International Journal of Pure and Applied Mathematics
Volume 119 No. 16 2018, 4063-4072
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Gathering Hazy Dots In Favour Of Digital Evidences

Aishwarya R
TIFAC-CORE in Cyber Security
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
ashravi94@gmail.com

Ashok Kumar Mohan
TIFAC-CORE in Cyber Security
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
m_ashokkumar@cb.amrita.edu

M. Sethumadhavan
TIFAC-CORE in Cyber Security
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
m_sethu@cb.amrita.edu

Abstract— **Metadata is commonly quoted as 'data within data' and the same can be accustomed in digital forensics as 'evidence within evidence'. Metadata plays a very important roles in digital forensics from seizure to documentation. It can provide substantiate information about digital media to locate possible evidences that looks like hazy dots in crime scene investigation. It can also disclose the information on whether someone had tried to hide, delete, or obscure the data. It can be used to automatically correlate documents from different sources. MAC timestamp helps in reconstructing the timeline when the actual incident started based on the modified, accessed and created time slots of the evidences. Tracing of the path will be difficult while analyzing evidence pool with terabytes and petabytes of evidences for a single case. In order to simplify this normalization sequence, argument or proof is obtained via a series of lemma's and sorted out into three categories namely similarity packet, similarity group and association group. These lemma's along with the metadata association reduces the number of redundant metadata and hence simplifies the process to trackback the origin. The efficiency with respect to this method is evaluated from heterogeneous files via are collected from five different sources as the evidence pool. Based on the efficiency observed, this process can be tweaked accordingly and can be automated for solving crime scenes related to digital evidences in a considerably less amount of time and resources put into practice.**

Keywords— *Metadata Association, MAC timings, Digital Forensic Investigation, Similarity pocket, Similarity group, Association group*

I. INTRODUCTION

Digital forensics deals with the recovery and investigation of evidences collected from digital devices. The basic steps followed in digital forensics are to recognize the need, define the problem, collect data, analyze the data, develop, test and select the hypothesis [1]. Something made by or using an art is termed as an artefact (noun: artifact) and it is commonly cited for digital evidences in cyberspace.

The datasets for an investigation starts with the collection of digital artifacts including, files like office document files (docx, xlsx, pptx and mdb), portable document format (pdf), all types for images (jpeg, tiff, svg, ico, bmp and png), audio (.acc and mp3), video (avi and mpg), executables (exe, bin, bat, dmg and apk) and compressed archives (zip, rar and tar.bz).These are the majority of file types analyzed in crime scenes but are not limited to extensions alone. Each follows different format to store the metadata [2] which is a vital information of each document. It has both advantages as well as disadvantages as it might disclose sensitive information [3]. The opening step in investigation mandates the verification of evidences over the CIA triage. The artifacts ought to be analyzed for Confidentiality, Integrity and Availability (CIA) of the resources by extracting and analyzing its own metadata information. Locard's Exchange principle is the basic kick starter for all forensic science investigation. As Dr. Edmond Locard coined it as follows, "Every contact leaves a trace". So eventually at least any one of the CIA factor over an artifact is directly influenced by the perpetrator (suspect) of a crime. Even if the actual data remains unaffected such as no change in its checksum or hash, but the corresponding metadata reflects the malicious action carried out on the evidence.

During the course of a digital forensic investigation metadata responds to the investigator queries based on What (is the data), When (it is created/accessed/modified), Where (it is stored), Who (created it), How (it is created), Which (type of data it is) and Why (combination of all above questionnaires will reassemble the answer for this) [19]. The best practices is to make use of metadata to locate a precise evidence from a pool of massive heterogeneous data

sources from all three types of standard metadata types namely administrative type, descriptive type and structural type. The ultimate need for metadata in digital forensic is that it assists proper forensic analysis via drawing attention to file patterns, ascertain timeline over a period of prosecution, and helps to connect the dots in identifying the anticipated data.

Metadata offers additional evidence that is capable of identifying any changes made in an evidence. There are various tools available for each type of storage devices. Taking a forensic image of any evidence is best way to maintain the integrity of the data [4]. Any file that is accessed does not affect the integrity of the file as there is no change made to the original file, but it changes the accessed time in MAC parameters of file metadata [6]. MAC timing corresponds to Access time (A-time) , Modification time (M-time), Creation time (C-time). Formulating is unique to its operating system and the corresponding file system in which it stores the MAC times. Metadata in different time zones acts as a nightmare for digital forensic investigators as date and time formats are purely dependent on the file system and the user who access it.

