Determining the Origin of Downloaded files Using Metadata Associations

Sriram Raghavan[1] and S. V. Raghavan [2]
[1] Secure Cyber Space
[2] Department of Computer Science & Engg., IIT Madras, Chennai, INDIA
Email: sriram.raghavan@securecyberspace.org; svr@cs.iitm.ernet.in

Abstract—Determining the "origin of a file" in a file system is often required during digital investigations. While the problem of "origin of a file" appears intractable in isolation, it often becomes simpler if one considers the environmental context, viz., the presence of browser history, cache logs, cookies and so on. Metadata can help bridge this contextual gap. Majority of the current tools, with their search-and-query interface, while enabling extraction of metadata stops short of leading the investigator to the "associations" that metadata potentially point to, thereby enabling an approach to solving the "origin of a file" problem. In this paper, we develop a method to identify the origin of files downloaded from the Internet using metadata based associations. Metadata based associations are derived though metadata value matches on the digital artifacts and the artifacts thus associated, are grouped together automatically. These associations can reveal certain higher-order relationships across different sources such as file systems and log files. We define four relationships between files on file systems and log records in log files which we use to determine the origin of a particular file. The files in question are tracked from the user file system under examination to the different browser logs generated during a user's online activity to their points of origin in the Internet.

Index Terms—metadata association, association group

I. Introduction

During forensic analysis, one is required to determine answers to six fundamental questions with regard to the digital artifacts present in digital evidence, *what*, *where*, *when*, *how*, *who* and *why* [3]. Answers pertaining to *what, who* and *when* are usually determined by examining the individual digital artifacts and their metadata. Metadata in a digital artifact is a commonplace for recording some important information pertaining to the nature of a digital artifact and its value is deep-seated in digital forensics [1]. Answers pertaining to *where*, *how* and *why* are usually fairly involved and require some detailed analysis. In this paper, we are concerned with the "where" question in regards to files suspected to be downloaded from the Internet. This can be of particular relevance when analyzing thumbnail images. Usually, thumbnail images are not subjected to detailed analysis unless the origin of the image files indicates suspicious behavior on the part of a user. To ascertain this, it is necessary to determine the origin of all the files downloaded from the Internet. We propose an automated method using metadata based associations to determine and group the source URL with each suspected file during analysis.

When a source of digital evidence is examined using traditional forensic tools, they deal with a monolithic forensic image of the source; the forensic image is examined using a forensic toolkit like Encase or FTK to examine the file system contents. Thereafter, each file is individually analyzed and its metadata is examined. By virtue of the monolithic nature of the forensic image, the files are examined in isolation. Naturally, when a user activity spans different sources, even as simple as a user file system and log files, the monolithic nature limits the ability to establish event linkages that are necessary to determine the origin of downloaded files.

Conventionally, the effort required to relate a file from the user file system with one or more log records and ascertain the nature of the events recorded in the logs is largely in the realms of a human investigator. This task requires significant man-hours of effort conducting multiple query-based searches using one or more tools in analyzing the digital artifacts [7]. In regards to the task of identifying the source of a downloaded file, unless the investigator has pre-existing knowledge of the source URL or the location is stored in the file metadata, it is likely to be missed. Besides, the user's Internet activities can only be deciphered when the browser logs are examined. Since history logs only record web access records, an investigator would require to simultaneously search the browser cache and history logs and compare them against the files in the user's computer to determine the origin of a file. As the number of suspected files increase, the task becomes demanding. This calls for an automated method that can track user activities across sources and group the related events to achieve a specific objective; in this case, identifying the source of downloaded files. The rest of the paper is organized as follows: In Section II, we review related work and motivate the use of metadata associations to track related artifacts in digital evidence. In Section III, we describe the problem this paper attempts to solve and in Section IV, we present a discussion on the nature of analysis and categorize the metadata across different sources of digital evidence for this purpose. In Section V, we introduce the

Manuscript received August 15, 2013; revised October 30, 2013.
Corresponding author email: sriram.raghavan@securceyberspace.org.
doi:10.12720/jcm.8.12.902-910

different types of metadata associations and define 3 types of artifact relationships that help identify event sequences during online activities. In Section VI, we describe our experimental method. In Section VII, we analyze our results and present a comparative assessment in relation to existing tools. In Section VIII, we provide a summary of our work and define scope for future research based on metadata associations.

II. Related Work

Present day tools are focused on finding pieces of evidence from different sources but do not integrate the information determined to aid in analysis [7]. File system contain metadata associated with file activity which is independent of file content and forensic tools extract these metadata to identify the owner, MAC timestamps, access privileges and so on. However, these tools do not, under normal circumstances, extract or use application metadata from files, which also contain valuable information and they do not seem to correlate the different metadata across files and alert an examiner when related artifacts are discovered during analysis [7], [8]. However, there is some research in identifying user's file and application activity by comparing against volatile memory. Case *et al*. [2] and Cohen [4] propose forensic analysis tools that map network sockets with memory dumps to identify active network connections. Windows registry and current active files can also be discovered from memory [6], [11]. However, these techniques require prior knowledge of the nature of network connections or the file contents to establish the mapping. Since such information is usually available during the forensic examination phase, determining patterns on-the-fly becomes rather involved.

A necessary functionality for forensic and analysis tools is to combine multiple attributes to derive semantic relationships between the various digital artifacts [8], [10]. Metadata based associations to determine relationships in digital artifacts can track user events. By identifying and grouping such related digital artifacts, one can reconstruct event chains using digital time-lining based on the artifact timestamps [9]. In this paper, we extend the definition of metadata to include log and network packet trace attributes and apply the metadata association model to determine the source of downloaded files and group them together during analysis. We adopted the AssocGEN engine [1] [8] to demonstrate it on a synthetic usage scenario involving a user downloading image files from an unknown domain on the Internet and compare our approach to corroborate the results using browser and network analysis tools.

III. Problem Description

Given a snapshot of a user's file system, it is necessary to determine the origin of the files discovered. Fig. 1 displays the *Downloads* folder on the user file system where we are interested in the origin of the digital image highlighted. When suspecting a file as a downloaded resource from the Internet, we search for other locations on the user's computer where a copy (temporary) of the file can be discovered.

