Challenges, Opportunities and a Framework for Web Environment Forensics

Mike Mabey, Adam Doupé, Ziming Zhao, Gail-Joon Ahn

Chapter 2

CHALLENGES, OPPORTUNITIES AND A FRAMEWORK FOR WEB ENVIRONMENT FORENSICS

Mike Mabey, Adam Doupé, Ziming Zhao and Gail-Joon Ahn

Abstract The web has evolved into a robust and ubiquitous platform, changing almost every aspect of people's lives. The unique characteristics of the web pose new challenges to digital forensic investigators. For example, it is much more difficult to gain access to data that is stored online than it is to access data on the hard drive of a laptop. Despite the fact that data from the web is more challenging for forensic investigators to acquire and analyze, web environments continue to store more data than ever on behalf of users.

This chapter discusses five critical challenges related to forensic investigations of web environments and explains their significance from a research perspective. It presents a framework for web environment forensics comprising four components: (i) evidence discovery and acquisition; (ii) analysis space reduction; (iii) timeline reconstruction; and (iv) structured formats. The framework components are non-sequential in nature, enabling forensic investigators to readily incorporate the framework in existing workflows. Each component is discussed in terms of how an investigator might use the component, the challenges that remain for the component, approaches related to the component and opportunities for researchers to enhance the component.

Keywords: Web environments, forensic framework, timelines, storage formats

1. Introduction

The web has transformed how people around the globe interact with each other, conduct business, access information, enjoy entertainment and perform many other activities. Web environments, which include all types of web services and cloud services with web interfaces, now offer mature feature sets that, just a few years ago, could only have been

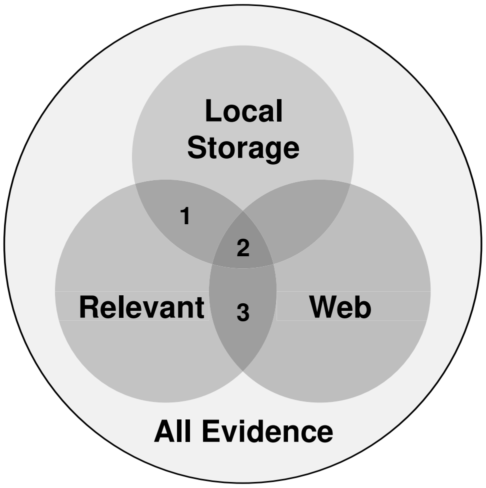

Figure 1. Types of evidence acquired during investigations.

provided by software running on a desktop computer. As such, the web provides users with new levels of convenience and accessibility, which have resulted in a phenomenon that critically impacts digital forensic investigations – people are storing less and less data on their local devices in favor of web-based solutions.

Current digital forensic techniques are good at answering questions about the evidence stored on devices involved in an incident. However, the techniques struggle to breach this boundary to handle evidentiary data that is stored remotely on the web. As Figure 1 illustrates, if forensic investigators depend only on the storage of the devices they seize as evidence, they will miss relevant and potentially vital information. Region 1 and 2 in the figure correspond to what a digital forensic investigator typically seeks – relevant artifacts that reside on the seized devices originating from: (i) programs and services running on the local devices; and (ii) the web, such as files cached by a web browser or email client. Region 3 corresponds to relevant data that the suspect has stored on the web, but the data cannot be retrieved directly from the seized devices. Everything outside the top and right circles represents non-digital evidence.

Modern cyber crimes present challenges that traditional digital forensic techniques are unable to address. This chapter identifies five unique challenges that web environments pose to digital forensic investigations: (i) complying with the rule of completeness (C0); (ii) associating a suspect with online personas (C1); (iii) gaining access to the evidence stored

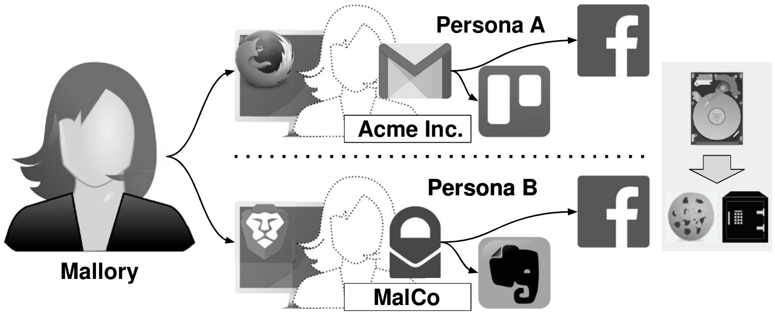

Figure 2. Motivating scenario.

online (C2); (iv) giving the evidence relevant context in terms of content and time (C3); and (v) integrating forensic tools to perform advanced analyses (C4). Currently, forensic investigators have no strategy or framework to guide them in their analysis of cases involving devices and users, where the evidentiary data is dispersed on local devices and on the web.

This chapter proposes a framework designed for conducting analyses in web environments that addresses challenges C0 through C4. The framework, which is readily integrated into existing workflows, enables a digital forensic investigator to obtain and give relevant context to previously-unknown data while adhering to the rules of evidence.