II. Applicability of metadata in File System

File Allocation Table (FAT) file system utilises only tables and does not work well with directories when it stores large number of files. While New Technology File System (NTFS) uses tables for storing metadata and it uses B-tree for directories to find files quickly. XFS uses small tables for metadata and b-trees to find the required table. B-TRee File System (BTRFS) and Hierarchical File System (HFS) by default employs only B-trees. However in all file system, it maintains an index like record which stores metadata information without considering how the search for this record is organized and what pointer is directed to this particular record. Once the index like record is lost, damaged or missing then it is impractical to recover the whole of metadata in its actual form in which the resident data structure have stored
it locally.

III. STANDARD METADATA DATASET AND MAC TIMINGS

When there is large amount of data it is difficult to access all of them individually, so the simplest way is to create datasets which makes classification easy [7]. A standard is followed for collecting the metadata datasets which is called as PREMIS. It is commonly known as the implementation strategies working group, responsible for setting standards on preservation of metadata. It also acts as the data dictionary that defines a collection of metadata where the majority of the repositories archiving digital objects will be recorded and preserved for future use. It is the benchmark formulation of identifying the core metadata elements always expressed in terms of XML schemas. The model has its root on the OAIS and expresses the relationship among five digital preservation entities that is to refer intellectual entities, objects, events, agents and rights. MAC timestamp has dissimilar declaration for all diverse file formats.

In FAT file system, the creation time stores at a rate of every ten milliseconds, the last modified time at a rate of every two seconds and the access time at a rate of every single day. Whereas in NTFS file system, the creation time and the last modified/access time is recorded at rate of every hundredth nanosecond, but the changes in access time can be stored in memory capable of surviving up to an hour before the file is being pushed to the disk. When a file is renamed MAC timestamp does not get affected. When moved in between folders it only changes the modified time and between disks the accessed time gets changed. When copied the accessed and created time changes. When modified only the modified time changes. When created a new MAC time is added and finally when it is deleted there are no changes.

IV. Association between metadata

In general, a typical digital evidence will provide a MAC timestamp information which plays an important role as it is possible to identify whether the data is tampered or not. They follow a unified timeline in order to identify if the file is tampered or not. When the file is analyzed in Windows system then the created, accessed and modified information is setup accordingly whereas in Linux system it is changed to Unix epoch time format. Thereby, in this way the evidences from heterogeneous sources are digitally mapped on reconstructing the timeline to solve the crime.

Timelines are generally obtained from the metadata of the digital artifacts facing the following challenges while dealing with timestamps across heterogeneous sources. The timelines are recorded differently for different digital artifacts and so the representation and resolution of timestamps will also differ accordingly. Here the origin of timestamp is mandated and upon failure it can lead to inconsistency. In case of inconsistencies the PIM (Provenance Information Model) is used to normalize the timestamp to interpret it to a generic standard format.

At the time of investigation, it is required to know the nature of the artifacts and also how they are related to each other. Using keyword searches it is not possible to cover all the possible associations between the artifacts and logically opposite result is expected if it is known in prior, as the grouping is likely to be

unsuccessful. For example, when any incidents take place there is chance that any of the two files might be related, if so then the time interval between them will give an idea whether it is related or not [15] and [16].

Metadata is essential for developing the semantic web which can be used for decision making and to gain results. Using this method it is possible to keep track on various evidences [12]. In case, there occurs any investigation based on social media the concept of Web Ontology will be useful as it can generate the user activity information, which can be compared with the suspect information to see if any relation can be found [13].

In near future, the concept of metadata will be viral in social media to find solutions to most of the investigation as it can give the largest possible information of an individual. The universal storage and faster access of all these kind of data cannot be made possible locally, so it is recommended to employ the intuition of cloud services which is a simple and scalable method to store, access and share the data globally.

V. EVALUATION OF ACADEMIC DEBATES

A file system is basically used for controlling the data that is stored and retrieved on the storage medium. Without the prior knowledge on the working of the file system, the metadata collected for mapping the evidences will be of no use or it might conclude with a false result. When a files metadata information is stored and accessed it is observed that they are handled uniquely over different file systems.