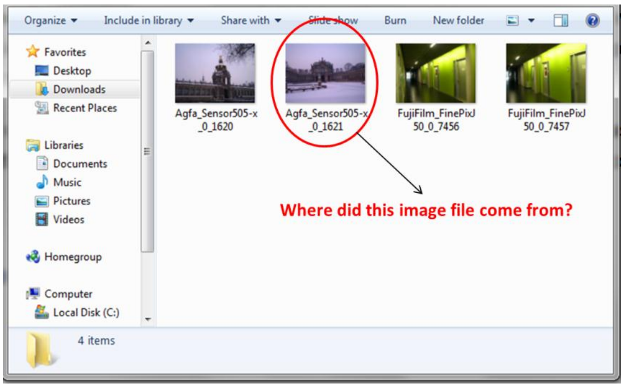


Fig. 1. Snapshot of the user's file system containing some digital image files.

In this case, we discover the presence of a copy in the temporary files folder corresponding to the user's Internet Explorer browser activity. Fig. 2 illustrates the discovery of an identical copy of the image file in the temporary files folder. Having determined the existence of at least one file in the temporary internet files folder, we extract the browser cache and seek file matches and determine their respective attributes, as metadata.

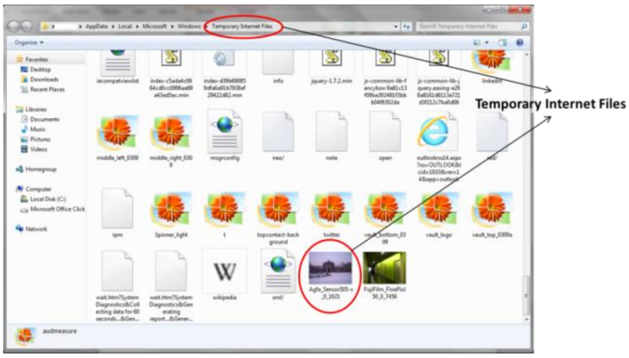


Fig. 2. Snapshot of the user's temporary internet files

The problem, therefore, is to develop a method to automatically identify all the files from the user file system which are present in the temporary internet folder and thereby analyze the Internet browser logs for related events to determine the source URLs corresponding to each downloaded image file.

IV. Nature of Analysis

The quest in the analysis of digital evidence is the identification of the events leading to the reported incident, the nature of these events and their attribution to individual(s). For our discourse, an event refers to actions

[1] Source freely available for download from http://sourceforge.net/projects/assocgen/

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 digital artifacts, or accessing or modifying existing digital artifacts(s). Typically the following are observed when a new event occurs:

- On a file system, an event can create a new file, or access or modify one or more aspects of an existing file.
- On a log file, an event usually creates a new log record. Existing log records are preserved, untouched.
- During a network packet capture session, an event captures a new network packet. Existing network packets are preserved, untouched.

If a new digital artifact is created as a result of an event, its occurrence is reflected in the metadata that are also created along with the digital artifact. If an existing artifact is modified as a result of an event, its occurrence is reflected in the change in values of the metadata linked to that artifact. Therefore, irrespective of the type of event, its effect can be perceived in the metadata.

The analysis is concerned with finding answers to the questions that relate to *what*, *when*, *where*, *how*, *who* and *why* [3]. Naturally the process of analysis is driven by methods to find these answers. The most common form of grouping metadata is timestamps with owner for files, username for logs or IP address for network packet traces. The motivation beyond this grouping is evident since it helps one find answers to *who* and *when*. To determine answers to *what*, *where*, and *how*, the artifacts are individually analyzed with perhaps, keyword filtering. However, this can be an extended process and may require multiple back-and-forth activities to determine the exact nature of the events recorded in evidence.

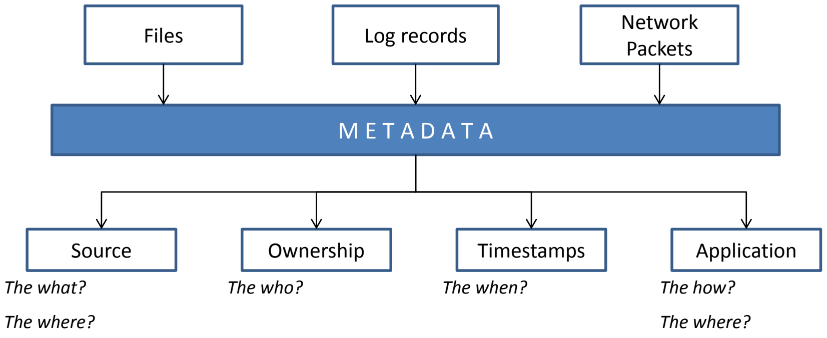


Fig. 3. Metadata families pertinent to forenic analysis

When an event creates or modifies more than one digital artifact, identifying the metadata that pertain to the event across these artifacts will elicit the relationships that exist between them. Therefore, focusing on the appropriate metadata across the digital artifacts, one can reconstruct the event(s). Naturally, it is necessary to determine the classes of metadata from such artifacts that can provide specific answers to the questions raised during forensic analysis. Typically, questions of the type “what” or “where” relate to the source of the artifact and the metadata that identify such sources are potential candidates for finding the answers. The “who” question identifies an individual who is or a system that is attributed to an artifact. The “when” question relates to the time-related event(s) that affected an artifact and the timestamps in metadata can provide such answers. The “how” question pertains to describing other aspects pertaining to an artifact when an event affecting the artifact was observed. Therefore, metadata that identify such situational information are likely candidates. This is diagrammatically illustrated in Fig. 3.

V. Determining Associations and Relationships using Metadata

In this section, we introduce the theory relating to the identification of associations using metadata and the discovery of relationships between the concerned artifacts to trace the origin of files from the Internet.

A. *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:

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

Partial association: When a particular metadata value in one digital artifact matches partially with the corresponding metadata on another artifact, for a value of STRING type, a *partial association* is said to occur between the artifacts 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 s1 match exactly with the corresponding characters in s2 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

Threshold association: When a particular metadata value in one digital artifact differs with the corresponding metadata on another artifact, 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 artifacts 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 artifact on which the comparison is pivoted, is identified.

Date association: When a particular metadata value in one digital artifact, for a value of DATE type, is matched against with the corresponding metadata on another artifact, it defines a *date association* between the said artifacts for that metadata. Such a date association can occur in 4 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 artifact 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 artifact 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.

