extended to Du, X., Le, Q., & Scanlon, M. (2020, June) techniques that produce a relevancy count for individual artifact similarity using individual file system metadata and their connected MAC timestamp actions carried out by the perpetrator of a digital crime scene.

ACKNOWLEDGMENT

The research work is supported by the Ministry of Electronics and Information Technology, Government of India initiative Visvesvaraya Ph.D. Scheme for Electronics and IT intended for IT/ITES under Cyber Security category in cooperation with TIFAC-CORE in Cyber Security research centre at Amrita Vishwa Vidyapeetham, a private deemed-to-be-university and Institute of Eminence based on Coimbatore, Tamil Nadu, India.

REFERENCES

- Agrawal, N., Bolosky, W. J., Douceur, J. R., & Lorch, J. R. (2007). A five-year study of file-system metadata. *ACM Transactions on Storage*, 3(3), 9. doi:10.1145/1288783.1288788
- Akreml, A., Sriti, M. F., Sallay, H., & Rouached, M. (2020). Ontology-Based Smart Sound Digital Forensics Analysis for Web Services. In *Digital Forensics and Forensic Investigations: Breakthroughs in Research and Practice* (pp. 497-520). IGI Global.
- Amato, F., Castiglione, A., Cozzolino, G., & Narducci, F. (2020). A semantic-based methodology for digital forensics analysis. *Journal of Parallel and Distributed Computing*, 138, 172–177. doi:10.1016/j.jpdc.2019.12.017
- Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2011). A sift-based forensic method for copy-move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, 6(3), 1099–1110. doi:10.1109/TIFS.2011.2129512
- Amrita-TIFAC-Cyber/Digital-Forensics/UMAM-DF (Unique Metadata Association Model - Digital Forensics) datasets. (2020)<https://github.com/Amrita-TIFAC-Cyber/Digital-Forensics>
- Amuchi, F., Al-Nemrat, A., Alazab, M., & Layton, R. (2012, October). Identifying cyber predators through forensic authorship analysis of chat logs. In *2012 Third Cybercrime and Trustworthy Computing Workshop* (pp. 28-37). IEEE. doi:10.1109/CTC.2012.16
- Balasubramanian, V., Doraisamy, S. G., & Kanakarajan, N. K. (2016). A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems*, 46(1), 121–145. doi:10.1007/s10844-015-0356-5
- Bhattacharya, S., Kaluri, R., Singh, S., Alazab, M., & Tariq, U. (2020). A Novel PCA-Firefly based XGBoost classification model for Intrusion Detection in Networks using GPU. *Electronics (Basel)*, 9(2), 219. doi:10.3390/electronics9020219
- Buchholz, F., & Spafford, E. (2004). On the role of file system metadata in digital forensics. *Digital Investigation*, 1(4), 298–309. doi:10.1016/j.diin.2004.10.002
- Casey, E. (2009). *Handbook of digital forensics and investigation*. Academic Press.
- Chhabra, G. S., Singh, V. P., & Singh, M. (2018). Cyber forensics framework for big data analytics in an IoT environment using machine learning. *Multimedia Tools and Applications*, 1–20.
- Daniel, L., & Daniel, L. (2012). Digital forensics for legal professionals. *Syngress Book Co, 1*, 287–293.
- Databook: Turning Big Data into Knowledge with Metadata at Uber. (n.d.). <https://eng.uber.com/databook/>
- Du, X., Le, Q., & Scanlon, M. (2020, June). Automated Artefact Relevancy Determination from Artefact Metadata and Associated Timeline Events. In *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)* (pp. 1-8). IEEE. doi:10.1109/CyberSecurity49315.2020.9138874
- Gopalakrishnan, A., Vineti, E., Mohan, A. K., & Sethumadhavan, M. (2018). The Art of Piecewise Hashing: A Lemma Toward Better Evidence Provability. *Journal of Cyber Security and Mobility*, 7(1), 109–130. doi:10.13052/jcsm2245-1439.719
- Grajeda, C., Breiting, F., & Baggili, I. (2017). Availability of datasets for digital forensics—And what is missing. *Digital Investigation*, 22, S94–S105. doi:10.1016/j.diin.2017.06.004
- Krishnan, K. (2013). *Data warehousing in the age of big data*. Newnes.
- Liu, Z., Lai, Z., Ou, W., Zhang, K., & Zheng, R. (2020). Structured optimal graph-based sparse feature extraction for semi-supervised learning. *Signal Processing*, 170, 107456. doi:10.1016/j.sigpro.2020.107456
- Mohammed, H., Clarke, N., & Li, F. (2016). *An automated approach for digital forensic analysis of heterogeneous big data*. Academic Press.
- Mohan, A. K., & Venkataraman, D. (2017, January). Forensic future of social media analysis using web ontology. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-6). IEEE. doi:10.1109/ICACCS.2017.8014682