In the traditional system, the digital evidence are examined to describe the condition of the files and simple relation between them in the case of digital forensics. Merely by searching the keywords or using the regular expression, the information about the set of files are retrieved. Then at the time of analyzing the files, it is essential to determine the type of association that it has with the neighbours and also to get the answer for the elementary questions like what, when, where, how, why and who. Hence, to make the above process simpler the concept of metadata is being used, which is automated and also generates unique values that can be stored and retrieved outside the source file but travels along everywhere piggybacking with the file [9].

A forensic integration architecture (FIA) is a framework used to extract the evidence source and store data with various origin from diverse sources. It finds the evidence information from a variety of sources that allows an investigator to rebuild the past by constructing the events as a well structured timeline. It focuses on a distinctive approach facilitated by a number of layers in a specific sequence. It is open and extendable which makes it simpler to embrace all runtime technological changes [10].

At the time of investigation, it is required to analyse the digital artifacts which were modified, accessed and created which is in the proximity at the time of incident response. However, it is not necessary to rely only on succession timestamps across the sources, since timestamps have an assortment of representation across various time zones and it also depends on the file system. Basically, there are three category of challenges that appear while dealing with timestamps from heterogeneous sources i.e., time of modification , time of accessing and time of creation of the file [8].

The difference in the origin of each and every digital evidence sources will result in different events and for that reason the representation and determination of the timestamp be at variance. So we need to get a unified timestamp value for getting accurate result. Provenance Information Model (PIM) is one which records the time data associated with a homogeneous source to analyze the shifts based on the individual timestamp to get values with respect to a single time zone. This concept is used in a framework called UniTIME which accepts the digital evidence as input and convert them into one or more homogeneous sources based on the PIM information. Unified timeline is created by mapping all timestamps with respect to PIM[8].

The three major phases of UniTIME are as shown below:

1. To determine the uneven UTC time value for timestamps by mapping Provenance Information Model to the respective homogeneous sources
2. To determine the local time zone information built on PIM
3. To confirm the timestamp-based declarations which is recorded in PIM for each homogeneous source

There is another similar development called AssocGEN to smooth the progress of digital investigation to determine the original file after the unified timeline is obtained. It is an essential analysis engine that uses the metadata in order to determine the association between artifacts which fit into the evidence files, system logs and network pcap dumps. It is a Java based cross-platform cooperative prepared for the identification of metadata associations to group the related artifacts based upon the investigators ease of use [11]. There are three basic layers in FIA for abstraction: the digital evidence layer, metadata parser layer cum digital artifact traversal and the evidence composition layer.

AssocGEN is constructed considering the three main algorithms namely Similarity Pocket, Similarity Group and Association Group which is the foundation for the approach over metadata association scheme proposed in this execution.

VI. APPROACH FOR METADATA ASSOCIATION

Analytics using metadata associations for digital investigations[5] is the theoretical approach on forming the metadata association between the artifacts with a better efficiency.

The metadata association between the digital artifacts using the 13 Lemma's is proved to be more efficient based on theoretical approach. The Lemma's are as follows:

a. Lemma 1

The minimum size of a similarity pocket for any metadata k ∈ [1,*M*] *is* 2.

***b.* Lemma 2**

The maximum size of a similarity pocket for any metadata k ∈ [1,*M*] is N_i, where N_i is the number of artifacts in source s_i.

***c.* Lemma 3**

The similarity pockets $sp_t^{ik}{}'s$ are mutually exclusive for all $t \in [N]$

d. Lemma 4

The largest number of similarity pockets which can be formed for some metadata $k \in [1,M]$ is $N_i/2$.
In this occurrence ,the total values of similarity pocket is limited to $1 \leq |sp_k^i| \leq (N_i - n)/2.$

e. Lemma 5

The largest number of similarity pockets that can be created across all M metadata is $MN_i/2$.
Similar to the above Lemma , this is result for the set of all metadata and sources.

f. Lemma 6

The similarity groups $sg_t^i{}'s$ are mutually exclusive for all $t \in N$. When two similarity groups sg_x^i and sg_y^i is taken for $x, y \in \mathbb{N}$ it is observed that the similarity pocket for sg_x^i and sg_y^i does not co-incide, indicating that the artifacts do not have same set of artifacts. Hence, $sg_x^i \cap sg_y^i = \emptyset \; \forall \; x, y \in \mathbb{N}$.

g. Lemma 7

The colossal number of similarity groups which can be present in a given source is $N_i/2$.
These metrics are reduction factor and grouping efficiency.
Reduction Factor = Ratio of similarity group by total number of similarity pockets i.e., 1/2M.
Grouping efficiency = 1.0 - 1/2M.

