CSIT (March 2014) 2(1):49–64
DOI 10.1007/s40012-014-0046-4

Eliciting file relationships using metadata based associations for digital forensics

Sriram Raghavan · S. V. Raghavan

Received: 5 September 2013 / Accepted: 27 June 2014 / Published online: 1 August 2014

Abstract In the conventional system of analysis that is concerned with digital forensics, content is analyzed to describe the state of files in digital evidence and ascertain their relevance. Such content analysis is carried out using "searching". When searching a file or for a file, use of keywords is the norm. When the exact words are not known, one may use regular expression search which uses a more flexible language for describing a set of keywords that fit a pattern. During analysis, there is also a need to identify all types of associations that exist between the files to answer the six fundamental questions of what, when, where, how, who and why. If the keywords and pattern have limited scope, an examiner often has very little to go on. Metadata contains information that represents the state of a file, even if partially. Besides, metadata based search is amenable to automation by virtue of the ubiquitous nature of metadata. During analysis, metadata can be used to ascertain the nature of digital photographs that were processed using software and identify digitally generated images that resemble original photographs. Metadata can also be used to identify word processing documents that were derived from other documents and stored as a duplicate or after modification in such a way that traditional techniques cannot detect. Often what is needed is the ability to identify section(s) of the evidence where relevant information appears to reside. Metadata based matches give rise to file relationships that encapsulate the event sequence among related files aiding in the discovery. This paper proposes a method to automatically identify associations among the files in digital evidence at the syntactic and semantic levels using metadata. We apply this method to identify metadata associations from collections of image files and word processing documents and elicit inter-file relationships for the purpose of identifying interesting or relevant files from large file collections in digital evidence. We demonstrate that the file relationships identified using metadata help in the identification of doctored photographs and copied documents.

Keywords Metadata · Metadata association · Metadata family · File relationship · Association group · Association index

S. Raghavan (✉)
Secure Cyber Space, Brisbane, QLD, Australia
e-mail: sriram.raghavan@securecyberspace.org
URL: http://www.securecyberspace.org

S. V. Raghavan
Department of Computer Science & Engineering,
Indian Institute of Technology Madras, Chennai, India
e-mail: svr@cs.iitm.ernet.in

1 Introduction

Digital evidence is present ubiquitously in cyber space today. Rapid advancements in digital technology over the past decade and the resultant multiplicity in file formats and log formats have rendered forensic analysis a challenge. Besides, applications also create multiple temporary files and logs hand-in-hand with regular files. Consider a scenario where a user downloads a set of digital photographs from the Internet, edits them and markets them as originals. This could normally be construed as IP theft. There is no single tool in current technology that can enable an examiner to detect such activities [16, 26]. While tools to detect an edited image are available they fall short of detecting the activity sequence. However, it is necessary to bring into evidence the original photographs while making a case. In order to do so, it is essential to find the

original photographs from the user's computer and group them with the edited duplicate. Where possible, one can assume that the deleted files can be recovered using data carving technology [7] with complete or partial embedded metadata. The regular files along with the recovered deleted files can provide a complete set for analysis during an investigation. The research challenge in our work is to identify all related files to a suspected file stored on the user's computer and determine if there is evidence of image doctoring. Traditional techniques have used classification to organize files according to similarity in content. While classification may classify original photographs from doctored copies, during analysis it is necessary to relate the edited photograph with its original to demonstrate the fact. The metadata found in sources of digital evidence allows one to identify such associations without having to do exhaustive analysis using metadata based matches.

This paper proposes a method to automatically identify associations among the files in digital evidence at the syntactic and semantic levels using metadata. For example, if we consider the scenario described above, we identify associations between different sources based on metadata to elicit evidence; they could involve files that are likely to have been downloaded, the origin of the downloads and doctoring of digital photographs if any, by determining relationships across. We identify metadata associations using value matches between the digital images across different classes to form groups of associated image files called '*association groups*' [27]. *An association group is a set of files such that each file in that group has at least one metadata association with one other file in the same group*. The association groups are then analyzed with regard to the six questions [7] of *what, who, when, where, how and why* that are relevant during digital forensic analysis.

1.1 Motivation for finding associations in digital evidence

Often during investigations, it may be necessary to holistically consider all related files rather than in isolation. This requires that for each file (file), the related files and log entries are identified and grouped together for analysis. This approach can be useful during analysis of files for the: (*i*) purposes of establishing the provenance, or (*ii*) purposes of identifying a pattern in the creation, modification, access or deletion of files and the nature of the files themselves, or (*iii*) purposes of analyzing related groups that can aid in the identification of said group's relevance to further investigation. Such a task can be achieved in two different ways: (*i*) using the actual *content* in the files and identifying matches across files; or (*ii*) using the attributes describing the files, or *metadata,* and identifying matches or similarities in them. The former is computationally intensive and is often used in literature whenever deep file analysis is needed [13, 17, 18, 32]. On the contrary, the latter remains largely unexplored. We focus on the metadata to identify metadata based associations across files.

1.2 Classification versus association

When analyzing a collection of files, typically, analysis can involve classification—Image classification, for instance, that involves grouping of similar digital image files, can be performed in many ways; image source-based, image dimension-based, digital camera-based, image timestamp-based, and so on. Hitherto, analysis has been focused on artifacts belonging to the same source [3, 10, 13, 23, 32] and the techniques employed apply classification using both a single parameter and multiple parameters. However, such classification is predominantly syntactic and often the burden of determining related digital images among a single category falls on the individual. When confronted with heterogeneous artifacts or even files belonging to the same category but belonging to different technology generations, classification requires an additional step by the examiner to "link up" the files. In practice, one may have to classify the files repeatedly, often using different parameters, before a pattern emerges. However, the intelligence relating to different types of classification which are likely to reveal such insight is not readily available [14].

Association based analysis focusses on identifying those digital images which are likely to occur across such groups and not bound by, albeit not precluded to, the rules defined by traditional classification. This is illustrated in Fig. 1. On the same collection of digital image files, while a classifier may take a classification parameter as seed input, the metadata based associations approach does not need any input. Besides, a classifier may give rise to image classes containing similar image files that are homogeneous with regard to that classification parameter, and the metadata association generates groups that contain image files "related" based on their metadata, which in turn relate to the events that affected them.

This paper proposes a method to *automatically identify syntactic and sematic associations* in collections of digital image files and word processing documents to elicit inter-file relationships for the purpose of identifying doctored images and derived copies of documents *using metadata*. We do this by identifying metadata associations using value matches between the files to form groups of associations called association groups [27]. We define six types of associations among files and categorize the metadata in word processing documents and digital image files into metadata families conducive to forensic analysis. We

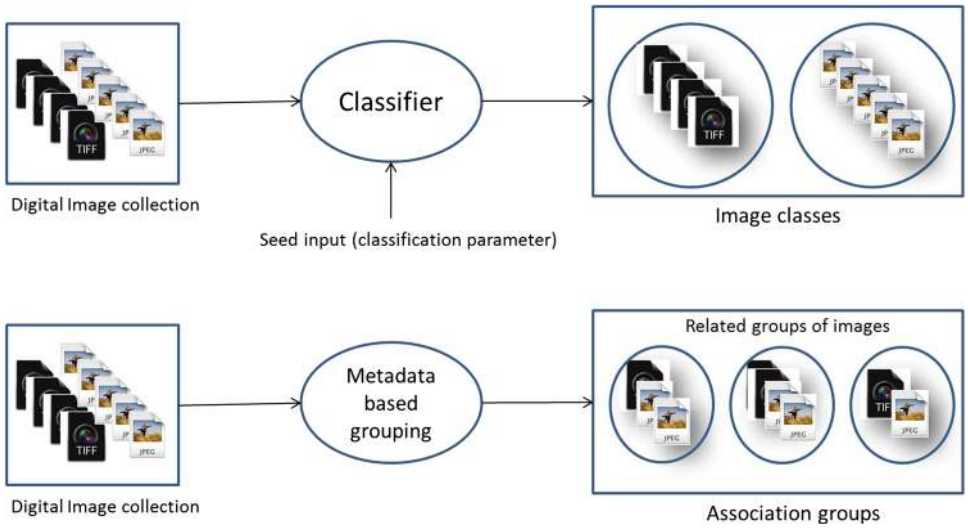

Fig. 1 Illustrating the differences in image classification versus association

propose two algorithms in this work that make use of the file relationships to group files that share source-based association and identify doctored files during analysis. In the sequel, we illustrate the role of metadata in digital investigations and describe the use of metadata to determine associations.