B. Artifact Relationships for tracing Online Activities

When we determine metadata associations across artifacts, it underlines the relationship between the artifacts which can reveal the nature of activities recorded. In this section, we define four types of artifact relationships based on metadata associations to discern the nature of online user activities. We identify metadata associations using value matches between the digital images across different classes to form groups of associated image files called '*association groups*' [9]. 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.

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, DOCX, .BAK or .TMP. The relationship is denoted by R_e and it may be expressed as $f_1 R_e f_2$ and read as $f_1 \Leftrightarrow f_2$. 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. $f_1 R_e f_2 \Leftrightarrow f_2 R_e f_1$
2. $(f_1 R_e f_2) \wedge (f_2 R_e f_3) \Rightarrow (f_1 R_e f_3)$

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

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 $f_1 R_s f_2$. By definition this relationship is commutative and associative. Therefore,

1. $f_1 R_s f_2 \Leftrightarrow f_2 R_s f_1$
2. $(f_1 R_s f_2) \wedge (f_2 R_s f_3) \Rightarrow (f_1 R_s f_3)$

When multiple such files ($f_1, f_2, f_3, \ldots f_n$) 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 $R_s\ (f_1, f_2, f_3, \ldots f_n)$.

Download Relationship: When the filename of a file *f* on the user file system generates a source metadata family metadata match with a download resource *r* recorded in a browser cache log, we define a *download relationship* indicating the download of the resource *r* to the user file system. The relationship is denoted by R_d and expressed as $f R_d r$ indicating the creation of *f* implies the download of resource *r*.

Happens Relationship: When a metadata match occurs on the ownership metadata family of log files such as the log records of the web history and cache logs of a web browser, we define a *happens relationship* indicating the happening of a web page visit prior to the download of the specified resource on the cache log. The relationship is denoted by R_h and expressed as $x\ \mathrm{R}_h\ y$ where x is the digital artifact corresponding to the browser history log and y is the artifact corresponding to the browser cache log. In general, for two events x, y, time(x) < time(y) indicating that x happened before y, and $y \Rightarrow x$. By definition, the relationship is not commutative but associative;

$$(x\mathrm{R_h}y) \wedge (y\mathrm{R_h}z) \Rightarrow (x\mathrm{R_h}z);\text{ and}$$

$$\{(z \Rightarrow y) \wedge (y \Rightarrow x)\} \Rightarrow (z \Rightarrow x)$$

VI. Method

Scenario. Set up a virtual machine and create a user account. Generate the following constructed scenario. Login to the user account and set up a user email account. Using the Internet browser, browse the Internet, arbitrarily choose a website and view the images on that website. Download images to the user's computer. Access user's email and view the messages and attached images using the browser. Download some of the attachment image files to the user's computer. Capture steps browsing activity using Wireshark network packet capture.

Analysis. Isolate the computer and create a virtual machine snapshot of the user file system. Isolate the Internet browser history and cache logs for analysis. Use AssocGEN tool and load the different sources, viz., user disk image, browser history and cache logs, network packet capture. Traverse the user's computer and determine the files containing *existence* relationships. Determine *download* relationships on the files and establish *happens* relationships from the corresponding browser logs. Using metadata associations, identify all relationships to determine the source of the files. Group each image file with its corresponding source URL from the browser logs. Repeat the steps using a different browser and note the observations using this method.

Corroboration. Examine the sources using traditional forensic and analysis tools. Use FTK to examine the file system forensic image. Use web analysis tools to examine the Internet browser logs. Determine the origin of the images discovered on the user's computer. Corroborate the results of the web analysis tools against Wireshark packet capture analysis.

A. *Basis of the Experiment*

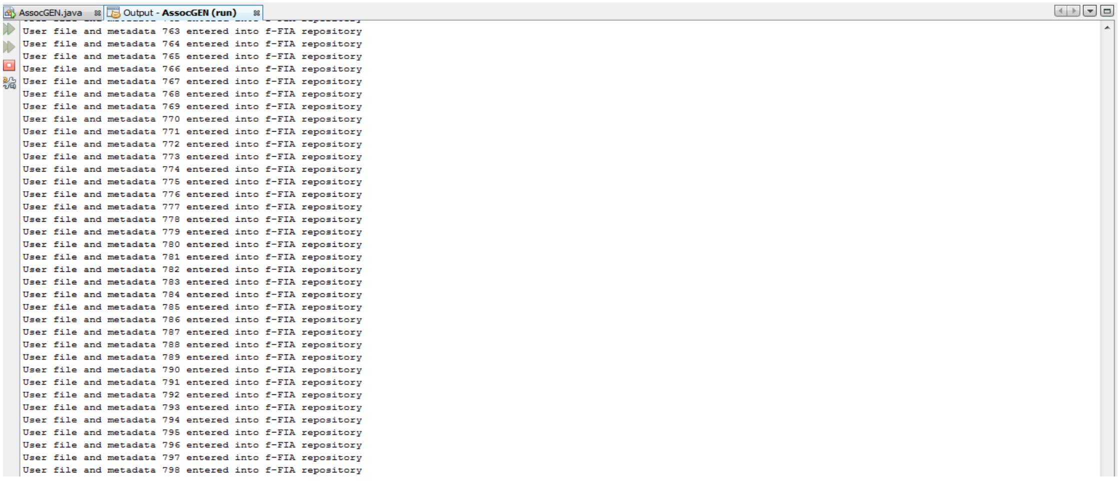
The AssocGEN tool, by design, segregates user documents, temporary Internet files, system and application logs including browser history and cache logs, network trace as distinct sources. Since AssocGEN is developed based on metadata associations, digital artifacts from different sources recording different stages of an activity can be accessed through associations. While any standard forensic tool treats browser logs are mere files, AssocGEN treats each log record as independent user activity and enables the identification of events that occurred affecting the user files as well as the logs concerned. The artifacts, associated via their metadata, are then grouped together depending on the nature of activities being investigated. The dataset generated from the scenario described in this experiment is summarized in Table I.