2. Motivating Scenario

Figure 2 presents a motivating scenario. Mallory, an employee at Acme Inc., is using company resources to start a new business, MalCo. This action is a violation of Acme's waste, fraud and abuse policies, as well as the non-compete agreement that she has signed. Mallory knows that eventually her computer may be analyzed by the IT department for evidence of her actions to provide grounds for Acme claiming ownership of MalCo after it is launched. Therefore, she uses various web services whenever she works on her new company to minimize the evidence left on her computer at Acme.

Mallory conscientiously segregates her web browsing between the work she does for Acme and what she does for MalCo, even using different web browsers. This segregation effectively creates two personas: (i) Persona A (Acme); and (ii) Persona B (MalCo).

When Mallory takes on Persona A, she uses Firefox as her web browser. Because Acme uses Google's G Suite, her work email is essentially a Gmail address. Mallory's team at Acme uses Trello to coordinate their activities and Facebook to engage with their clients. She used the Gmail address to create her accounts on Trello and Facebook.

When Mallory assumes Persona B to work on MalCo, she is careful to only use the Brave web browser. For her MalCo-related email, she created an account with Proton Mail because of its extra encryption features. She used her Proton Mail address to create accounts on Evernote and Facebook. In Evernote, Mallory stores all her MalCo business plans, client lists and product information. Using her Persona B Facebook account, Mallory has secretly contacted Acme customers to gauge their interest in switching to MalCo after it launches.

3. Unique Forensic Challenges

This section discusses the five principal challenges that web environments pose to digital forensic investigations. For convenience, the five challenges are numbered C0 through C4.

3.1 Rule of Completeness (C0)

The rules of evidence protect victims and suspects by helping ensure that the conclusions drawn from the evidence by forensic investigators are accurate. The completeness rule states that evidence must provide a complete narrative of a set of circumstances, setting the context for the events being examined to avoid "any confusion or wrongful impression" [14]. Under this rule, if an adverse party feels that the evidence lacks completeness, it may require the introduction of additional evidence "to be considered contemporaneously with the [evidence] originally introduced" [14].

The rule of completeness relates closely to the other challenges discussed in this section, which is why it is numbered C0. By attempting to associate a suspect with an online persona (C1), an investigator increases the completeness of the evidence. The same is true when an investigator gains access to evidence stored on the web (C2).

The rule of completeness can be viewed as the counterpart to relevant context (C3). By properly giving context to evidence, an investigator can ensure that the evidence provides the "complete narrative" that is required. However, during the process of giving the evidence context, the investigator must take care not to omit evidence that would prevent confusion or wrongful impression.

3.2 Associating Online Personas (C1)

When an individual signs up for an account with an online service provider, a new persona is created that, to some degree, represents who the individual is in the real world. The degree to which the persona accurately represents the account owner depends on a number of factors. Some attributes captured by the service provider (e.g., customer identification number) may not correlate with real-world attributes. Also, a user may provide fraudulent personal information, or may create parody, prank, evil-twin, shill, bot or Sybil accounts.

The challenge faced by a forensic investigator is to associate a persona with an individual in order to assign responsibility to the individual for the actions known to have been performed by the persona. If an investigator is unable to establish this link, then the perpetrator effectively remains anonymous.

In addition to being difficult to make an explicit link to an individual, it is also difficult to discover personas in the first place, especially if the forensic investigator only (at least initially) has access to data from the devices that were in the suspect's possession. This difficulty arises because web environments tend to store very little (if any) data on a user's local devices that may reveal a persona.

In Mallory's case, the data left behind that could reveal her personas resides in browser cookies and her password vault. After determining the online services associated with these credentials, the investigator still must find a way to show that it was actually Mallory who created and used the accounts. This is a more difficult task when many users share the same computer.

3.3 Evidence Access (C2)

An investigator could determine that a service provider would be likely to have additional data created or stored by the suspect. In this case, the typical course of action is to subpoena the service provider for the data. However, this option is available only to law enforcement and government agencies. If an investigation does not merit civil or criminal proceedings, corporate and non-government entities are essentially left to collect whatever evidence they can on their own.

While many web services provide APIs for programs to access data, no unified API is available to access data from multiple web services nor should such an API exist. Since web services are so disparate, a unique acquisition approach has to be developed for each web service. Moreover, because there is no guarantee that APIs will remain constant, it may be necessary to revise an approach every time the service or its API change.

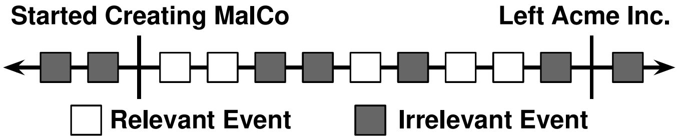

Figure 3. Timeline of Mallory's actions.

3.4 Relevant Context (C3)

The objective of a digital forensic investigator is to distill evidence down to the artifacts that tell the story of what happened during an incident by increasing the relevance of the contexts of artifacts. A context comes in two forms, both of which are critical to an investigation.

The first form is thematic context, which effectively places labels on artifacts that indicate their subjects or themes. An investigator uses the labels to filter out artifacts that are not relevant to the investigation, thereby focusing on artifacts that help prove or disprove the suspect's involvement in the incident. A common tool for thematic context is a keyword search, in which the investigator enters some keywords and the tool searches the file content and returns instances that match the provided text or related text (if the tool uses a fuzzy-matching algorithm).