Phil Harvey. (2005). <https://exiftool.org/>

Prem, T., Selwin, V. P., & Mohan, A. K. (2017, April). *Disk memory forensics: Analysis of memory forensics frameworks flow*. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE.

Raghavan, S. (2013). Digital forensic research: Current state of the art. *CSI Transactions on ICT*, 1(1), 91–114. doi:10.1007/s40012-012-0008-7

Raghavan, S. (2014). *A framework for identifying associations in digital evidence using metadata* (Doctoral dissertation). Queensland University of Technology.

Raghavan, S., Clark, A., & Mohay, G. (2009, January). FIA: an open forensic integration architecture for composing digital evidence. In *International Conference on Forensics in Telecommunications, Information, and Multimedia* (pp. 83-94). Springer. doi:10.1007/978-3-642-02312-5_10

Raghavan, S., & Raghavan, S. V. (2009, September). Digital Evidence Composition in Fraud Detection. In *International Conference on Digital Forensics and Cyber Crime* (pp. 1-8). Springer.

Raghavan, S., & Raghavan, S. V. (2013a, November). A study of forensic & analysis tools. In *2013 8th International Workshop on Systematic Approaches to Digital Forensics Engineering (SADFE)* (pp. 1-5). IEEE. doi:10.1109/SADFE.2013.6911540

Raghavan, S., & Raghavan, S. V. (2013b, November). AssocGEN: engine for analyzing metadata-based associations in digital evidence. In *2013 8th International Workshop on Systematic Approaches to Digital Forensics Engineering (SADFE)* (pp. 1-8). IEEE. doi:10.1109/SADFE.2013.6911541

Raghavan, S., & Raghavan, S. V. (2014). Eliciting file relationships using metadata-based associations for digital forensics. *CSI Transactions on ICT*, 2(1), 49-64.

Raghavan, S., & Raghavan, S. V. (2016, January). Reconstructing tabbed browser sessions using metadata associations. In *IFIP International Conference on Digital Forensics* (pp. 165-188). Springer. doi:10.1007/978-3-319-46279-0_9

Raghavan, S., & Raghavan, S. V. (2017). Analytics using metadata associations for digital investigations. *CSI Transactions on ICT*, 5(3), 315–338. doi:10.1007/s40012-017-0174-8

Raghavan, S., & Saran, H. (2013, November). UniTIME: Timestamp interpretation engine for developing unified timelines. In *2013 8th International Workshop on Systematic Approaches to Digital Forensics Engineering (SADFE)* (pp. 1-7). IEEE.

Rajendiran, K., Kannan, K., & Yu, Y. (2020). Applications of Machine Learning in Cyber Forensics. In *Confluence of AI, Machine, and Deep Learning in Cyber Forensics* (pp. 29-46). IGI Global.

RM, S. P., Maddikunta, P. K. R., Parimala, M., Koppu, S., Reddy, T., Chowdhary, C. L., & Alazab, M. (2020). An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Computer Communications*.

Thanekar, S., Subrahmanyam, K., & Bagwan, A. (2016). A study on digital forensics in Hadoop. *Indonesian Journal of Electrical Engineering and Computer Science*, 4(2), 473. doi:10.11591/ijeecs.v4.i2.pp473-478