TABLE I. SUMMARY OF THE EVIDENCE ANALYZED AND THEIR CHARACTERISTICS

Characteristics	
User files	47, 699 (30 GB)
Temporary Internet files	8916
Browser history	115832
Browser cache	128624
Network packets in trace	35035

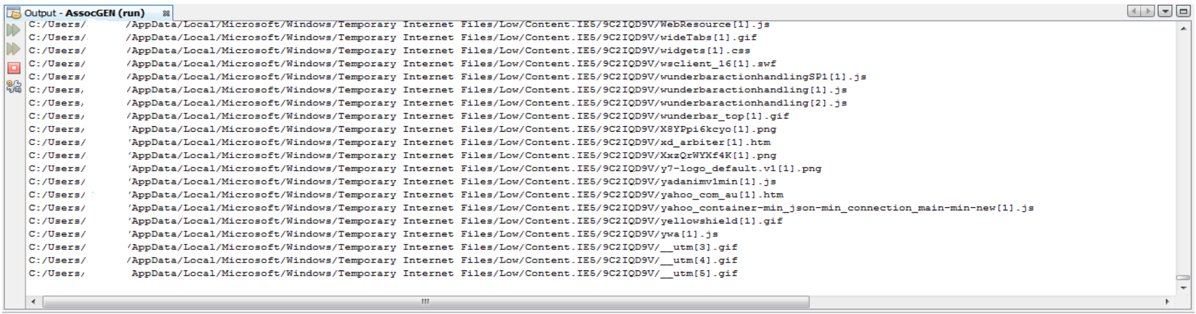
Using metadata associations, the log records are grouped with the related files tracing the event sequences to help an examiner. When a particular file is identified, all log events pertaining to that file can be identified and grouped and if this happens to relate to the Internet browser logs, then the source URL and the domain can be identified.

B. *Observations*



```
User file and metadata 763 entered into f-FIA repository
User file and metadata 764 entered into f-FIA repository
User file and metadata 765 entered into f-FIA repository
User file and metadata 766 entered into f-FIA repository
User file and metadata 767 entered into f-FIA repository
User file and metadata 768 entered into f-FIA repository
User file and metadata 769 entered into f-FIA repository
User file and metadata 770 entered into f-FIA repository
User file and metadata 771 entered into f-FIA repository
User file and metadata 772 entered into f-FIA repository
User file and metadata 773 entered into f-FIA repository
User file and metadata 774 entered into f-FIA repository
User file and metadata 775 entered into f-FIA repository
User file and metadata 776 entered into f-FIA repository
User file and metadata 777 entered into f-FIA repository
User file and metadata 778 entered into f-FIA repository
User file and metadata 779 entered into f-FIA repository
User file and metadata 780 entered into f-FIA repository
User file and metadata 781 entered into f-FIA repository
User file and metadata 782 entered into f-FIA repository
User file and metadata 783 entered into f-FIA repository
User file and metadata 784 entered into f-FIA repository
User file and metadata 785 entered into f-FIA repository
User file and metadata 786 entered into f-FIA repository
User file and metadata 787 entered into f-FIA repository
User file and metadata 788 entered into f-FIA repository
User file and metadata 789 entered into f-FIA repository
User file and metadata 790 entered into f-FIA repository
User file and metadata 791 entered into f-FIA repository
User file and metadata 792 entered into f-FIA repository
User file and metadata 793 entered into f-FIA repository
User file and metadata 794 entered into f-FIA repository
User file and metadata 795 entered into f-FIA repository
User file and metadata 796 entered into f-FIA repository
User file and metadata 797 entered into f-FIA repository
User file and metadata 798 entered into f-FIA repository
```

Fig. 4. AssocGEN analysis engine processing user file system and loading metadata into f-FIA repository

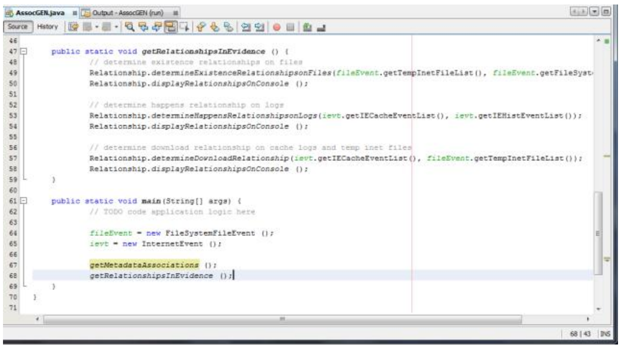


```
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/WebResource[1].js
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wideTabs[1].gif
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/widgets[1].css
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wsclient_16[1].swf
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wunderbaractionhandlingSP1[1].js
C:/Users,      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wunderbaractionhandling[1].js
C:/Users,      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wunderbaractionhandling[2].js
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/wunderbar_top[1].gif
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/X8YPpi6kcyo[1].png
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/xd_arbiter[1].htm
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/XxzQrWYXf4K[1].png
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/y7-logo_default.v1[1].png
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/yadanimv1min[1].js
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/yahoo_com_au[1].htm
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/yahoo_container-min_json-min_connection_main-min-new[1].js
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/yellowshield[1].gif
C:/Users/       'AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/ywa[1].js
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/__utm[3].gif
C:/Users/      /AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/__utm[4].gif
C:/Users/       AppData/Local/Microsoft/Windows/Temporary Internet Files/Low/Content.IE5/9C2IQD9V/__utm[5].gif
```

Fig. 5. AssocGEN processing the temporary internet files

We use AssocGEN to first process the user file system and load the files and their metadata (after parsing) into the *f*-FIA repository which is the repository associated with the AssocGEN engine storing the metadata associations. Fig. 4 is a snapshot of AssocGEN loading the user files into the repository. Once the user files are completed, AssocGEN tracks all internet based activity which includes traversing the temporary internet files folder used by the web browser to temporarily store downloaded web resources. Fig. 5 is a snapshot of AssocGEN traversing the temporary internet files before parsing the metadata and loading them into the *f*-FIA repository.