The second form of context is temporal context, which places an artifact in a timeline to indicate its chronological ordering relative to events in the non-digital world as well as other digital artifacts. Creating a timeline provides an investigator with a perspective of what happened and when, which may be critical to the outcome of the investigation.

Although these forms of context have always been important objectives for digital forensic investigators, web environments make it much more difficult to create contexts because web users can generate artifacts and events at a higher pace than traditional evidence. Furthermore, the web has diverse types of data, such as multimedia, many of which require human effort or very sophisticated software to assign subjects to the data before any thematic context can be determined.

Figure 3 shows Mallory's actions in creating MalCo. Identifying the relevant events from the irrelevant events provides thematic context. Temporal context is provided to events by placing them in chronological order and creating a window of interest by determining the points at which Mallory engaged in inappropriate behavior.

3.5 Tool Integration (C4)

Researchers have long decried the shortcomings of the two types of tools that are available to digital forensic investigators. The first type,

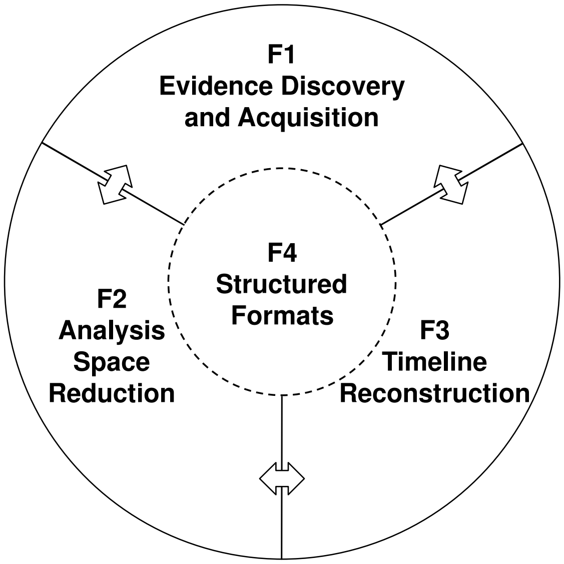

Figure 4. Web environment forensics framework.

one-off tools, are usually designed to perform very specific actions or analyses; they may not have very good technical support or may be outdated, poorly documented or have other issues. The second type, monolithic tools, seek to cover as many use cases as possible in a single package. While these tools often enjoy the benefits of commercial software, their vendors have an obvious interest in keeping the details about the tools and underlying techniques proprietary to maintain a competitive edge. Also, monolithic tools often do not support scripting, automation and importing/exporting data from/to other tools [5, 12, 33].

Given the complexity of the situation, it is unreasonable to expect a single tool or technique to address the challenges that hinder web environment forensics. Therefore, it is clear that forensic tools designed to properly accommodate evidence from web environments will have to overcome the status quo and integrate with other tools to accomplish their work.

4. Web Environment Forensics Framework

Figure 4 presents the proposed web environment forensics framework. It incorporates four components that are designed to directly address the challenges discussed in Section 3. The four components are: (i) evidence discovery and acquisition (F1); (ii) analysis space reduction (F2); (iii)

Table 1. Challenges addresed by the framework components.

	F1	F2	F3	F4
Rule of Completeness (CO)	✓	✓	✓	–
Associating Personas (C1)	✓	–	–	–
Evidence Access (C2)	✓	–	–	–
Relevant Context (C3)	–	✓	✓	–
Tool Integration (C4)	–	–	–	✓

timeline reconstruction (F3); and (iv) structured formats (F4). The components provide a digital forensic investigator with: (i) previously-unknown data related to an incident; and (ii) the relevant context of the incident.

Table 1 identifies the challenges addressed by the four components. Components F1, F2 and F3 interrelate with each other non-sequentially, meaning that the sequence in which an investigator could use the components is not dictated by the components themselves, but by the flow of the investigation and the investigator's needs. In fact, after an investigator completes one component, he may subsequently need one, both or neither of the other two components. However, as will be discussed later, component F4 relates to the components F1, F2 and F3 in a special way.

The non-sequential relationships between components F1, F2 and F3 enable an investigator to incorporate the components into an existing workflow as needed and in a manner that befits the investigation. For example, after acquiring new evidence from the web, it may be necessary to narrow the focus of the investigation, which, in turn, may tell the investigator where to find new, previously-inaccessible evidence, thus creating the sequence F1 → F2 → F1. Similarly, an investigator may use acquired data to reconstruct a timeline of events, which may be most useful after it is reduced to the periods of heightened activity. With a focus on these events, it may then become necessary to create a timeline of even finer granularity or to acquire new evidence specific to the period of interest. The sequence of these steps is F1 → F3 → F2 → F3.

The remainder of this section describes the objectives of each component in the framework, the investigator's process for fulfilling the component, the research challenges that impede progress on the component, related approaches and key research opportunities for the component.

4.1 Evidence Discovery and Acquisition (F1)

The objective of framework component F1 is to overcome the challenges involved in: (i) establishing associations between a suspect and online personas (C1); and (ii) gaining access to the evidence stored in web services by the personas (C2). It is important to note that component F1 does not attempt to discern whether or not the data is relevant to the investigation. Instead, the focus is to discover and acquire web-based evidence created by the suspect, but not stored on the seized devices; this is evidence that would not otherwise be accessible to the investigator. Of course, component F1 also helps an investigator comply with the rule of completeness (C0).