2 Related work

Metadata refers to *data about the data* that is stored in digital media. Metadata is the information about the data contained in a source, be it a file, folder, hard disk drive, logs or network traffic and *is independent of the content it describes*. For instance, metadata for a file contains information regarding the filename, location of the file, file size, content type, application type, ownership, access privileges, date and timestamps and so on. Metadata, by virtue of recording the partial state of a file, contain information of forensic value [6]. Metadata can be considered as sets of name-value pairs. As metadata describes attributes regarding the data, it is useful to group files with the same values for attributes together. Similar description exists for log file related metadata.

2.1 Types of metadata

Metadata contain information relating to *who*, *how* and *when* the files were created or modified or accessed [5]. In files, information relating to filename, location, file extension, size, MAC timestamps,[1] author (group), and word count, etc. are recorded as file metadata. Some metadata may also provide additional attributes such as content length, total edit time, line count, last saved and printed timestamp, author group, last author, creator, publisher, etc. and the granularity is often dependent on the application. Two important types of file metadata are *file system metadata* or metadata generated by the file system regarding that file and *application metadata* or metadata generated by specific applications about the content stored on such files.

File system metadata record information that relate to the file system and help it manage the file within that file system. Buchholz and Spafford [5] provide a qualitative treatment of file system metadata and their importance in digital forensics which reemphasizes the ability to answer the Casey's six questions [7].

Application metadata is a blanket name given to information that applications store regarding the files they operate on. Application metadata are strongly reliant on the type of file they describe, i.e., application metadata for a text file differs significantly from that of a Microsoft document or a JPEG image file. Brand et al. [4] categorized application metadata into three categories, viz., *descriptive*, *structural* and *administrative* metadata, each referring to specific domains within an application. NISO presented an overview of the different metadata structures and Microsoft Office documents have imbibed this specification into their documents which resulted in the OOXML metadata.

2.2 Use of file metadata in digital forensics

Alvarez [1] used EXIF metadata in digital photographs to verify authenticity of a picture and determine whether it was altered. Castiglione et al. [8] highlighted the information that can be obtained from the Microsoft Compound Document File Format (MCDFF) and lists some metadata useful in forensic investigations. Rowe and Garfinkel [29]

[1] MAC timestamps indicate when a file was created (C), when it was last modified (M) and when it was last accessed (A).

developed a tool that used directory and file metadata to determine anomalous files on a large corpus. The tool used *fiwalk* to traverse the corpus and compute statistical characteristics on metadata containing numerical values. The analysis resulted in the identification of misnamed and duplicate copies of files. Chow et al. [9] evaluate file system MAC timestamp rules and Koen and Olivier [20] applied them to validate files for copy or move actions. Willassen [31] designed a method to compare file system MAC timestamps to detect antedating.

2.3 Metadata for grouping files

Boutell and Luo [3] used EXIF metadata in digital photographs to classify camera types. Minack et al. [24] evaluated image-related metadata based search on personal image collections. Liu et al. [22] proposed a feature combination method to classify digital images that combined image content and EXIF metadata based on linear-discriminant-analysis (LDA) for digital photograph management.

Bohm and Rakow [2] discussed the different aspects of classifying multimedia documents based on document metadata. Multimedia documents can be classified into six orthogonal categories, viz., representation of media type, content description, content classification, document composition, document history and document location. Fathi et al. [13] and Denecke et al. [10] classified documents based on author and title in document metadata and Toyama et al. [30] built a system that utilized geographic information in location metadata (or geotags) to classify digital photographs with same location information. Maly et al. [23] proposed a method to classify documents based on layout metadata.

The major challenges associated with file analysis can be summarized as the following:

1. Device used to create one or more file
2. Software used in creating or processing files
3. Users or owners of one or more files
4. Time instants when the files were operated on

When a file or a collection of files is analyzed, some of the questions that need to be answered in reference to the six forensic questions listed in [7] may require: (*i*) the identification of one or more devices involved in the creation/transport of files, (*ii*) software or list of software used in creation and or editing/doctoring (as the case may be) of the files, (*iii*) the owner and author or list of owners and authors who are associated with one of more of the files, and (*iv*) the time instants when one of more of the files were operated on in any way during its life cycle. Often the analyses appear fragmented when such information is discerned from the files in isolation. However, using metadata based associations, we believe that files group together to elicit the underlying context. Often patterns emerge which can be valuable during analysis.

Traditionally, such analysis was conducted by classifying the files in questions and conducting searches based on known information. Another topic of relevance in the context of forensic analysis is file authentication that usually involves extensive computation [18, 19]. However, the sheer volumes of digital evidence analyzed today render the process of discovery through query and search or even classification infeasible [14, 16]. File classification focuses on syntactic organization; the goal of digital forensics however, is the identification of all event sequences and determining the files relevant to those sequences. For instance, it may be necessary to generate a list of all files that were created at a particular location determined based on the EXIF *lat-long* information of a digital photograph or determine doctored photographs and group them with their originals or identify the different versions of a document that exist and determine the original (oldest) document using a timeline. This requires an approach which can identify not just identical files [29] but also other forms of file associations. We demonstrate this approach on *unknown* collections of digital image files and word processing documents. We discuss the types of associations that this paper is concerned with in the sequel.

3 Types of metadata associations

Metadata associations can arise out of different types of matches in the metadata value and with regard to that, there can be 4 basic types of associations based on value, viz., *exact association*, *partial association*, *threshold association* and *date association*. These are elaborated below:

3.1 Exact association

When a particular metadata value in one file matches exactly with the corresponding metadata on another file, irrespective of the type of value, an *exact association* is said to occur between the files for that metadata.

3.2 Partial association

When a particular metadata value in one file matches partially with the corresponding metadata on another file, for a value of STRING type, a *partial association* is said to occur between the files for that metadata. Such a partial association can be of three different types.

Left sequence For two strings s_1 and s_2 such that $s_1 \neq s_2$, if two or more characters from the left in s_1 match

exactly with the corresponding characters in s_2, that defines a *left sequence partial association* between s_1 and s_2.

E.g. s_1 = **SAM**UEL s_2 = **SAM**SON

Right sequence For two strings s_1 and s_2 such that $s_1 \neq s_2$, if two or more characters from the right in s_1 match exactly with the corresponding characters in s_2, that defines a *right sequence partial association* between s_1 and s_2.

E.g. s_1 = WILLIAM**SON** s_2 = ROBERT**SON**

Anywhere in the middle For two strings s_1 and s_2 such that $s_1 \neq s_2$, if two or more characters in s_1 match exactly with the corresponding characters in s_2 and do not match at either the left or right ends, that defines a *middle sequence partial association* between s_1 and s_2.

E.g. s_1 = I**NTRI**GUE s_2 = CO**NTRI**EVE

3.3 Threshold association

When a particular metadata value in one file differs with the corresponding metadata on another file, for a value of NUMERIC type, such that the difference occurs within a pre-defined threshold, a *threshold association* is said to occur between the files for that metadata. Such a threshold association may occur either with a value greater than or less than the specified threshold. As such, the nature of the difference in value is only relevant, if the file on which the comparison is pivoted, is identified.

3.4 Date association

When a particular metadata value in one file, for a value of DATE type, is matched against with the corresponding metadata on another file, it defines a *date association* between the said files for that metadata. Such a date association can occur in four different types.

At time t For two timestamps t_1 an t_2, if their values match to the last degree of resolution that can be determined within technological constraints, then an *at t date association* is said to occur. The value is taken as reference time *t*.

Before time t For two timestamps t_1 and t_2 such that $t_1 \neq t_2$, when it is determined that one timestamp is less than the other, then a *before t date association* is said to occur. In this case, the file corresponding to the larger timestamp value is taken as reference on which the comparison is pivoted and its value is taken as reference time *t*.