After the files are processed, AssocGEN extracts the browser history and cache events which are, likewise, loaded into the repository with their respective attributes. After this, the analysis engine makes a procedure call to generate all metadata associations. Once the associations and generated and stored into the repository, a procedure call is made to discern the relationships that exist among the associations which can provide the origins of the image files in question. Fig. 6 is a snapshot of the execution logic in AssocGEN to determine metadata associations in evidence followed by the extraction of relevant relationships leading to the determination of the origin of the image files.



```

    public static void getRelationshipsInEvidence () {
            // determine existence relationships on files
            Relationship.determineExistenceRelationshipsonFiles(fileEvent.getTempInetFileList(), fileEvent.getFileSyst
            Relationship.displayRelationshipsOnConsole ();

            // determine happens relationship on logs
            Relationship.determineHappensRelationshipsonLogs(ievt.getIECacheEventList(), ievt.getIEHistEventList());
            Relationship.displayRelationshipsOnConsole ();

            // determine download relationship on cache logs and temp inet files
            Relationship.determineDownloadRelationship(ievt.getIECacheEventList(), fileEvent.getTempInetFileList());
            Relationship.displayRelationshipsOnConsole ();
    }

    public static void main(String[] args) {
            // TODO code application logic here

            fileEvent = new FileSystemFileEvent ();
            ievt = new InternetEvent ();

            getMetadataAssociations ();
            getRelationshipsInEvidence ();
    }
}

```

Fig. 6. AssocGEN code logic

The *existence relationships* R$_e$ are determined to exist between the user files and their copies in the temporary internet files folder, the *happens relationships* R$_h$ are determined to exist between the browser logs obtained from the browser history and cache and the *download relationships* R$_d$ are determined to exist between the browser cache and the temporary files. The relationships determined from the metadata associations for the Internet Explorer browser is tabulated in Table II. The results were found to be identical when the experiment was repeated using the Mozilla Firefox browser.

Since we were primarily interested in establishing the origin of the digital image files discovered on the user's file system, we only focused on those set of image files, 142 in number. These were the digital image files that were discovered in the temporary files folder of the user's computer. When we compared the browser logs (history ad cache), we derived 424 relationships which identified 424 unique resources that were visited and downloaded. These records provided the *happens* relationship R$_h$. A similar relationship was also determined between the browser cache and the temporary files folder giving rise to 424 unique files being discovered on the temporary files folder. These included the 142 digital image files and other web resources such as validation scripts (.js) and bitmap images (.bmp). For the sake of this exercise, we only focused on identifying those *download* R$_d$ and *happens* R$_h$ relationships identified between the user's computer and the web domain ascertained as the origin. Other activities including normal web browsing activities of the user were omitted.

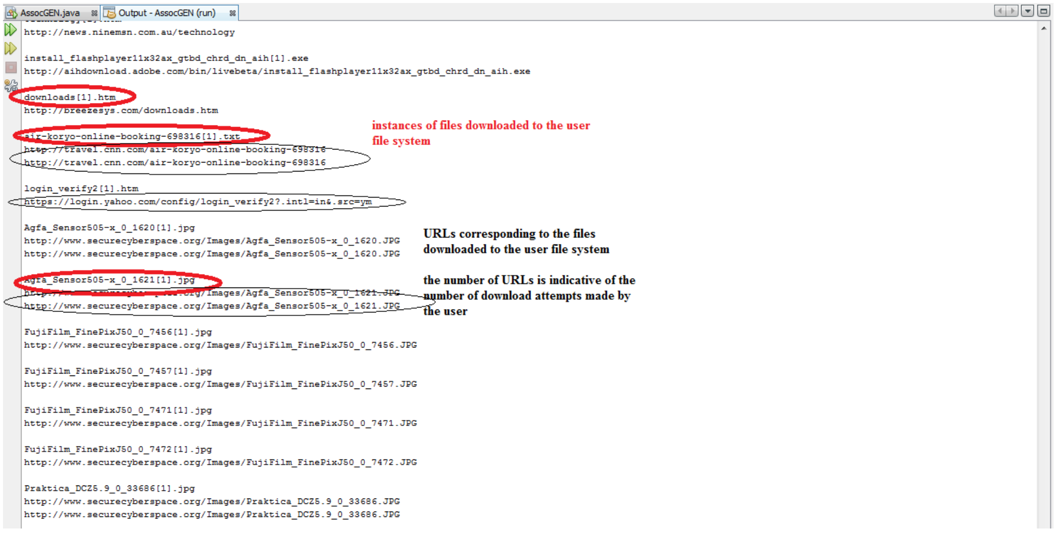
TABLE II. TABULATING THE DISCOVERED METADATA BASED RELATIONSHIPS IN EVIDENCE

Type of artifact relationships	**Number of relationships discovered**
Existence relationships R$_e$	142
Happens relationships R$_h$	424
Source relationships R$_s$	3
Download relationships R$_d$	424

The relationships also identified 3939 digital image files on the user's file system which were captured using three distinct digital still cameras, namely, AgfaSensor 505, FugiFilm_FinePixJ50, and Pracktika_DCZ5.9 as determined from their EXIF metadata. These digital image files indicated 3 respective source relationships with the digital image files whose origin is our subject of discussion. Each of these files contained timestamps that preceded the duration of the experiment. The log records relating to these files were not available in the corresponding browser history and cache logs which we determined was owing to a 7-day history window determined from the browser's registry settings. Therefore, only *source* relationships based on camera EXIF metadata were determined for these files.

VII. Analysis

Once the relationships are determined in evidence, the file origin are determined by mapping the web page linked to the download of the resource leading to the identification of the files stored in the temporary internet files folder and the presence in the user file system. Fig 6 shows the pairings of the image files discovered on the user's computer and their respective web page origins. In each grouping that is shown in Fig. 7, the image file name is printed first followed by the URI corresponding to the web page visit in the browser log. The groupings where multiple URIs are listed with a digital image indicate multiple visit counts that represents the number of additional copies that were downloaded to the user file system. In all, there were 142 digital images that were downloaded from the specified web domain consisting and also consisted of 282 other web resources such as validation scripts (.js) and bitmap files (.bmp) which were discovered on the temporary files folder. Besides, the metadata associations determined that the user file system also contained 3939 digital image files which were taken with 3 different digital still cameras (AgfaSensor 505, FugiFilm_FinePixJ50, and Pracktika_DCZ5.9) and exhibited source relationships with the digital image files downloaded from the specified web domain. These findings seemed to indicate that these digital images were also likely to have been downloaded from the same web domain, although there is no current trace of this in the evidence other than the image relationships determined.



```
http://news.ninemsn.com.au/technology

install_flashplayer11x32ax_gtbd_chrd_dn_aih[1].exe
http://aihdownload.adobe.com/bin/livebeta/install_flashplayer11x32ax_gtbd_chrd_dn_aih.exe

downloads[1].htm
http://breezesys.com/downloads.htm

air-koryo-online-booking-698316[1].txt
http://travel.cnn.com/air-koryo-online-booking-698316
http://travel.cnn.com/air-koryo-online-booking-698316

login_verify2[1].htm
https://login.yahoo.com/config/login_verify2?.intl=in&.src=ym

Agfa_Sensor505-x_0_1620[1].jpg
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1620.JPG
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1620.JPG

Agfa_Sensor505-x_0_1621[1].jpg
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1621.JPG
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1621.JPG

FujiFilm_FinePixJ50_0_7456[1].jpg
http://www.securecyberspace.org/Images/FujiFilm_FinePixJ50_0_7456.JPG

FujiFilm_FinePixJ50_0_7457[1].jpg
http://www.securecyberspace.org/Images/FujiFilm_FinePixJ50_0_7457.JPG

FujiFilm_FinePixJ50_0_7471[1].jpg
http://www.securecyberspace.org/Images/FujiFilm_FinePixJ50_0_7471.JPG

FujiFilm_FinePixJ50_0_7472[1].jpg
http://www.securecyberspace.org/Images/FujiFilm_FinePixJ50_0_7472.JPG

Praktica_DCZ5.9_0_33686[1].jpg
http://www.securecyberspace.org/Images/Praktica_DCZ5.9_0_33686.JPG
http://www.securecyberspace.org/Images/Praktica_DCZ5.9_0_33686.JPG
```