Investigator Process (F1). The investigator's process for fulfilling component F1 comprises two actions: (i) discovery; and (ii) acquisition.

- **Discovery:** In order to discover previously-inaccessible evidence, an investigator has to analyze the storage of the devices in custody for clues that connect the user to evidence stored on the web. Example clues include web session cookies, authentication credentials and program-specific artifacts such as those collected by the community and posted at `ForensicArtifacts.com`. Finding and identifying these artifacts requires a sound understanding of their critical characteristics and, in some cases, a database of artifact samples to facilitate efficient comparison.

  In the case of authentication artifacts with certain formats, the process of discovery can be automated, relieving an investigator from attempting manual discovery, which does not scale well. However, even with automated discovery, it may be necessary for the investigator to manually determine the service to which a credential gives access. For example, if a user stores the username and password in a text file, even if the artifact has the structure that enables a program to accurately extract the credentials, it may require a human to consider the context of the file (name of the directory or file) in order to derive the corresponding service.

- **Acquisition:** After the investigator discovers an authentication artifact and identifies the corresponding service, it is necessary to devise a means to acquire data from the service. Given the variety of web services, an approach for acquiring data from one source may not apply directly to other sources. Investigators and tool developers need to understand which principles are transferable and design their workflows and tools to be as general-purpose as possible [26]. They should also leverage structured storage formats (F4) for the acquired evidence.

Challenges (F1). The discovery and acquisition actions of component F1 face unique challenges:

- **Discovery:** The task of discovering evidence in the web has some challenges. First, the volume of data a suspect can store on the web is nearly unlimited. Not only does this present a challenge in terms of storage requirements for holding the evidence, but it also makes the task of analyzing it more complex.

  Second, the boundaries of the data set are nebulous in a geographical sense as well as in terms of the service that maintains the data. In contrast, the boundaries of hard drive storage (e.g., total number of sectors) are well-defined and an investigator can identify the boundaries easily via simple analysis of the disk media. However, it is difficult for an investigator to find a starting point for discovering evidence in a web environment. In contrast, any investigator knows that the best place to start analyzing a desktop computer is its hard drive. The best analog for evidence in the web is for the investigator to start with what can be accessed, which, in most instances, is a device with storage, such as a smart phone, computer, laptop, GPS device or DVR. However, it is also possible that the devices possessed by a suspect contain no information about where their data is stored on the web.

  A third challenge occurs when a suspect has many accounts on the web – accounts with multiple web services and multiple accounts with a single service. While it is possible that all the user accounts are active and accessible, it is more likely that some accounts have been suspended or deactivated due to inactivity, intentional lockout, unsuccessful authentication attempts or other circumstances. Furthermore, with so many web services and accounts, it is not uncommon for an individual to forget that an account was created with a particular web service months or years after the fact. It is unlikely that the data from an inactive or forgotten account would play a critical role in an investigation, but this illustrates the challenge of discovering all the data created by a user on the web. The existence of a large number of user accounts also makes it more difficult to evaluate their relevance, although this challenge relates more directly to component F2.

- **Acquisition:** Acquiring data presents its own set of challenges. First, the data stored by web services changes continually. This is especially true when the data is automatically generated on behalf of a user. With the continued proliferation of Internet of Things

> devices, forensic investigators are likely to see an ever-increasing amount of automatically generated data for the foreseeable future. Such data is not dissimilar to evidence that requires live acquisition, but it may be more fragile and require special care and handling.
>
> The other key challenge to acquiring data from a service provider involves actually accessing the data (discussed in Section 3.3). Since a unified API is not available for acquiring data from web services, considerable manual effort is required on the part of an investigator to understand and interface with each service.

Related Approaches (F1). Very few approaches are currently available to an investigator to complete component F1 of the framework, and even fewer are automated [22]. Dykstra and Sherman [11] have evaluated the efficacy of forensic tools in acquiring evidence from an Amazon EC2 instance. In general, the tools did well considering they were not designed for this type of evidence acquisition. However, the approach only works for instances under the control of the investigator at the guest operating system, virtualization and host operating system layers, not at the web application layer.

Research Opportunities (F1). Artifact repositories such as `ForensicArtifacts.com` and `ForensicsWiki.org` are valuable resources for forensic investigators. However, a critical shortcoming is that the information they contain is only suitable for human consumption, meaning that it is not currently possible for automated tools to leverage the data hosted on these sites. Future research should focus on converting the information to a structured format (F4) with the necessary semantics to facilitate automation.

Although each web service has its own set of APIs, it may be possible, through a rigorous study of a wide range of services, to create an abstraction of the various calls and create a generic and reusable method that facilitates acquisition.

4.2 Analysis Space Reduction (F2)

Not every evidence artifact is equally important to an investigation. Investigators would greatly benefit from assistance in identifying and focusing on the most relevant artifacts (C3). When irrelevant data is filtered in a triaging process, an investigator can save time and effort in completing the analysis – this is the motivation and the objective of component F2.