After time t For two timestamps t_1 and t_2 such that $t_1 \neq t_2$, when it is determined one timestamp is greater than the other, then an *after t date association* is said to occur. In this case, the file corresponding to the smaller timestamp value is taken as reference on which the comparison is pivoted and its value is taken as reference time *t*.

Between time instants t' and t" For two timestamps t_1 and t_2, if we can determine pre-defined time instants *t*' and *t*" such that $t' < t_1$, $t_2 < t$", then a *between t' and t" date association* is said to occur.

4 Metadata based analysis of file collections

The aim in the analysis of digital evidence is identification of the events leading to a reported incident, the nature of these events and their attribution to individual(s). For our discourse, an event refers to actions that are directly performed by an individual on any digital device. Examples of such events are creating a file, modifying a file, sending an email, logging into a server, visiting a website, downloading a file, etc. Each event can result in creating new files, or accessing or modifying existing file(s). If a new file is created as a result of an event, its occurrence is reflected in the metadata that are also created along with the file. If an existing file is modified as a result of an event, its occurrence is reflected in the change in values of the metadata linked to that file. Therefore, irrespective of the type of event, its effect can be perceived in the metadata linked to the metadata. When an event creates or modifies more than one file, identifying the metadata that pertain to the event across these files will elicit the relationships that exist between them. Therefore, focusing on the appropriate metadata across the files, one can reconstruct the event(s).

4.1 Conducting forensic analyses on file collections

During a digital investigation that involves the analysis of collection(s) of digital images, many forensic questions can be raised, some of which are listed below:

1. How many sources can be identified from the file metadata? How many files belong to each of these identified sources?
2. How many files show evidence of being doctored? How many Internet downloaded files show evidence of doctoring? What editing software was used in each case?
3. Are there other "similar" files where source metadata is incomplete? How many other metadata match for such files?
4. Which of the files were downloaded from the Internet? If so, can the source of these files be identified?

While some of these questions can be answered in part or whole using traditional classification, often it is up to an examiner to analyze the individual classes to identify inter-

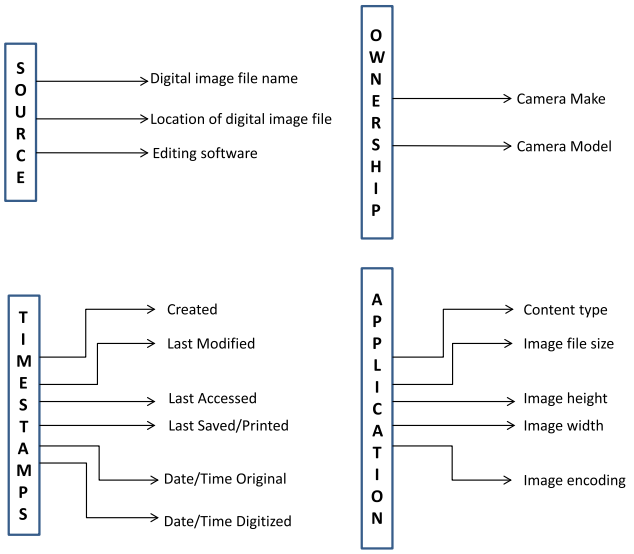

Fig. 2 Digital image metadata tags of interest in digital investigations

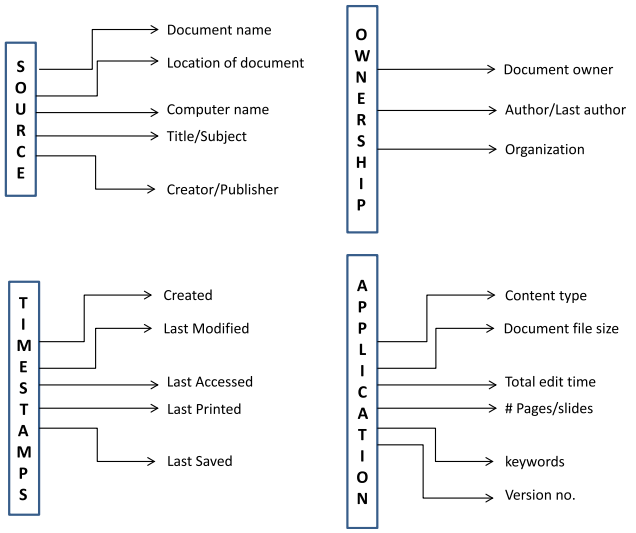

Fig. 3 Word processing document metadata tags of interest in digital investigations

file relationships. We believe that identifying such relationships can help a forensic examiner infer the nature of activities that led to the existence of the files being analyzed. To determine answers to such questions, it is necessary to recognize that no single classification method can provide all the answers and it is necessary to determine relationships between the files to extract all higher-order associations that exist both within a particular source class and across such classes. Such a task requires exhaustive classification using all individual parameters (from metadata) as well as all combinations of multiple parameters to determine where the files overlap and group them. The association groups generated from the metadata based associations, on the other hand, achieve this task readily and simplify the task of identifying related files to a search task within an association group. Through its automation, the metadata based associations [27] integrates this task and eliminates the need to manually identify such related files during analysis.

4.2 Metadata and metadata families for file collections

We identify the digital image metadata at their respective metadata families relevant during forensic analysis in Fig. 2. A collection of digital image files can be organized according to the image file names and their respective locations on a particular source of digital evidence. The metadata that allow one to do that belong to the *source* metadata family. Another metadata pertaining to this family, viz., 'software' metadata is usually found in digital images if the images were edited. When this metadata value is present and there are no discernible EXIF markers, it could indicate a digitally generated image file.

Digital image files also require to be identified based on the device used to record or capture the digital image files [11] and the metadata that allow us to do that are the EXIF metadata Camera make and model metadata tags. The EXIF metadata [12] in the digital image files store information about the digital still camera and technical details about how a digital photograph was captured. Such groupings not only identify all the cameras used in generating the collection, but they can be used to identify the number of digital images generated by camera of a particular make and model. These metadata belong to the *ownership* metadata family.

The MAC timestamps and the EXIF timestamps, where available, belong to the *timestamp* metadata family and identify events corresponding to creation, modification and access of the image files.

Image dimensions can help one gauge the granularity of digital image files and is a useful pre-analysis metric; higher the image dimensions, better the level of detail in the image file. Such metadata and those such as image file size and image content type that provide information regarding the features of digital image files belong to the *application* metadata family.

Digital image files do not store author information; rather record the details pertaining to devices such as digital still cameras, computers and computer-based software used in creating or editing these images. As a result, the software and camera devices are identified as source and ownership information pertaining to namesake metadata families in our experiments.

We identify the document metadata at their respective metadata families relevant during forensic analysis in Fig. 3. A collection of word processing documents can be organized

according to the image file names and their respective locations on a particular source of digital evidence. As discussed earlier, title or subject metadata can often throw light on understanding if the document has been used as a template in creating the material while leaving the metadata untouched. 'Creator' and 'Publisher' metadata help identify some of the additional software used in generating the content. Such metadata belong to the *source* metadata family.

In documents, it may be necessary to identify the author(s), their affiliations with an organization or company, when and who last modified the document and so on. The metadata that allow one to do that belong to the *ownership* metadata family. The MAC timestamps and the document timestamps, where available, belong to the *timestamp* metadata family and identify events corresponding to creation, modification and access of the word processing documents. Metadata such as number of pages, slides etc., retain some content context. 'keywords' is another metadata, if available, which could potentially provide alternate keywords to examiners while exploring related documents or other files from one or more sources of digital evidence. Such metadata that provide information regarding the features of word processing documents belong to the *application* metadata family.

4.3 File relationships based on metadata associations

When we determine metadata associations across files, it underlines the relationship between the files which can reveal the nature of activities recorded. In this section, we define six types of file relationships based on metadata associations to conduct analysis.