***h.* Lemma 8**

The reduction factor r_{sp} and the grouping efficiency n_{sp} are independent of the mean.

i. Lemma 9

The grouping efficiency n_{sp} and the reduction factor r_{sp} are independent of the variance**.**

j. **Lemma 10**

The grouping efficiency n_{sp} and the reduction factor r_{sp} are independent of the distribution.

k. Lemma 11

The reduction factor is directly proportional to $|sp_k^i|$ and the constant of proportionality is $\frac{1}{MN_i}$.

***l.* Lemma 12**

The reduction factor $r_{sg} = \frac{1}{(\mu+1)}$.

m. Lemma 13

r_{sp} and n_{sg} are independent of variance.[5]

VII. POC USING METADATA ASSOCIATION

In this case, the "Blue Whale Game" has been considered for collecting the datasets and the result is obtained using the metadata values for all file types. In this process of collecting the datasets, five different sources were considered namely a desktop computer, personal laptop, mobile device, cloud storage and a USB device. All the different set of files as mentioned below were collected in these respective devices and the three algorithms were applied to trace back to the original file. The datasets that are considered for this case is a part of an entire dataset that consisted of the following across the five different sources shown in the Table 1.

TABLE I

File Type	Count	Category/Vendor
ppt	3716	MS office document
jpg	15327	Image file
pdf	14483	Portable document file
apk	10	Application file
zip	5225	Archive
png	46445	Image file

mp4	62	Video file
gif	19806	Image file
doc	5612	MS document file
mp3	24	Audio file
bmp	321	Image file

During an investigation, analysis is done on the acquired source of information which are a collection of digital data denoted as D, said to be a set of sources and s_i is the i_{th} source acquired for analysis

D = { S_1,S_2,….S_5}

In our investigation the sources are
S_1 = Desktop
S_2 = Laptop
S_3 = Mobile
S_4 = Cloud
S_5 = USB

Each source contains one or more number of digital artifacts represented as $N_1, N_2, N_3, \dots N_{100.}$ Here, the artifacts ranges from 1 to 100.

For which each digital artifact in a source is referred as $a_1^i, a_2^i, a_3^i, \dots$ and so on. Each of these artifacts have metadata values, which is referred as m_j^i.

$S_i = \{ a_j^i | a_j^i$ is j[th] digital artifact belonging to sources $S_{i,j} \in [1, N] \}$ i ∈ [1,S].

Source S_i can also be represented as a collection of metadata of each artifacts that belong to S_i,

$S_i = \{ m_j^i | \forall j \in [1, N_i] \}$ m_j^i is the metadata vector such that $m_j^i \leftrightarrow a_j^i \}$ $i \in$ [1,S].

Further the metadata can be depicted as M-vector where m_k^{ij} is the kth value.

$m_j^i = \{ m_k^{ij} | k \in [1, M] \}$ j ∈ [1,N], i∈[1,S]

If the metadata size is not constant then the remaining values in the vector M is appended with null.

A. Similarity Pocket

Given: A Source S_1, S_2, S_3, S_4, S_5
$A^i = a_1^i, a_2^i, \dots . a_{100}^i$ set of digital artifacts of source s_i
$M^i = m_1^i, m_2^i, \dots . m_{100}^i$, set of metadata vectors corresponding to A_i, i = ranges from 1 to 100

Output: sp_i, set of all similarity pockets generated on source s_i.

If $A_i \rightarrow [a_1, a_2, a_3, \dots a_{100}]$ of Source S_1(Windows)

2. If $m_1 \rightarrow [m_1^{11}, . m_5^{15}, . m_1^{1100}, . m_5^{1100}]$ of S_1

Then

$sp_t^{11} \rightarrow [a_1, a_3, a_6 \dots]$
$sp_t^{12} \rightarrow [a_{12}, a_{34} . .]$
$sp_t^{13} \rightarrow [a_{41}, a_{2}, . .]$
$sp_t^{14} \rightarrow [a_{56}, a_{70} \dots]$
$sp_t^{15} \rightarrow [a_{90}, a_{10} \dots]$

Repeat the above steps for S_2, S_3, S_4, S_5.
$sp^i \rightarrow \{ sp_t^{11}, . sp_t^{15}, . sp_t^{21}, . sp_t^{25}, .$
$sp_t^{31}, . sp_t^{35}, . sp_t^{41}, . sp_t^{45}, . sp_t^{45}, . sp_t^{51}, . sp_t^{55} \}$

B. Similarity Group

Given : $sp_i = [sp_1^{1k}, sp_2^{2k}, \dots, sp_{100}^{5k}]$, set of similarity pockets generated for all metadata $k \in [1,5]$ on source s_i.