Although component F2 removes evidence from view, the process helps an investigator comply with the rule of completeness (C0). This is because the narrative of the evidence is unfettered by irrelevant artifacts.

While analysis space reduction through improved thematic context can benefit forensic analyses of digital evidence of all types, due to the virtually limitless storage capacity, analyses of evidence from web environments stand to gain particular performance improvements from the incorporation of component F2.

Investigator Process (F2). There are two general approaches to reducing the analysis space of evidence: (i) classification; and (ii) identification.

- **Classification:** This approach involves the categorization of evidentiary data and indicating the types of data that are of interest and are not of interest. Classification is the more common form of thematic context and aligns well with the example provided in Section 3.4. Forensic investigators may also wish to classify or separate artifacts according to when they were created, modified or last accessed, in which case, techniques from component F3 would be helpful.

- **Identification:** This approach reduces the analysis space by determining what exactly comprises the evidence; this is especially important when evidence is encrypted or otherwise unreadable directly from device storage. The primary task is more about identifying the data rather than classifying it or determining its relevance. Nevertheless, identification is still a method for providing thematic context because it enables an investigator to determine if the data is relevant to the investigation or not. The main difference is that, instead of identifying the subject of the data directly, the investigator determines the subject from the identity of the data.

  One method to reduce the analysis space via identification is to use information about data (i.e., metadata) to eliminate what the data cannot be, incrementally approaching an identification via true negatives. This approach is applicable only when the set of possibilities is limited (i.e., the approach does not apply to arbitrary files created by a user).

Because the ultimate goal of component F2 is to end up with less (but more relevant) evidence than the original data set, F2 tools may export their results in the same format as the data input to the tools. This provides the benefit that F2 tools can be incorporated in existing workflows without having to change how other tools process data. However, even in cases where the reduction of the analysis space yields data of a different type than the input (e.g., via the identification approach), tools should still use structured formats for the reasons discussed in Section 4.4.

Challenges (F2). Reducing the analysis space in an automated manner requires the careful consideration of a number of factors. First, the implication here is that an algorithm is given the responsibility to understand the nature of the evidence and make a judgment (albeit a preliminary one) concerning its bearing on an individual's guilt or innocence. While false positives reduce the analysis space in a sub-optimal manner, a false negative obscures a relevant artifact from the investigator's view and could alter the outcome of the investigation, which is, of course, unacceptable.

Exculpatory evidence, which suggests innocence, is particularly sensitive to false negatives because it is inherently more difficult to identify than inculpatory evidence, which, by definition, tends to suggest guilt. In other words, evidence that exonerates a suspect is more difficult to interpret in an automated fashion because it may not directly relate to the incident under investigation, it may require correlation with evidence from other sources or it may be the absence of evidence that is of significance.

In addition to the challenges related to the accuracy of the tools that reduce the analysis space, it is also important to consider the fact that the volume of data stored by a suspect on the web may be very large. Even after an accurate reduction to relevant data, the size of the resulting data set may still be quite large and time-consuming for analysis by a human investigator.

Related Approaches (F2). Researchers have developed several data classification techniques such as object recognition in images and topic identification of documents [16]. Another classification example is the National Software Reference Library (NSRL) [20], which lists known files from benign programs and operating systems. By leveraging the National Software Reference Library to classify evidence that is not of interest, an investigator can reduce the analysis space by eliminating from consideration files that were not created by the user and, thus, do not pertain to the investigation.

Research Opportunities (F2). Perhaps the most important potential research topic related to reducing analysis space is developing methods to minimize false positives without risking false negatives. Such an undertaking would clearly benefit from advances in natural language processing, computer vision and other artificial intelligence domains. The better a tool can understand the meaning of digital evidence, the more likely it would accurately minimize false negatives.

Because people regularly use multiple devices on a typical day, the evidence they leave behind is not contained on a single device. Forensic investigators would benefit greatly from improved cross-analytic techniques that combine the evidence from multiple sources to help correlate artifacts and identify themes that otherwise would have been obscured if each source had been analyzed individually.

Researchers have already demonstrated that it is possible to identify encrypted data without decrypting it [15, 24]. Although such approaches may not be well-suited to every investigation involving encrypted data, the fact that it is possible under the proper circumstances demonstrates there are research opportunities in this area.

4.3 Timeline Reconstruction (F3)

The objective of framework component F3 is to improve the temporal context of the evidence by reconstructing the incident timeline, giving the artifacts a chronological ordering relative to other events. This timeline, in turn, helps tell a more complete story of user activities and the incident under investigation. The additional information also contributes to a more complete narrative, helping satisfy the rule of completeness (C0).

Investigator Process (F3). The first step in reconstructing a timeline from web environment data is to collect all available evidence that records values of time in connection with other data. This task requires F1 tools and methods. Accordingly, all the challenges and approaches discussed in Section 4.1 apply here as well.

All the collected timeline information should be combined into a single archive or database, which would require a unified storage format (F4) that accommodates the various fields and types of data included in the original information. However, because the information originates from several sources, the compiled timeline may include entries that are not relevant to the investigation. In this case, it would be beneficial to leverage component F2 approaches to remove entries that do not provide meaningful or relevant information, thereby improving the thematic context of the evidence. Similarly, if a particular time frame is of significance to an investigation, removing events that fall outside the window would improve the temporal context of the evidence.