4.3.1 Existence relationship

When a metadata match occurs in the source metadata family for metadata *filename* or *Title/Subject* of the file between files f_1 and f_2, where f_1 and f_2 reside on different homogeneous sources, we define an *existence relationship* between the files. The files themselves need not belong to the same application type, but only contain the metadata that leads to a metadata association, e.g., .DOC and .DOC or .DOC and .BAK or .TMP. The relationship is denoted by R_e and it may be expressed as $f_1 \mathrm{R}_e f_2$ and read as $f_1 \Longleftrightarrow f_2$. By definition this relationship is commutative and associative. The association groups containing such relationship pairs in evidence are referred to as existence association groups. Therefore,

1. $f_1 \mathrm{R}_e f_2 \Longleftrightarrow f_2 \mathrm{R}_e f_1$
2. $(f_1 \mathrm{R}_e f_2) \wedge (f_2 \mathrm{R}_e f_3) \Longleftrightarrow (f_1 \mathrm{R}_e f_3)$

When multiple such files (f_1, f_2, f_3, … f_n) exhibit an identical association between each other, e.g., produce a metadata match for the same value of filename, we represent this relationship as R_e ($f_1, f_2, f_3, \ldots f_\mathrm{n}$).

4.3.2 Source relationship

When a metadata match occurs in the source metadata family between files f_1 and f_2, where f_1 and f_2 belong to the user file system, we define a *source relationship* between the files indicating that the files were likely to be created on the same source as identified the respective metadata. The relationship is denoted as R_s and is expressed as $f_1 \mathrm{R}_s f_2$. By definition this relationship is commutative and associative. Therefore,

1. $f_1 \mathrm{R}_s f_2 \Longleftrightarrow f_2 \mathrm{R}_s f_1$
2. $(f_1 \mathrm{R}_s f_2) \wedge (f_2 \mathrm{R}_s f_3) \Rightarrow (f_1 \mathrm{R}_s f_3)$

When multiple such files (f_1, f_2, f_3, … f_n) exhibit an identical association between each other, e.g., produce a metadata match for the same value of computer name or software, we represent this relationship as R_s ($f_1, f_2, f_3, \ldots f_\mathrm{n}$).

4.3.3 Parallel occurrence relationship

When a metadata match occurs in the timestamp metadata family between two files f_1 and f_2, where f_1 and f_2 belong to the user file system, we define a *parallel occurrence relationship* indicating that the two files f_1 and f_2 were accessed at the same time instant identified by the matching value of the timestamps in their metadata. The relationship is denoted by R_{po} and expressed as $f_1 \mathrm{R}_{po} f_2$. By definition, this relationship is commutative and associative. Therefore,

1. $f_1 \mathrm{R}_{po} f_2 \Longleftrightarrow f_2 \mathrm{R}_{po} f_1$
2. $(f_1 \mathrm{R}_{po} f_2) \wedge (f_2 \mathrm{R}_{po} f_3) \Rightarrow (f_1 \mathrm{R}_{po} f_3)$

When multiple such files (f_1, f_2, f_3, … f_n) exhibit an identical association between each other, e.g., produce a metadata match for at least one timestamp on the same value, we represent this relationship as R_{po} ($f_1, f_2, f_3, \ldots f_\mathrm{n}$).

4.3.4 Structure similarity relationship

When a metadata match occurs in the application metadata family between two files f_1 and f_2, where f_1 and f_2 belong to the user file system, we define a *structure similarity relationship* indicating that the two files f_1 and f_2 have identical or equivalent attributes. The relationship is denoted by R_{ss} and expressed as $f_1 \mathrm{R}_{ss} f_2$. By definition, this relationship is commutative and associative. Therefore,

1. $f_1 \mathrm{R}_{ss} f_2 \Longleftrightarrow f_2 \mathrm{R}_{ss} f_1$
2. $(f_1 \mathrm{R}_{ss} f_2) \wedge (f_2 \mathrm{R}_{ss} f_3) \Rightarrow (f_1 \mathrm{R}_{ss} f_3)$

When multiple such files (f_1, f_2, f_3, … f_n) exhibit an identical association between each other, e.g., produce a metadata match for the same value of content type or file size, we represent this relationship as R_{ss} ($f_1, f_2, f_3, \ldots f_\mathrm{n}$).

4.3.5 Unauthenticated modification relationship

When two files f_1 and f_2 differ in metadata only with respect to the structural composition of the files and the software exclusively present in only one of the files, it indicates an *unauthenticated modification relationship* denoted by R_{ua} and expressed as $f_1 R_{ua} f_2$. The relationship, by definition is commutative.

4.3.6 Majority relationship

When two files f_1 and f_2 have an unauthenticated modification relationship, in the presence of a third file f_3 which contains a source relationship with either f_1 or f_2, then that pair of files is said to exert a *majority relationship*, denoted by R_m over the other file. Therefore, if

$(f_1 R_{ua} f_2) \wedge (f_1 R_s f_3) \Rightarrow (f_1, f_3) R_m f_2$.

When these relationships are determined across files on the same homogeneous source, it results in similarity pockets if exactly one metadata match is discovered or similarity groups in the case of multiple metadata matches. Across multiple sources, as in the case of existence relationship, this would result in association groups.

In the sequel, we introduce our metric to evaluate the effectiveness of grouping based on metadata associations.

4.4 Metrics and measurements

To evaluate the effectiveness of the metadata associations on a given dataset, we introduced a parameter called the *association index* (*ai*). The association index *ai* for a file on a particular source or dataset is defined as the fraction of the number of files on that source that can be discovered using the metadata associations generated with a given file and applied iteratively each discovered file exhaustively. By definition, a value $ai = 0$ indicates that the file in question is isolated and $ai = 1.0$ indicates that the file is highly connected and all files are related to the said file. The following relationships hold with regard to the association index *ai*:

$$0 \leq ai \leq 1.0 \tag{1}$$

$$ai = \frac{1}{\sum i}\left(\sum_i \frac{|agi|}{N}\right) \tag{2}$$

where $\sum_i |agi|$ is the number of files in the association groups as determined using file i as the seed, N represents the total number of files being considered and $\sum i$ is the total number of groups formed as a result of the association based grouping. In our experiments, on a given source, we compute the association indices for all the files on the source and determine the mean *ai* value that is assigned to the source.

To assess the effectiveness of the metadata associations generated on file collections, we define *effort margin* $\mathbf{r}$ and its complement *grouping efficiency* $\boldsymbol{\eta}$ as metrics. The effort margin is a measure of the fraction of effort as against the individual file analysis when conducting forensic analysis. The effort margin is computed as the ratio of sum of the number of association groups to the number of groups to be analyzed in the worst case.[2] The value ranges from 0 to 1, where 0 represents zero effort for the examiner and 1 represents effort identical to that when carried out with traditional forensic tools for individual file analysis. The effort margin can take a value 1, if and only if all the files remain unassociated after applying the model leading to a separate group for each file. The effort margin can take a value 0 only theoretically since the least value for the numerator in the ratio is 1 which results when all the files are grouped together.

The grouping efficiency is a measure of the degree of closeness between the files in digital evidence, across all sources. It is computed as $1.0 - \mathbf{r}$. The value for grouping efficiency ranges from 0 to 1, where 0 represents that no association groups were generated, implying that all the files remained unassociated, while a value of 1 represents that all the files were grouped together. The grouping efficiency can take a value 1 only theoretically since the effort margin can only take non-zero values in practical scenarios. In short,

$$\textit{Effort margin } \mathbf{r} = \frac{\text{Number of association groups}}{\text{Number of association groups in the worst case}} \tag{3}$$

$$\textit{Grouping efficiency } \boldsymbol{\eta} = 1 - \mathbf{r} \tag{4}$$

In the sequel, we propose algorithms based on identifying metadata associations for discovering files source and identifying doctored files during analysis.

4.5 Algorithms

The relationships can exist based on an exact value match between two or more files of the same type on the same homogeneous source or across heterogeneous files based on a value match established through a metadata equivalence relationship on the corresponding metadata names across sources. In our algorithms (described below), the free-running variable t accounts for the different disjoint groups generated within a source and t is a member of the set of natural numbers $\mathbf{N}$. In order to identify the files belonging to the same source, we apply the *source relationship* as described in Algorithm 1.