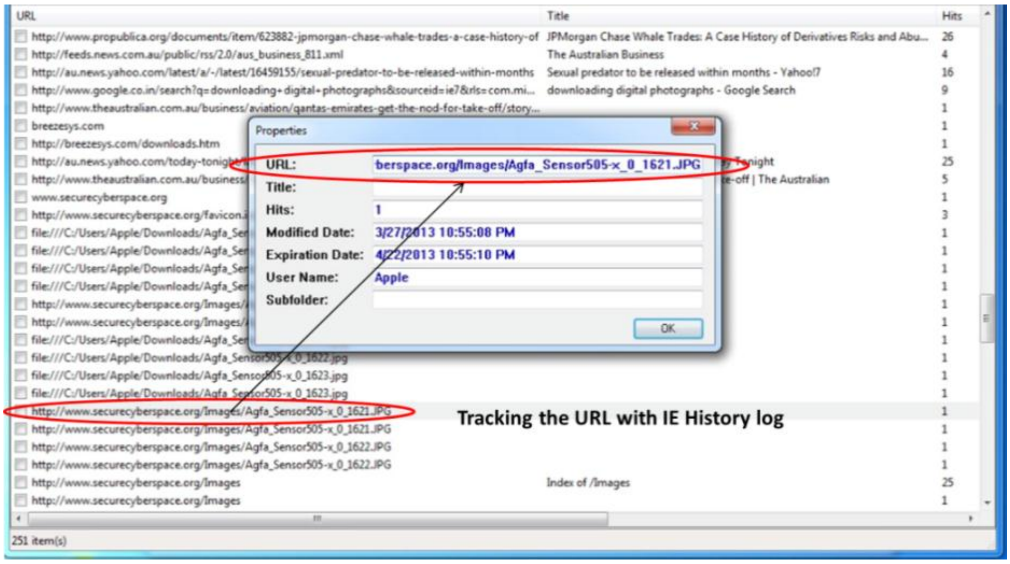
instances of files downloaded to the user file system

URLs corresponding to the files downloaded to the user file system

the number of URLs is indicative of the number of download attempts made by the user

Fig. 7. AssocGEN pairing the image files with their respective web page origins

To corroborate the findings, we analyze the browser history logs (Fig. 8) and determine the origin by tracking the URL in the attribute corresponding to the resource in question. In the figure below, visitation of the image file on a website is identified which also provides us with a URL.



URL	Title	Hits
http://www.propublica.org/documents/item/623882-jpmorgan-chase-whale-trades-a-case-history-of	JPMorgan Chase Whale Trades: A Case History of Derivatives Risks and Abu...	26
http://feeds.news.com.au/public/rss/2.0/aus_business_811.xml	The Australian Business	4
http://au.news.yahoo.com/latest/a/-/latest/16459155/sexual-predator-to-be-released-within-months	Sexual predator to be released within months - Yahoo!7	16
http://www.google.co.in/search?q=downloading+digital+photographs&sourceid=ie7&rls=com.mi...	downloading digital photographs - Google Search	9
http://www.theaustralian.com.au/business/aviation/qantas-emirates-get-the-nod-for-take-off/story...		1
breezesys.com		1
http://breezesys.com/downloads.htm		1
http://au.news.yahoo.com/today-tonigh	Tonight	25
http://www.theaustralian.com.au/business	e-off | The Australian	5
www.securecyberspace.org		1
http://www.securecyberspace.org/favicon.i		3
file:///C:/Users/Apple/Downloads/Agfa_Se		1
file:///C:/Users/Apple/Downloads/Agfa_Se		1
file:///C:/Users/Apple/Downloads/Agfa_Se		1
file:///C:/Users/Apple/Downloads/Agfa_Se		1
http://www.securecyberspace.org/Images/		1
http://www.securecyberspace.org/Images/		1
file:///C:/Users/Apple/Downloads/Agfa_Se		1
file:///C:/Users/Apple/Downloads/Agfa_Sensor505-x_0_1622.jpg		1
file:///C:/Users/Apple/Downloads/Agfa_Sensor505-x_0_1623.jpg		1
file:///C:/Users/Apple/Downloads/Agfa_Sensor505-x_0_1623.jpg		1
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1621.JPG		1
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1621.JPG		1
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1622.JPG		1
http://www.securecyberspace.org/Images/Agfa_Sensor505-x_0_1622.JPG		1
http://www.securecyberspace.org/Images	Index of /Images	25
http://www.securecyberspace.org/Images		1

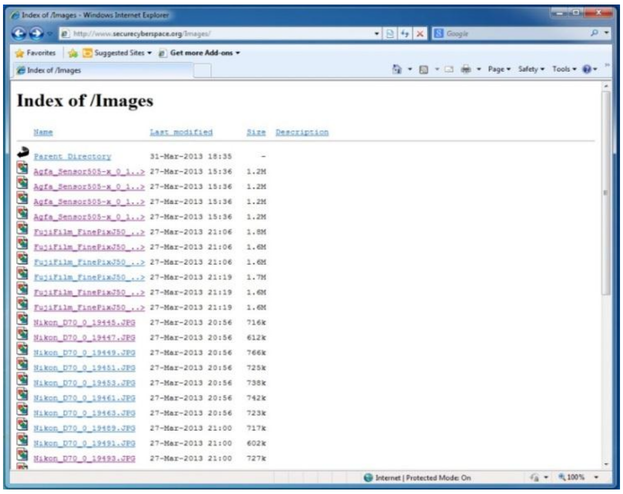
251 item(s)