Output : sg_i, set of all similarity groups generated on source s_i.

1. If $sp_t^{ik} = [a_1, a_4, \dots . a_{100}]$ of source S_1 (Windows) for all metadata
then
2. $sg_t^1 = [sp_t^{1k}, sp_t^{2k}, sp_t^{3k}, sp_t^{4k}, sp_t^{5k}]$ of source S_1.
3. Repeat above for all sources S_2,S_3,S_4,S_5
4. sg_i = set of all sg_t^i for all sources and all metadata

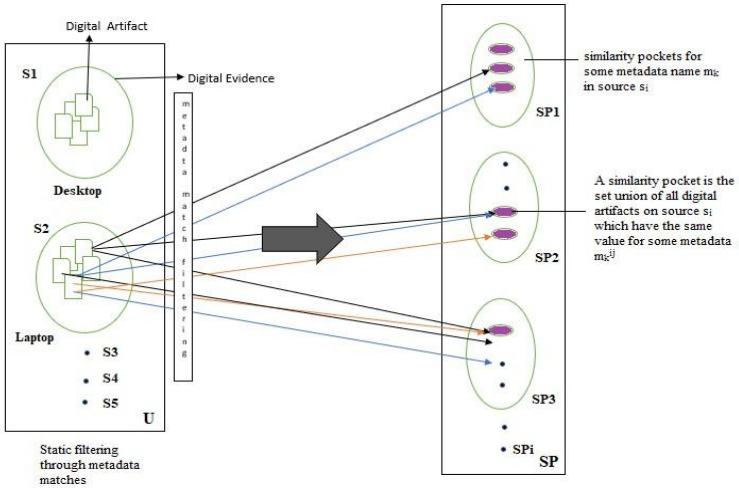


Figure. 1: Similarity Pocket

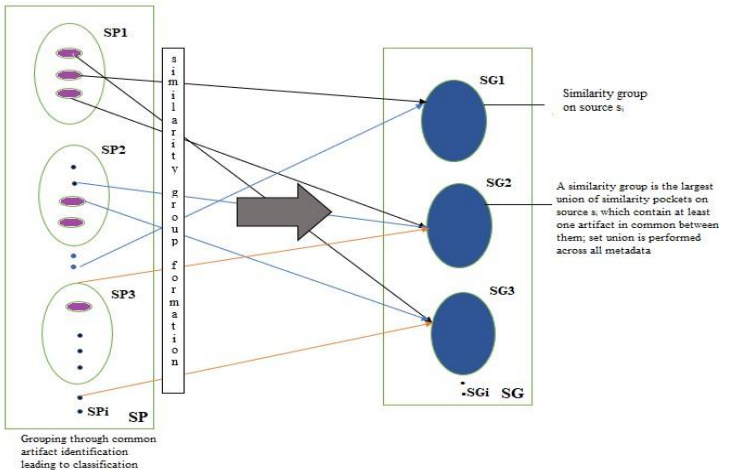


Figure. 2: Similarity Group

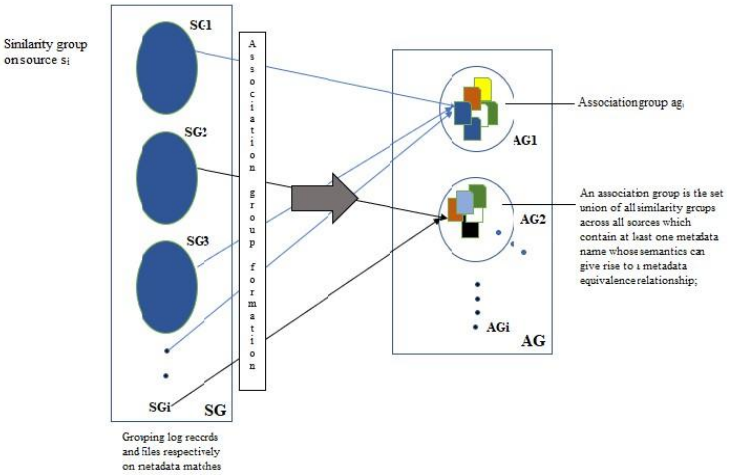


Figure. 3: Association Group

C. Association group

Given : $SG = [sg_1^i, sg_2^i, \dots sg_{100}^i]$ set of all sets of similarity groups across all sources s_i for $i \in [1,S]$ in D.