[2] In the worst case, number of association groups equals the number of files in the source.

Source Identification Algorithm

Given: $S = \{s_1, s_2, s_3, \ldots s_N\}$, the set of all sources of digital evidence
$F_i = \{f^i_1, f^i_2, f^i_3, \ldots f^i_{Ni}\}$, the set of all files belonging to source s_i, $i \in [1, N]$
$M_i = \{m^i_1, m^i_2, m^i_3, \ldots, m^i_{Ni}\}$, the set of all metadata vectors corresponding to each $f^i_j \in F_i$, $j \in [1, N_i]$
Set L of metadata corresponding to the *source metadata family* on each source s_i
Output: A list of sources stored in *orgn* and the corresponding sets of files SP^i
begin algorithm
$orgn \leftarrow 0$; $SP^i \leftarrow \emptyset$
for each $s_i \in S$ **do**
repeat
for each $f^i_j \in F_i$ **do**
for each $m^{ij}_k \in m^i_j$ and $m^{ij}_k \in L$ **do**
$sp^{ik}_t \leftarrow \{f^i_j \mid j \in [1, N_i], (\exists v, m^{ij}_k = v)\}$
end for
end for
$SP^i \leftarrow \{sp^{ik}_t \mid k \in [1, M], i \in [1, N], t \in \mathbf{N}\}$
$orgn \leftarrow$ list v of values corresponding to each sp^{ik}_t in SP^i
until $|orgn| = |SP^i|$
end for
Generate a list *orgn* of individuals or devices from all $sp^{ik}_t \in SP^i$ where $j \in [1, N_i]$, $k \in [1, M]$
Display *orgn*, SP^i as outputs
end algorithm

For this algorithm, the list L maintains a list of those metadata that record values corresponding to source devices or software that were used to generate the file it was attributed to. The source device or software can be different from the source of digital evidence that contains a digital image file. Where necessary, the metadata equivalence relationships are established across digital image files that contain the same metadata value for differing metadata tag names.

Having grouped files that demonstrate the same source associations, it may be necessary to also determine some files from that set which are modified. Typically, this can imply that files belonging to some source were doctored using same software. However when two image files demonstrate the software edited relationship, it may need to be established, with the presence of a third file, that in conjunction with the first file exerts a majority relationship. This is because, with regard to digital image files where this relationship holds forensic value, sometimes the absence of metadata can imply software activity, as in the case of digitally generated image files and image files downloaded from the Internet [28]. In order to identify all files that were edited with a particular piece of software, we apply the *unauthenticated modified relationship* and for each pair, identify a third file, two of which can exert a *majority relationship* over the third for the 'Software' in the source metadata family as per Algorithm 2.

Edits Identification Algorithm

Given: $S = \{s_1, s_2, s_3, \ldots s_N\}$, the set of all discrete homogeneous sources of digital evidence
$F_i = \{f^i_1, f^i_2, f^i_3, \ldots f^i_{Ni}\}$, the set of all files belonging to source s_i, $i \in [1, N]$
$M_i = \{m^i_1, m^i_2, m^i_3, \ldots, m^i_{Ni}\}$, the set of all metadata vectors corresponding to each $f^i_j \in F_i$, $j \in [1, N_i]$
Output: A list *sftw* of software and corresponding sets of files SP^i
begin algorithm
for each $s_i \in S$ **do**
repeat
for each $f^i_j \in F_i$ **do**
for each $m^{ij}_k \in m^i_j$ corresponding to the j^{th} file $f^i_j \in F_i$ **do**
$sp^{ik}_t \leftarrow \{f^i_j \mid j \in [1, N_i], (\exists v, m^{ij}_k = v)\}$
end for
end for
$SP^i \leftarrow \{sp^{ik}_t \mid k \in [1, M], t \in \mathbf{N}\}$
Extract unauthenticated modification relationship $\{(f^i_j, f^i_k) \mid f^i_j \, R_{ua} \, a^i_k, j \in SP^i, k \in SP^i\}$ for each file f^i_j from sp^{ik}_t in SP^i
for each (f^i_j, f^i_k) pair identified **do**
Identify a third file f^i_n such that $f^i_n \, R_m f^i_j$ for sp^{ik}_t in SP^i
end for
stfw $\leftarrow$ source metadata name corresponding to the software that established the modified relationship R_m on triad f^i_j, f^i_k, f^i_n from sp^{ik}_t in SP^i
until $|sftw| = |SP^i|$
Display *sftw*, SP^i for s_i as outputs
end for
end algorithm

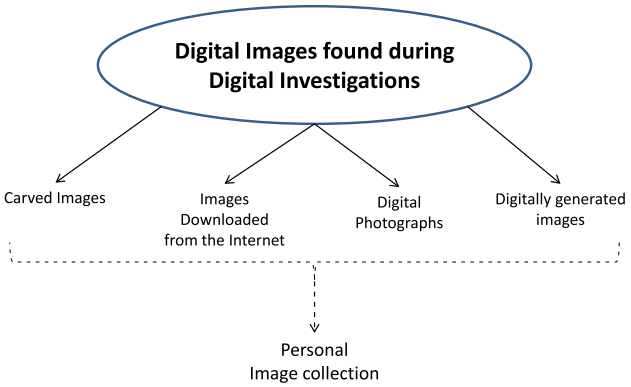

Fig. 4 Different probable sources for digital images discovered in digital evidence

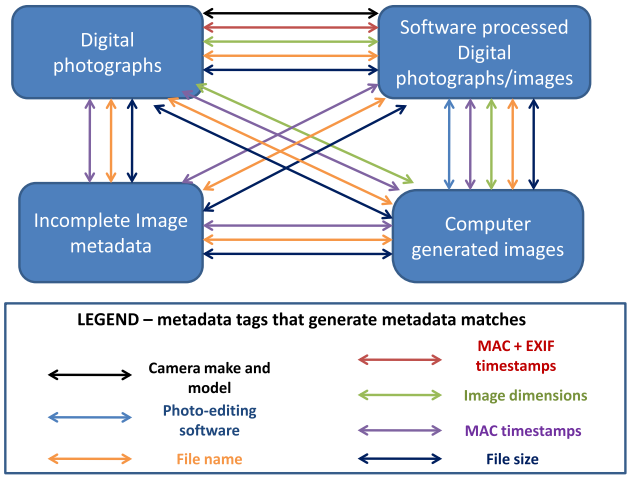

Fig. 5 Illustrating possible metadata associations between the different lists

In the sequel, we discuss the nature of metadata associations that can be identified on digital image and word processing document datasets and apply our algorithm to identify the sources using metadata association. We also apply our software edits algorithm on the result to identify doctored files and present our findings.

5 Analysis of digital images using metadata associations

While examining a source of digital evidence for digital images, an examiner is likely to discover images from different sources, viz., images recovered from carved data [25], images that are digital photographs, images edited or digitally generated using software and images downloaded from the Internet. These different types of digital images are shown in Fig. 4.

Each collection of images has a different level of metadata associated with it that can either enhance or impede the grouping. Usually, images from carved data have incomplete or no metadata and hence a grouping based on metadata is likely to result in a high effort margin and low grouping efficiency. Images from the Internet can be downloaded in several ways and popular methods include downloading images from Google image search results and downloading compressed archives from where the images are then extracted. While the Google database may not include image metadata unless it is voluntarily provided during uploading, archives usually omit image metadata during compression. As a result, the chances that metadata is present in such images is likely to be low, which could also lead to a high effort margin **r** and low grouping efficiency **η**. Images that are digital photographs store a variety of metadata provided by digital technology for better management. As these images are rich in metadata, they are likely to result in low **r** and high **η**. Digitally generated images and those edited by software are increasingly storing valuable information in the image metadata and hence fall in the same category for **r** and **η**. In any personal collection, the images found are usually a mixture of digital images across such sources, and hence the grouping efficiency is determined by the majority fraction of image sources.

We develop a systematic method to group digital image files in a given collection to identify doctored and digitally generated images using metadata based associations and identifying image file relationships. Doctored image files are copies of digital photographs which are processed using image editing software such that its relationship with the original photograph is not apparent. As a consequence, identifying such image files as doctored copies and relating it back to the original photographs by grouping them together remains a challenge.