After the timeline information has been compiled and filtered, it is necessary to establish the relationships between entries. Establishing the sequence of events is a simple matter if everything is ordered chronologically. Other types of relationships that may prove insightful include event correlations (e.g., event a always precedes events b and c) and clustering (e.g., event x always occurs close to the time that events y and z occur). Finally, an investigator may leverage existing analysis and visualization tools on the timeline data, assuming, of course, that they are compatible with the chosen storage format.

Challenges (F3). The analysis of traditional digital evidence for timeline information is well-researched [3, 8, 19, 28]. However, current approaches may not be directly applicable to web environments due to the inherent differences. For example, timeline reconstruction typically incorporates file metadata as a source of time data and previous work has demonstrated that web service providers regularly store custom metadata fields [26]. These metadata fields are typically a superset of the well-known modification, access and creation (MAC) times, and include cryptographic hashes, email addresses of users with access to the file, revision history, etc. Clearly, these fields would be valuable to investigators, but they are not accommodated by current timeline tools.

For forensic tool developers to incorporate these fields into their tools, they would have to overcome some additional challenges. Since web service providers use different sets of metadata fields, it would be critical to devise a method that supports diverse sets of fields. One approach is to create a structured storage format (F4) with the flexibility to store arbitrary metadata fields. Another approach is to unify the sets of metadata fields using an ontology such that the metadata semantics are preserved when combining them with fields from other sources.

Another challenge to incorporating metadata from web environments in timeline reconstruction is that the variety of log types and formats grows as new devices emerge (e.g., Internet of Things devices). Many of these devices perform actions on behalf of their users and may interface with arbitrary web services via the addition of user-created “skills” or apps. Forensic researchers have only recently begun to evaluate the forensic data stored on these devices [13, 23].

Finally, as with any attempt to reconcile time information from different sources, it is critical to handle differences in time zones. While it is a common practice for web services to store all time information in the UTC mode, investigators and tools cannot assume that this will always be the case. Reitz [25] has shown that correlating data from different time zones can be a complicated task.

Related Approaches (F3). As mentioned above, it is uncertain if current approaches to timeline reconstruction would assist investigators with regard to evidence from web environments; in fact, the research literature does not yet contain any approaches designed for this purpose. However, because the first step in timeline reconstruction is to collect data with time information, some cloud log forensic approaches may provide good starting points.

Marty [17] presents a logging framework for cloud applications. This framework provides guidelines for what is to be logged and when, but it requires application developers to be responsible for the implementations. As such, this approach may complement other logging methods, but it may not be directly applicable to web environments.

Research Opportunities (F3). The visualization of timeline data is an active research area [8, 21, 29, 31] and there will always be new and better ways to visualize timeline data. For example, virtual and augmented reality technologies may help forensic investigators to better understand data by presenting three-dimensional views of timelines.

One unexplored aspect of timeline reconstruction is the standardization of the storage format (F4) of the data that represents timelines. Separating the data from the tool would facilitate objective comparisons of visualization tools and enable investigators to change tools without having to restart the timeline reconstruction process from scratch.

When reconstructing a timeline from multiple sources, there is always the chance that a subset of the time data will correspond to an unspecified time zone. A worthwhile research topic is to develop an approach that elegantly resolves such ambiguities.

4.4 Structured Formats (F4)

Structured formats provide a means for storing information that is not specific to a single tool or process, thereby facilitating interoperability and integration (C4). Structured formats also enable comparisons of the outputs of similar tools to measure their consistency and accuracy, which are key measurements of the suitability of forensic tools with regard to evidence processing.

Component F4 is positioned at the center of the framework because components F1, F2 and F3 all leverage some type of storage format, even if the format itself is not a part of each component. For example, after discovering and acquiring new evidence, a tool must store the evidence in some manner; clearly, the format in which the evidence is stored should have a generic, yet well-defined, structure. The structure used by the acquisition tool does not change how it performs its principal task, but it is a peripheral aspect of its operation.

Structured formats are critical to the proper functioning of the proposed framework. In order for a tool that provides one component to communicate with another tool that provides a different component, the two tools must be able to exchange data that they can both understand. Defining a structured format for the data is what makes this possible.

Investigator Process (F4). Structured formats are intended to facilitate tool interoperability and integration. Therefore, a forensic investigator should rarely, if ever, have to work directly with structured formats.

Challenges (F4). In order to realize the benefits, a structured format must satisfy three conditions. First, it must precisely represent the original evidence without any loss of information during conversion. Second, there must be a way to verify that the data conforms to the format specifications. This leads to the third condition, which requires that the specifications must be published and accessible to tool developers.

While many storage formats exist, none of them is perfect or covers every use case. As in the case of software engineering projects, format designers are constantly faced with the need to compromise or make trade-offs; this, in turn, makes them less suitable for certain circumstances. For example, some storage formats fully accommodate the file metadata fields used by Windows filesystems, but not Unix filesystems. This illustrates how difficult it can be to incorporate the correct level of detail in a format specification [6]. In this regard, open-source formats have an advantage in that the community can help make improvements or suggest ways to minimize the negative effects of trade-offs.