Output : AG, set of all association groups generated on D

1.If $sg_t^1 \in SG$ for all sources and all metadata then
2.ag_t is union of all sg_i with minimum one metadata match
3. AG is set of all ag_t for all sources

VIII. Results

According to the theoretical approach using the lemma facts the values of similarity pocket with respect to number of metadata has to go on increasing with increase in metadata value but practically it might not be possible, it may generate slightly different values.

Hence, the theoretical approach is assumption based on lemma's and when implementing such concept it has to be verified practically to know its nature by calculating the efficiency. Similar considerations will be in the case of similarity group with respect to the number of sources.

IX. Conclusion and Future Work

MAC timestamp plays a very important role in classifying and associating the metadata. The actual timeline is also reconstructed by using the MAC timestamp by arranging it in sequential order. The lemma facts and metadata association made is easy to evaluate and efficiency of the expected result is expected to be better than the traditional method. Our work in progress is to apply the algorithms and proposed Lemma's to the dataset collected in large scale of evidence pools. The efficiency will be analyzed, so as to propose an automated process to reduce the time taken and other resources invested in solving a cyber crime.

REFERENCES

[1] Raghavan, S. (2013). Digital forensic research: current state of the art. *CSI Transactions on ICT*, 1(1), 91-114.

[2] Orozco, A. L. S., Gonzlez, D. M. A., Villalba, L. J. G., Hernndez-Castro, J. (2015). Analyzis of errors in exif metadata on mobile devices. *Multimedia Tools and Applications*, 74(13), 4735-4763

[3] Payne, D. (2013). Metadata: The good, the bad, and the misunderstood. *GPSolo*, 30, 51.

[4] Raghavan, S., Raghavan, S. V. (2013, November). A study of forensic analyzis tools. *In Systematic Approaches to Digital Forensic Engineering (SADFE)*, 2013 Eighth International Workshop on (pp. 1-5). IEEE

[5] Raghavan, S., Raghavan, S. V. (2017). Analytics using metadata associations for digital investigations. *CSI Transactions on ICT*, 5(3), 315-338

[6] Castiglione, A., De Santis, A., Soriente, C. (2007). Taking advantages of a disadvantage: Digital forensics and steganography using document metadata. *Journal of Systems and Software*, 80(5), 750-764.

[7] Habermann, T., Kozimor, J., Gordon, S. (2017). Metadata Evaluation and Improvement: Evolving analyzis and Reporting.

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

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

[10] 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, Berlin, Heidelberg.

[11] Raghavan, S., Raghavan, S. V. (2013, November). AssocGEN: engine for analyzing metadata based associations in digital evidence. *In Systematic Approaches to Digital Forensic Engineering (SADFE)*, 2013 Eighth International Workshop on (pp. 1-8). IEEE.

[12] Venkataraman, D., Bharathi, P. D. (2017, January). Resource tracking in institutions using semantic web. *In Advanced Computing and Communication Systems (ICACCS)*, 2017 4th International Conference on (pp. 1-4). IEEE.

[13] Mohan, A. K., Venkataraman, D. (2017, January). Forensic future of social media analyzis using web ontology. *In Advanced Computing and Communication Systems (ICACCS)*, 2017 4th International Conference on (pp. 1-6). IEEE

[14] Raghavan, S., Raghavan, S.V (2013). Determining the origin of downloaded files using metadata associations. *Journal of Communications*, 8(12), 902-910

[15] 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, Berlin, Heidelberg.

[16] V. Sivasangari, Ashok Kumar Mohan, K. Suthendran and M. Sethumadhavan, "Isolating Rumors Using Sentiment Analysis", Journal of Cyber Security and Mobility,Vol.7,pp.181-199, 2018.