5.1 Expected behavior

Edited photographs generate many metadata associations since the images typically contain metadata pertaining to camera make and model and the photo-editing software. The MAC timestamps and EXIF timestamps can be used to generate a unified timeline of the digital images and validate the authenticity of photographs suspected to be edited when sufficient metadata is unavailable. The set of possible associations that can be identified among the various lists is illustrated in Fig. 5.

5.2 Observations

Digital photographs captured with the same camera generated many metadata associations all of which corresponded to source relationships. Edited photographs gave rise to the *unauthenticated* R_{ua} relationship which was later

Table 1 Results of applying the source identification algorithm on digital image file datasets

Dataset name	Dataset volume	Number of images in the dataset	Digital photographs	Edited with software	Computer generated images	Incomplete image metadata
Noakes' photograph dataset	374 MB	126	124	7	0	2
Personal image collection	1.6 GB	491	312	53	12	179
Digital corpora	6.8 GB	2,157	207	1,891	0	1,891
Assorted dataset	50 GB	100,000	75,000	24,875	25	125

confirmed after establishing *existence* R_e and *majority* R_m relationships with the original photographs stored in the application temporary folders. The existence relationship was established between the digital photograph and the temporary file while photograph and the temporary file exerted the majority relationship over the edited image from the collection. The results of applying Algorithm 1 to our digital image datasets are shown in Table 1.

Noakes' dataset[3] contained 124 digital photographs[4] of which 7 were processed with Adobe Photoshop and 2 images had insufficient image metadata. In the personal image collection,[5] we discovered overlapping sets with regard to the digital photographs and the software processed images. In the Digital corpora dataset,[6] we discovered 1891 images as belonging to Incomplete Image metadata and all digital photographs in this dataset were processed using Adobe Photoshop. In the Assorted dataset,[7] there were 25,000 edited and digitally generated image files that were created by processing the original photographs using photo-editing software. The results of applying Algorithm 2 the digital image datasets for identifying photographs that were digitally doctored or generated is tabulated in Table 2.

The digital image collections that contained digital photographs typically contained multiple photographs from the same digital camera. All digital photographs from the same camera generate source relationships between each other and consequently are grouped together in the same association group. Naturally, each digital image in that group finds all other digital images from the same group. Therefore, if one of the digital images in an association group had an *ai* value 0.3, all the other digital images in that group also had the same value. In general, we may state that each digital image had an *ai* value which is the fraction of the total number of digital images the dataset that were associated with that image. Therefore, datasets that contained digital photographs (both normal and edited) produced higher value for *ai* as against datasets that contained fewer digital photographs. In dataset #4, we observed high averages for the *ai* values since there were only few unassociated digital images. In datasets #2 and #3, we observed very low values for average *ai* since there were a significant number of unassociated digital images in these datasets.

In dataset #2, the software processed images also overlapped with the set of images that contained incomplete image metadata, primarily on the Software metadata tag. There were 179 images identified under the category of incomplete image metadata, however, 53 of those contained the software metadata tag. Some images in this collection were intended to be used as desktop background images. The image dimensions 800 × 600 and 1,080 × 800 were commonly found in computer generated images which are also the standard desktop resolution ratios on any computer monitor. All digital photographs that were processed using software, for instance the images under both categories in datasets 2, 3 and 4, were found to have similar image dimensions metadata.

In conclusion, we demonstrated the use of the metadata based associations to detect doctored and digitally generated image files in a collection. Our observations seem to indicate that detection rate is independent of the size of the dataset but depends on the quality of the metadata and the cohesiveness of the image files within the dataset, as measured by the association index *ai*.

6 Analysis of documents using metadata associations

We develop a systematic method to group word processing documents in a given collection to identify derived documents using metadata based associations and identifying

[3] http://code.google.com/p/metadata-extractor/source/browse?repo=sample-images.

[4] The dataset has been updated since and contains over 200 digital photographs.

[5] Obtained from a volunteer and includes carved image files, digital photographs, edited photographs, digitally generated images and Internet downloaded image files.

[6] Obtained from http://digitalcorpora.org/corpora/files.

[7] Created by combining image files from the digital corpora JPEG repo at http://digitalcorpora.org/archives/250 and the Dresden image database at http://forensics.inf.tu-dresden.de/ddimgdb/locations/jpeg_scene.

Table 2 Results of grouping the metadata associations to image datasets to detect doctored digital images

Dataset no.	Dataset volume	Number of images in the dataset	Association index (ai) $\frac{1}{\sum_i}\left(\sum_i \frac{\lvert agi\rvert}{N}\right)$	Number of edited or digitally generated files/number of detected files	Detection efficiency (%)
1	374 MB	126	0.42	7/6	85.7
2	1.6 GB	491	0.26	65/53	81.5
3	6.8 GB	2,157	0.001	1,891/207	10.9
4	50 GB	100,000	0.78	25,000/23,475	93.9

document relationships. Derived documents are those that were obtained from some original document, copied and stored either modified or unmodified such that its relationship with the original document is not apparent. As a consequence, identifying such documents as derived and relating it back to the original document by grouping them together remains a challenge. In addition to the identification of derived documents for detecting IP theft, it is useful to determine characteristics such as total number of authors, number of single author documents, number of authors who appear in exactly one file, most number of documents authored by a single individual and so on. Typically, classification techniques can identify these characteristics, each classification process uses unique parameters to determine the classes that exist [21]. However, during analysis, it is also necessary to identify documents related to those found in a particular class. For instance, if we were to classify all documents into Word documents, PowerPoint slides and Excel spreadsheets, how do we determine all the co-authors of a particular set of word document who have

1. authored single-author word documents?; and
2. co-authored PowerPoint slides or Excel spreadsheets?

If such authors exist, then are the set of co-authors identical or different? Some other questions that can be posed during analysis include how do we determine the PowerPoint slides that were created, modified, used or downloaded along with a word document and how many excel files were used during that time the document was edited? By their very nature, these questions necessitate one to study the relationships that exist in the documents, a task that requires content analysis, usually by an individual. Traditional forensic tools offer little help in identifying such critical information when analyzing document collections. We have identified 12 characteristics for document collections and propose the application of Algorithm 2 to the outputs generated from applying Algorithm 1 to our document datasets to determine the characteristics using the metadata. While we stipulate the traditional definition of the term sources for applying Algorithm 1, we modify the definition to mean the source groups obtained before applying Algorithm 2. The results are tabulated in Table 3.

6.1 Observations

For the desktop dataset, besides the metadata 'Author' and 'Organization/Company', the metadata 'Filesize' and 'Filename' generated the most number of metadata matches. After combing the overlapping similarity pockets, we discovered 108 association groups. In addition to this, there were 32 documents that were removed to the unclassified list as they lacked sufficient metadata. Such files were individually analyzed by examining the forensic image under FTK. Metadata association also leads to file grouping that reduces the number of independent documents for further analysis and it can help one triage a dataset and focus on a smaller set of relevant documents.

For the digital corpora dataset, there were not many common points with regard to where the documents were downloaded from and therefore, it resulted in a much larger set of association groups. Metadata 'Author' and 'Organization/Company' generated the most number of matches amongst their documents. Filesize matches, although present, had few other metadata matches and resulted in a small number of association groups. In all, we determined 1892 association groups and 209 documents in the unclassified list. Since these documents were downloaded from the Internet from diverse sources, the relative association factor was, expectedly, low. Notwithstanding, metadata matches and association groups enable one to group similar documents and analyze related documents together, eliminating the need to repeated or unnecessary analysis.