It is critical to the proposed framework and to the principle of composability that analysis tools use structured formats to store their results in addition to the evidence itself. This is the only way to support advanced analyses that can handle large evidence datasets, such as those originating from the web.

Related Approaches (F4). Several structured formats have been proposed for digital forensic applications over the years, most of them designed to store hard disk images [10, 32]. This section summarizes some of principal structured formats.

The Advanced Forensic Format (AFF) [9] provides a flexible means for storing multiple types of digital forensic evidence. The developers, Cohen et al., note in their paper that "[unlike the Expert Witness Forensic (EWF) file format], AFF [employs] a system to store arbitrary name/ value pairs for metadata, using the same system for both user-specified metadata and for system metadata, such as sector size and device serial number."

The Cyber Observable Expression (CybOX) [18] language was designed "for specifying, capturing, characterizing or communicating ... cyber observables." Casey et al. [6] mention in their work on the Digital Forensic Analysis Expression (DFAX) that CybOX can be extended to represent additional data related to digital forensic investigations. The CybOX Project has since been folded into version 2.0 of the Structured Threat Information Expression (STIX) specification [1].

The Cyber-Investigation Analysis Standard Expression (CASE) [7], which is a profile of the Unified Cybersecurity Ontology (UCO) [30], is a structured format that has evolved from CybOX and DFAX. CASE describes the relationships between digital evidence artifacts; it is an ontology and, therefore, facilitates reasoning. Because CASE is extensible, it is a strong candidate for representing evidence from web environments.

Digital Forensics XML (DFXML) [12] is designed to store file metadata, the idea being that a concise representation of metadata would enable investigators to perform evidence analyses while facilitating remote collaboration by virtue of DFXML's smaller size. Because it is written in XML, other schemas can extend DFXML to suit various scenarios.

Email Forensics XML (EFXML) [22] was designed to store email evidence in a manner similar to DFXML. Instead of storing email in its entirety, EFXML only stores the metadata (i.e., headers) of all the email in a dataset. EFXML was designed to accommodate email evidence originating from traditional devices as well as from the web.

The Matching Extension Ranking List (MERL) [15] is more specialized than the formats discussed above. Instead of storing evidence, MERL files store the analysis results from identifying extensions installed on an encrypted web thin client such as a Chromebook. MERL does not have the flexibility to store other kinds of data. However, unlike many of the other formats, it was created specifically for web-based evidence.

Of course, none of the formats were created to capture the diverse types of evidence in the web. However, some formats, such as AFF4, the latest version of AFF, may provide enough flexibility to store arbitrary types of web-based evidence.

Research Opportunities (F4). Buchholz and Spafford [4] have evaluated the role of filesystem metadata in digital forensics and have proposed new metadata fields that would assist in forensic examinations. A similar study conducted for web-based evidence would be of great use to the digital forensics community. As discussed above, each web service has its own custom metadata that serves its purposes, but it is important to understand how the metadata differ from service to service, which ones have value to investigators, how to unify (or at least reconcile the semantics of) the various fields and which fields would be useful if digital forensic investigators were able to make suggestions.

5. Related Work

The previous sections have discussed many approaches related to the individual components of the proposed framework. This section examines key approaches that relate to the framework as a whole.

Paglierani et al. [22] have developed a framework for automatically discovering, identifying and reusing credentials for web email to facilitate the acquisition of email evidence. Although their approach directly addresses the objectives of discovering and acquiring web evidence (F1) and provides a concise, structured format for storing email evidence (F4), their framework is tailored too closely to web email to be applied to web environments in general.

Ruan et al. [27] have enumerated several forensic challenges and opportunities related to cloud computing in a manner similar to what has been done in this research in the context of web environments. However, Ruan et al. do not provide a guide that could assist investigators in using the cloud for forensic examinations; instead, they only highlight the potential benefits of doing so. Additionally, although much of the modern web is built on cloud computing, the two are not synonymous. As such, many of the challenges listed by Ruan and colleagues, such as data collection difficulties, services depending on other services and blurred jurisdictions, apply to web environments, but the opportunities, such as providing forensics as a service, apply mainly to implementing forensic services in the cloud.

Birk and Wegener [2] provide recommendations for cloud forensics, separated by the type of cloud service provider, infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). Of these the most applicable to web environments is, of course, software as a service. However, Birk and Wegener place the responsibility for providing the means of forensic acquisition on cloud service providers. In contrast, the framework proposed in this chapter assists digital forensic investigators in understanding what they can accomplish even with uncooperative cloud service providers.

6. Conclusions

Conducting digital forensic analyses of web environments is difficult for investigators because of the need to comply with the rule of completeness, associate suspects with online personas, gain access to evidence, give the evidence relevant contexts and integrate tools. The framework presented in this chapter mitigates these challenges, guiding digital forensic investigators in processing web-based evidence using their existing workflows. Web environments provide exciting challenges to digital forensics and the area is ripe for research and innovation.

Acknowledgement

This research was partially supported by the DoD Information Assurance Scholarship Program and by the Center for Cybersecurity and Digital Forensics at Arizona State University.

References

[1] S. Barnum, Standardizing Cyber Threat Intelligence Information with the Structured Threat Information Expression (STIX), Technical Report, MITRE Corporation, Bedford, Massachusetts, 2014.