Properties

URL: berspace.org/Images/Agfa_Sensor505-x_0_1621.JPG
Title:
Hits: 1
Modified Date: 3/27/2013 10:55:08 PM
Expiration Date: 4/22/2013 10:55:10 PM
User Name: Apple
Subfolder:

Tracking the URL with IE History log

Fig. 8. Analysis of IE History - identifying the origin of download



Index of /Images - Windows Internet Explorer
http://www.securecyberspace.org/Images/
Favorites
Suggested Sites
Get more Add-ons
Index of /Images
Page
Safety
Tools
Internet | Protected Mode: On
100%

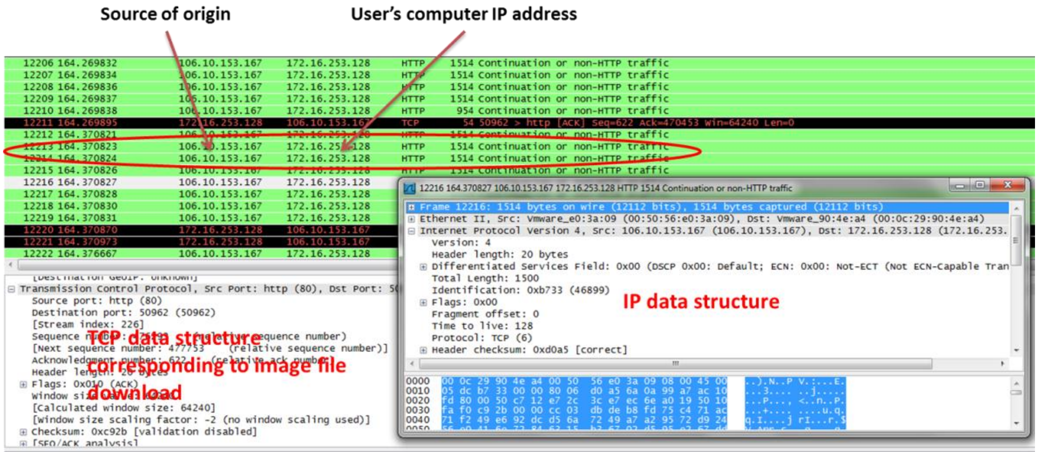
```
Index of /Images

Name                        Last modified      Size  Description

Parent Directory            31-Mar-2013 18:35     -
Agfa_Sensor505-x_0_1..>     27-Mar-2013 15:36  1.2M
Agfa_Sensor505-x_0_1..>     27-Mar-2013 15:36  1.2M
Agfa_Sensor505-x_0_1..>     27-Mar-2013 15:36  1.2M
Agfa_Sensor505-x_0_1..>     27-Mar-2013 15:36  1.2M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:06  1.8M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:06  1.6M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:06  1.6M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:19  1.7M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:19  1.6M
FujiFilm_FinePixJ50_..>     27-Mar-2013 21:19  1.6M
Nikon_D70_0_19445.JPG       27-Mar-2013 20:56  716k
Nikon_D70_0_19447.JPG       27-Mar-2013 20:56  612k
Nikon_D70_0_19449.JPG       27-Mar-2013 20:56  766k
Nikon_D70_0_19451.JPG       27-Mar-2013 20:56  725k
Nikon_D70_0_19453.JPG       27-Mar-2013 20:56  738k
Nikon_D70_0_19461.JPG       27-Mar-2013 20:56  742k
Nikon_D70_0_19463.JPG       27-Mar-2013 20:56  723k
Nikon_D70_0_19489.JPG       27-Mar-2013 21:00  717k
Nikon_D70_0_19491.JPG       27-Mar-2013 21:00  602k
Nikon_D70_0_19493.JPG       27-Mar-2013 21:00  727k
```

Fig. 9. Snapshot of the specified webpage corroborating the listed files in user's computer

To corroborate this finding, we visit the website (webpage snapshot illustrated in Fig. 9) and determine that the image file is indeed listed. In addition, we also note the presence of other files which are likely to be present on the user file system. When the findings are corroborated against the network packet trace, we obtain a similar assessment as illustrated in Fig. 10. However, if we were to incorporate the network trace as another source of evidence into AssocGEN, then the analysis engine will simply group the respective TCP sessions between the domain of origin and the user's computer and incorporate it into the association groups corresponding to the appropriate relationships. The corroboration of these results conclusively establishes that the image files analyzed were in fact downloaded from the specified website.