The number of distinct authors (characteristic #1) and number of distinct organizations (characteristic #7) are computed by counting the value field for the 'Author' and 'Company' metadata tags respectively. Since, author field is also multi-valued and a document can have more than one author when this is the case, each unique author is counted. Wherever multiple authors from the same organization are discovered, the individual similarity pockets are merged into association group(s). Thus, we integrate multiple association groups and the size of the largest similarity pocket for metadata tag 'Company' provides the organization generating the most number of documents. The largest multi pocket generated from the similarity

Table 3 Tabulating the results from determining dataset characteristics for the document datasets

Characteristic no.	Dataset characteristics	Desktop (976)	Digital corpora (2970)
	Association index (*ai*) $\frac{1}{\sum i}\left(\sum_i \frac{\lvert agi\rvert}{N}\right)$	0.21	0.004
	Number of derived documents	274	2,252
	Number of derived documents that were detected	170	202
	Detection efficiency (%)	62	8.9
1	No. of distinct authors	158	3,300
2	Most number of documents by one author	170	228
3	No. of authors who have authored more than one document	126	2,599
4	Most number of documents similarly named by a single author	36; Stefan	17; J. Scott Peterson[a]
5	Most number of documents of similar file size belonging to one author	98; Stefan	9; Jon Heal
6	Most number of organizations single author is affiliated with	4; Stefan	2[b]
7	No. of distinct organizations	71	1,098
8	Most number of documents generated within the same Organization	79; QUT	50; US Dept of Agriculture
9	Most number of authors from a single organization	13; QUT	11; US Dept of Agriculture
10	No. of organizations generating multiple documents	27	336
11	No. of distinct application names	16	20
12	No. of distinct document titles	207	1,703

[a] Since all the files in this repository we renamed (and named similarly) after they were downloaded by Garfinkel, this value is merely the single largest similarity pocket based on 'author'

[b] More than one author was found to be affiliated with two organizations. Since there is multiplicity, no name is specified

pockets for 'Author' and 'Company' will provide the values for characteristics #2 and #8. The number of non-singular similarity pockets identified for 'Author' and 'Company' will provide the values for characteristics #3 and #10.

By grouping the similarity pockets for 'Author' and 'Filename' similarity the size of the largest multi pocket provides the values for characteristic #4. If we substitute the similarity pockets generated by 'Filename' with those by 'Filesize' then, the largest multi pocket thus formed provides the values for characteristic #5. Superimposing the similarity pockets obtained from the 'Author' and 'Company' metadata will reveal the set of authors who share the same organization affiliation. The largest multi pocket formed by superimposing the similarity pockets for 'Author' with the ones for 'Company' provides the values for characteristic #9. Characteristics #11 and #12 are determined in much the same way as characteristics #1 and #2.

In conclusion, we demonstrated the use of the metadata based associations to detect derived word processing documents files in a collection. Our observations seem to indicate that detection rate is independent of the size of the dataset but depends on the quality of the metadata and the cohesiveness of the image files within the dataset, as measured by the association index *ai*. Additionally we were able to define and determine a set of 12 characteristics that help in triage of word processing document collections. These characteristics are obtained directly based on groupings obtained from metadata associations that can help an examiner focus on a relevant subset quickly during analysis.

7 Discussion

During forensic investigations, investigations often require information on the circumstances and conditions prevalent during periods of interest. The semantics associated with metadata usually relate to events (e.g., timestamps) and consequently, determining matching metadata values correspond to identifying identical or related events. In this paper, we have demonstrated the use of metadata based associations to automatically detect file relationships and group doctored image files and derived documents with the respective originals.

Metadata underlines the context to describe the *situational similarity* during the life cycle of the digital images stored in digital evidence. Using metadata associations, we can *automatically identify and group*:

1. a digital photograph and any altered version of itself together;
2. an edited image with digital generated images using a particular software;
3. digital image files with log records that identify the event sequence tracing the file download from the Internet;

4. a digital photograph or a digital generated image with image files that are related or similar containing partial metadata; and
5. all thumbnail image files.

The ability to automatically identify and group such related sets of digital image files based on metadata associations simplifies the process of analysis for an examiner. Metadata associations can be used to validate hypotheses by comparing different metadata values across the digital images from a known source and establish consistency among them. For instance, digital photographs taken with the same camera tend to have similar file names and possibly similar file sizes,[8] the metadata associations and the groups generated can be used to determine if such is the case. For photographs that do not adhere to this hypothesis, a detailed offline assessment can be conducted. This work assumes that the metadata used in identifying metadata associations are authentic and not subject to tampering. Where tampering or incomplete information is in play, it may be possible to detect those inconsistencies if extra intelligence is available a priori. This is a proposed direction as indicated in our future work.

7.1 Document relationships and analysis

Document metadata store a variety of information regarding who and how a document was created and operated on such as author, organization, document format, application type, application version, MAC timestamps and document timestamps. Such information are related to who created the document and how (formatting information) it was created. Document metadata may also record information about where it was created (geo-tagging), number of pages/slides, formatting type, encoding type and so on. Rowe and Garfinkel [29] have analyzed the same digital corpora *govdocs1* file repository [15] to determine anomalous documents. They compute statistical characteristics using directory metadata and identify the top and bottom 5 percentile in the repository as outliers. In our paper, we used different subsets of the same dataset and determined metadata associations to elicit file relationships for identifying doctored digital photographs on image collections and derived documents from document collections as instances of IP theft. Since metadata keep track of the events that influenced a file, identifying metadata associations among the files will help identify related events and files across files to assist an examiner in triage and quickly focusing on the relevant set of files for further examination.

[8] File sizes are similar only under the conditions all camera settings are identical for the photographs under question. In other cases, this cannot be guaranteed.

7.2 Digital image file relationships and analysis

When it is suspected that one or more digital image files were downloaded, this can be established by identifying the download R_d and the happens R_h relationship between the image files and the respective browser log files [28]. Digital image files that demonstrate an existence R_e relationship indicate the presence of another copy of that image and this can be used to determine duplicate image files in a collection. Besides, when such pairs of image files also exhibit source R_s relationship along with unmodified authentication R_{ua} relationship, an edited image file is likely to be present whose original image is identified using the existence R_e relationship. Naturally, during image analysis, these image files can be starting points when no other information is available regarding the image collection. Each camera make and model identified through source R_s relationship is a potential source of digital evidence discovered. Digital image files that demonstrate parallel occurrence R_{po} relationship are likely to have been operated on using some software if there is an exact metadata match and further analysis of the content may be warranted in cases where R_{ua} relationship is not observed. Digital image files which exhibit the structural similarity R_{ss} relationship are likely to possess identical image resolution capability and encoding indicating that their content can be analyzed using the same tool. This can be useful if an unknown application format is detected during the examination of the image collection. Images with incomplete image metadata, unless they contained illicit content, can be spared from unnecessary analysis. However, that may be ascertained only through content processing using an alternate tool.

7.3 False positives

One of the challenges that this approach can face, particularly in the absence of any information that can help an examiner target one's analysis, is when seemingly unconnected files generate metadata associations leading to crowded association groups. This is more likely in environments where multiple shared computers can contain shared metadata values. In our experiments which included files obtained from multiple computers on a single network, generated multiple partial matches with users from one organization or from varied versions of the same parent software and in such cases, the partial matches on the source metadata family were excluded from further search. Where a seed file was used to determine the association group, the false positive rate was under 10 % on the larger datasets. Mostly, the false positives resulted in additional processing of up to 20 extra files within an association group.

An approach to generate focus in such scenarios is to rank the metadata families prior to applying the algorithms. Another approach can be to specify the granularity with which the algorithm may require to proceed which can be specified using a metadata association engine [27]. The resultant groups can be hierarchically organized as preferred and analyzed depending on investigation requirements.

7.4 Trickery and deceit

In recent times, with increased online presence and cloud-based computing, files are transported over the Internet to cloud servers where files can be operated on. While this appears to be a simple extension to the remote server execution mode, the session-less protocols in-play today do not permit the retention of particular file(s) and suitably modify them when they are operated on under cloud environments; rather, it generates an *independent* file transfers which is then treated as *new* file creation activities that is then created on the user's file system. An associated challenge is that often not all metadata may be retained on the processed file. While this can be attributed to computational efficiency, it can have a bearing on ability to accurately identify relationships based on metadata. Consequently, when a processed file is compared against its original, only a few metadata associations are likely to be discovered. It is likely that an intelligent attacker may choose to exploit this technique to confuse a metadata based system into incorrectly attributing associations. Offline analysis may often reveal very little in the absence of extra intelligence. However, under such situations, the approach can resort to analyzing access logs locally and remotely, and also be able to trace live network traffic, it can used in the generation of metadata association groups which can be used to identify the provenance of the files concerned. In the event, complete tracking is not feasible, it may still be possible to detect anomalous associations and alert the user regarding the suspected anomaly. Such anomalies can be detected when associated files from the same association group diverge on the association index value. Finite-state Markov models seem to hold much promise in providing limited correction capabilities where anomalies are detected. The authors are currently exploring this route.

8 Conclusions and future work

In this paper, we studied the use of the metadata association model to analyze collections of digital image files. Using metadata belonging to the four metadata families, it is possible to determine digital image relationships through metadata associations to find answers to questions pertaining to the analysis of digital image collections. We illustrated the use of file relationships to identify doctored digital photographs from image collections and derived documents from document collections as instances of IP theft.

Since this work assumes authenticity of the metadata used for identifying metadata associations, investigating the applicability of this approach under *partial or incorrect* information is one of our directions for the future. The approach requires establishing a prior against which new associations, as they become available, can be compared for consistency. The authors are currently exploring Markov model-based approach in this regard. Besides, many files are being downloaded these days and during investigations, it may be necessary to ascertain their provenance. Using metadata associations to address this problem involves understanding the semantic relationships that exist between files on a file system, temporary folders and internet based logs. Such an approach can develop a technique for the automatic identification of file provenance on large file collections. This provides another direction for future work.

References

1. Alvarez P (2004) Using extended file information (EXIF) file headers in digital evidence analysis. Int J Digit Evid 2(3):1–5
2. Bohm K, Rakow TC (1994) Metadata for multimedia documents. In: Proceedings of ACM SIGMOD RECORD 1994, vol 23(4), 21–26
3. Boutell M, Luo J (2005) Beyond pixels: exploiting camera metadata for photo classification. Pattern Recognit Image Underst Photographs 38(6):935–946. doi:10.1016/j.patcog.2004.11.013 ISSN: 0031-3203
4. Brand A, Daly F, Meyers B (2003) Metadata demystified, The Sheridian and NISO Press, http://www.niso.org/standards/resources/Metadata_Demystified.pdf. Accessed 11 July 2013. ISBN: 1-880124-59-9, pp 1–19
5. Buchholz F, Spafford EH (2004) On the role of system metadata in digital forensics. Digit Invest 1(1):298–309
6. Carrier BD (2005) File system forensic analysis. Addison Wesley Publishers, New York ISBN: 0-32-126817-2
7. Casey E (2011) Digital evidence and computer crime: forensic science, computers and the internet. Academy Press Publications 3/e, Washington D.C ISBN: 978-0-12-374268
8. Castiglione A, De Santis A, Soriente C (2007) Taking advantages of a disadvantage: digital forensics and steganography using document metadata. J Syst Soft 80(5):750–764
9. Chow K, Law F, Kwan M, Lai P (2007) The rules of time on NTFS file system. In: Proceedings of the 2nd international workshop on systematic approaches to digital forensic engineering, April 2007
10. Denecke K, Risse T, Baehr T (2009) Text classification based on limited bibliographic metadata. In: Proceedings of the fourth IEEE international conference on digital information management, ICDIM 2009, ISBN: 978-1-4244-4253-9, 27–32
11. Digital Imaging Group Inc., (2001) DIG35 Specification: metadata for digital images, Version 1.1 April 16th 2001 Working Draft, Digital Imaging Group Inc., 2001-04-16

12. EXIF Specification Document (2002) JEITA CP-3451, Standard of Japan and Information Technology Association, Exif Version 2.2, http://www.exif.org/Exif2-2.PDF. Retrieved 12 July 2011
13. Fathi M, Adly N, Nagi M (2004) Web documents classification using text, anchor, title and metadata information. In: Proceedings of the international conference on computer science, software engineering, information technology, e-Business and Applications, 1–8
14. Garfinkel SL (2009) Digital forensic research: the next 10 years, Digital investigations. In: Proceedings of the 10th annual conference on digital forensic research workshop (DFRWS'10), Vol. 7(2010), S64–S73
15. Garfinkel SL, Farrell P, Roussev V, Dinolt G (2009) Bringing science to digital forensics with standardized forensic corpora, Digital investigation. In: Proceedings of the 9th annual conference on digital forensic research workshop (DFRWS'09), Vol. 6, S2–S11
16. Garfinkel S (2009) Automating disk forensic processing with Sleuthkit, XML and Python. In: Proceedings of the 2009 fourth international ieee workshop on systemmatic approaches to digital forensic engineering (SADFE 2009), Berkeley, California, ISBN: 978-0-7695-3792-4, 73–84
17. Jiang X, Walters A, Xu D, Spafford E, Buchholz F, Wang Y (2007) Provenance-aware tracing of worm break-in and contaminations: a process coloring approach. In: Proceedings of the 24th IEEE international conference on distributed computing systems, (ICDCS 2006), Lisbon, Portugal, ISBN: 0-7695-2540-7, 38
18. Kee E, Farid H (2010) Digital image authentication from thumbnails. In: Proceedings of the SPIE symposium on electronic imaging, media forensics and security, vol 7541, SPIE, (2010)
19. Kee E, Johnson MK, Farid H (2011) Digital image authentication from JPEG headers. IEEE Trans Inform Forensic Secur 6(3): 1066–1075
20. Koen R, Olivier M (2008) The use of file timestamps in digital forensics. In: Proceeding of the Information Security of South Africa (ISSA 2008), 1–16
21. Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the international conference on machine learning (ICML 1998), 296–304
22. Liu X, Zhang L, Li M, Zhang H, Wang D (2005) Boosting image classification with LDA-based feature combination for digital photograph management, J Pattern Recognit, Elsevier Science Publications, ISSN: 0031-3203, 38(6):887–901
23. Maly KJ, Zeil SJ, Zubair M (2007) Exploiting dynamic validation for document layout classification during metadata extraction. In: Proceedings of the IADIS international conference on world wide web and the internet (WWW/Internet 2007), ISBN: 978-972-8924-44-7, 261–268
24. Minack E, Paiu R, Costache S, Demartini G, Gaugaz J, Ioannou E, Chirita P-A, Nejdl W (2010) Leveraging personal metadata for desktop search: the Beagle ++ system, J Web Semantics, Elsevier Science Publications, ISSN: 1570-8268, 8(1):37–54
25. Palmer G (2001) A road map for digital forensic research: DFRWS Technical Report, DTR - T001-01 FINAL, DFRWS Technical Committee
26. Raghavan S, Raghavan SV (2013a) A study of forensic and analysis tools. In: Proceedings of the 2013 8th international workshop on systematic approaches to digital forensics engineering (SADFE), Hong Kong, China Nov 21-22, 2013
27. Raghavan S, Raghavan SV (2013b) AssocGEN: engine for analyzing metadata based associations in digital evidence. In: Proceedings of the 2013 8th international workshop on systematic approaches to digital forensics engineering (SADFE), Hong Kong, China Nov 21–22, 2013
28. Raghavan S, Raghavan S V (2013c). Determining the origin of downloaded files using metadata associations, J Commun, ISSN: 1796-2021, 8(12):902–910
29. Rowe NC, Garfinkel S (2011) Finding anomalous and suspicious files from directory metadata on a large corpus. In: Proceedings of the third international conference on digital forensics and cyber crime, ICDF2C 2011, Dublin, Ireland 2011
30. Toyama K, Logan R, Roseway A, Anadan P (2003) Geographic location tags on digital images. In: Proceedings of ACM Multimedia 2003, Berkeley, California, ISBN: 1-58113-722-2, 156–166
31. Willassen S (2008) Finding evidence of antedating in digital investigations. In: Proceedings of the third international conference on availability, reliability and security, ARES 2008, 26–32
32. Zander S, Nguyen T, Armitage G (2005) Self-learning IP traffic classification based on statistical flow characteristics. In: Proceedings of the Sixth ICST conference on passive active measurement, (PAM 2005), LNCS 3431, Springer-Verlag Publishers, Berlin, 325–328