Source of origin
User's computer IP address
TCP data structure corresponding to image file download
IP data structure

```
12206 164.269832   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12207 164.269834   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12208 164.269836   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12209 164.269837   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12210 164.269838   106.10.153.167   172.16.253.128   HTTP    954 Continuation or non-HTTP traffic
12211 164.269895   172.16.253.128   106.10.153.167   TCP      54 50962 > http [ACK] Seq=622 Ack=470453 Win=64240 Len=0
12212 164.370821   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12213 164.370823   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12214 164.370824   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12215 164.370826   106.10.153.167   172.16.253.128   HTTP   1514 Continuation or non-HTTP traffic
12216 164.370827   106.10.153.167   172.16.253.128
12217 164.370828   106.10.153.167   172.16.253.128
12218 164.370830   106.10.153.167   172.16.253.128
12219 164.370831   106.10.153.167   172.16.253.128
12220 164.370870   172.16.253.128   106.10.153.167
12221 164.370973   172.16.253.128   106.10.153.167
12222 164.376667   106.10.153.167   172.16.253.128
```

```
  [Destination GeoIP: Unknown]
⊟ Transmission Control Protocol, Src Port: http (80), Dst Port: 5(
    Source port: http (80)
    Destination port: 50962 (50962)
    [Stream index: 226]
    Sequence number: [illegible]    (relative sequence number)
    [Next sequence number: 477753    (relative sequence number)]
    Acknowledgment number: 622    (relative ack number)
    Header length: 20 bytes
  ⊞ Flags: 0x010 (ACK)
    Window size value: [illegible]
    [Calculated window size: 64240]
    [Window size scaling factor: -2 (no window scaling used)]
  ⊞ Checksum: 0xc92b [validation disabled]
  ⊞ [SEQ/ACK analysis]
```

```
12216 164.370827 106.10.153.167 172.16.253.128 HTTP 1514 Continuation or non-HTTP traffic

⊞ Frame 12216: 1514 bytes on wire (12112 bits), 1514 bytes captured (12112 bits)
⊞ Ethernet II, Src: Vmware_e0:3a:09 (00:50:56:e0:3a:09), Dst: Vmware_90:4e:a4 (00:0c:29:90:4e:a4)
⊟ Internet Protocol Version 4, Src: 106.10.153.167 (106.10.153.167), Dst: 172.16.253.128 (172.16.253.
    Version: 4
    Header length: 20 bytes
  ⊞ Differentiated Services Field: 0x00 (DSCP 0x00: Default; ECN: 0x00: Not-ECT (Not ECN-Capable Tran
    Total Length: 1500
    Identification: 0xb733 (46899)
  ⊞ Flags: 0x00
    Fragment offset: 0
    Time to live: 128
    Protocol: TCP (6)
  ⊞ Header checksum: 0xd0a5 [correct]

0000  00 0c 29 90 4e a4 00 50  56 e0 3a 09 08 00 45 00   ..).N..P V.:...E.
0010  05 dc b7 33 00 00 80 06  d0 a5 6a 0a 99 a7 ac 10   ...3.... ..j.....
0020  fd 80 00 50 c7 12 e7 2c  3c e7 ec 6e a0 19 50 10   ...P..., <..n..P.
0030  fa f0 c9 2b 00 00 cc 03  db de b8 fd 75 c4 71 ac   ...+.... ....u.q.
0040  71 f2 49 e6 92 dc d5 6a  72 49 a7 a2 95 72 d9 24   q.I....j rI...r.$
```

Fig. 10. Corroborating the findings with network trace analysis

VIII. Conclusions & Future Work

We demonstrated the use of metadata based associations to determine relationships between different sources of digital evidence, viz., user file system, browser logs and temporary internet files to determine the origin of digital image files downloaded from the Internet. The metadata in a file is used to track alternate copies of the file and log events that created / affected the file during a user's online sessions. Using metadata associations, we determined file-file, file-log event, log event-log event relationships which is then traced to the source URL of the file download.

In the future, we intend to extend this work to study the complexities associated with determining the source URLs under partial information. Research is currently underway to adapt the AssocGEN engine to operate under partial information.

References

[1] F. Buchholz and E. H. Spafford, "On the role of system metadata in digital forensics," *Digital Investigations*, vol. 1, no. 1, pp. 298-309, 2004.

[2] A. Case, A. Cristina, L. Marziale, G. G. Richard, and V. Roussev, "FACE: Automated digital evidence discovery and correlation," in *Proc. 8th Annual Digital Forensic Research Workshop*, (Supplement 1), 2008, pp. S65-S75.

[3] E. Casey, "Digital evidence and computer crime: Forensic science, computers and the internet," *Academy Press Publications 3/e*, 2011.

[4] M. I. Cohen, "PyFlag–An advanced network forensic framework," in *Proc. 8th Annual Digital Forensic Research Workshop*, (Supplement 1), 2008, pp. S112-S120.

[5] Digital Imaging Group Inc., DIG35 Specification – Metadata for Digital Images, Version 1.1 April 16th 2001 Working Draft, *Digital Imaging Group Inc.*, TR2001-04-16, 2001.

[6] B. Dolan-Gavitt, "Forensic analysis of Windows Registry in Memory," in *Proc. 8th Annual Digital Forensic Research Workshop*, (Supplement 1), 2008, pp. S26-32.

[7] S. L. Garfinkel, "Digital forensic research: The next 10 years," in *Proc. 10th Annual Conference on Digital Forensic Research Workshop*, vol. 7, 2010, pp. S64-S73.

[8] S. Raghavan and S. V. Raghavan, "A study of forensic and Analysis tools," in *Proc. 8th International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE 2013)*, Hong Kong, China, Nov 21-22, 2013.

[9] S. Raghavan and S. V. Raghavan, "AssocGEN: Engine for analyzing metadata based associations in Digital Evidence," (Accepted) in *Proc. 8th International Conference on Systematic*

Approaches to Digital Forensic Engineering (SADFE 2013), Hong Kong, China Nov 21-22, 2013.

[10] N. C. Rowe and S. Garfinkel, "Finding anomalous and suspicious files from directory metadata on a large corpus," in *Proc. Third International Conference on Digital Forensics and Cyber Crime*, Dublin, Ireland, 2011.

[11] R. B. van Baar, W. Alink, and A. R. Van Ballegooji, "Forensic memory analysis: files mapped in memory," in *Proc. 8th Annual Digital Forensic Research Workshop*, (Supplement 1), vol. 5, 2008, pp. S52-S57.

Sriram Raghavan is a Security and Forensic Consultant with Secure Cyber Space from where he consults on many issues the areas of Cyber Security and Digital Forensics. Sriram has been a working in the areas of digital forensics for over 7 years and specializes in the determinations of associations among digital evidence for the purposes of evidence compositions and event reconstruction. Sriram is finalizing his doctoral dissertation in this area for submission at the Queensland University of Technology, Brisbane where he has also been a senior researcher since 2008. In addition to digital forensics, Mr. Raghavan has been interested in Systems and Architecture and has al and has also donned the roles of an architect responsible for designing Mobile signaling systems and unstructured intelligence systems through is experience in Mobile Computing systems and Multi-agent AI. Mr. Raghavan can be reached at sriram.raghavan@securecyberspace.org

S. V. Raghavan is a professor of Computer Science and Engineering at the prestigious Indian Institute of Technology Madras and holds an adjunct professorship at the Indian Institute of Technology in Delhi. Prof. Raghavan is a well-known author of Multimedia and Networked Systems and fondly known as *Father* of Educational and Research Network (ERNet) in India. Prof. Raghavan has recently architected India's National Knowledge Network (NKN) which connects over 1500 research and educational institutions all over India. Prof. Raghavan is currently the Scientific Secretary in the Office of the Principal Scientific Advisor to the Govt. of India. Prof. Raghavan can be reached at svr@cs.iitm.ernet.in.