[2] D. Birk and C. Wegener, Technical issues of forensic investigations in cloud computing environments, *Proceedings of the Sixth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, 2011.

[3] F. Buchholz and C. Falk, Design and implementation of Zeitline: A forensic timeline editor, *Proceedings of the Digital Forensics Research Workshop*, 2005.

[4] F. Buchholz and E. Spafford, On the role of file system metadata in digital forensics, *Digital Investigation*, vol. 1(4), pp. 298–309, 2004.

[5] A. Case, A. Cristina, L. Marziale, G. Richard and V. Roussev, FACE: Automated digital evidence discovery and correlation, *Digital Investigation*, vol. 5(S), pp. S65–S75, 2008.

[6] E. Casey, G. Back and S. Barnum, Leveraging CybOX to standardize representation and exchange of digital forensic information, *Digital Investigation*, vol. 12(S1), pp. S102–S110, 2015.

[7] E. Casey, S. Barnum, R. Griffith, J. Snyder, H. van Beek and A. Nelson, Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language, *Digital Investigation*, vol. 22, pp. 14–45, 2017.

[8] Y. Chabot, A. Bertaux, C. Nicolle and T. Kechadi, A complete formalized knowledge representation model for advanced digital forensics timeline analysis, *Digital Investigation*, vol. 11(S2), pp. S95–S105, 2014.

[9] M. Cohen, S. Garfinkel and B. Schatz, Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow, *Digital Investigation*, vol. 6(S), pp. S57–S68, 2009.

[10] Common Digital Evidence Storage Format Working Group, Survey of Disk Image Storage Formats, Version 1.0, Digital Forensic Research Workshop (`www.dfrws.org/sites/default/files/survey-dfrws-cdesf-diskimg-01.pdf`), 2006.

[11] J. Dykstra and A. Sherman, Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust and techniques, *Digital Investigation*, vol. 9(S), pp. S90–S98, 2012.

[12] S. Garfinkel, Digital forensics XML and the DFXML toolset, *Digital Investigation*, vol. 8(3-4), pp. 161–174, 2012.

[13] J. Hyde and B. Moran, Alexa, are you Skynet? presented at the *SANS Digital Forensics and Incident Response Summit*, 2017.

[14] Legal Information Institute, Doctrine of completeness, in *Wex Legal Dictionary/Encyclopedia*, Cornell University Law School, Ithaca, New York, 2018.

[15] M. Mabey, A. Doupé, Z. Zhao and G. Ahn, `dbling`: Identifying extensions installed on encrypted web thin clients, *Digital Investigation*, vol. 18(S), pp. S55–S65, 2016.

[16] F. Marturana and S. Tacconi, A machine-learning-based triage methodology for automated categorization of digital media, *Digital Investigation*, vol. 10(2), pp. 193–204, 2013.

[17] R. Marty, Cloud application logging for forensics, *Proceedings of the ACM Symposium on Applied Computing*, pp. 178–184, 2011.

[18] MITRE Corporation, Cyber Observable Expression (CybOX) Archive Website, Bedford, Massachusetts (`cybox.mitre.org`), 2017.

[19] S. Murtuza, R. Verma, J. Govindaraj and G. Gupta, A tool for extracting static and volatile forensic artifacts of Windows 8.x apps, in *Advances in Digital Forensics XI*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 305–320, 2015.

[20] National Institute of Standards and Technology, National Software Reference Library (NSRL), Gaithersburg, Maryland (`www.nist.gov/software-quality-group/national-software-reference-library-nsrl`), 2018.

[21] J. Olsson and M. Boldt, Computer forensic timeline visualization tool, *Digital Investigation*, vol. 6(S), pp. S78–S87, 2009.

[22] J. Paglierani, M. Mabey and G. Ahn, Towards comprehensive and collaborative forensics on email evidence, *Proceedings of the Ninth International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 11–20, 2013.

[23] J. Rajewski, Internet of Things forensics, presented at the *Endpoint Security, Forensics and eDiscovery Conference*, 2017.

[24] A. Reed and M. Kranch, Identifying HTTPS-protected Netflix videos in real-time, *Proceedings of the Seventh ACM Conference on Data and Application Security and Privacy*, pp. 361–368, 2017.

[25] K. Reitz, Maya: Datetimes for Humans (`github.com/kennethreitz/maya`), 2018.

[26] V. Roussev, A. Barreto and I. Ahmed, API-based forensic acquisition of cloud drives, in *Advances in Digital Forensics XII*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 213–235, 2016.

[27] K. Ruan, J. Carthy, T. Kechadi and M. Crosbie, Cloud forensics, in *Advances in Digital Forensics VII*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 35–46, 2011.

[28] B. Schneier and J. Kelsey, Secure audit logs to support computer forensics, *ACM Transactions on Information and System Security*, vol. 2(2), pp. 159–176, 1999.

[29] J. Stadlinger and A. Dewald, Email Communication Visualization in (Forensic) Incident Analysis, ENRW Whitepaper 59, Enno Rey Netzwerke, Heidelberg, Germany, 2017.

[30] Z. Syed, A. Padia, T. Finin, L. Mathews and A. Joshi, UCO: A unified cybersecurity ontology, *Proceedings of the Workshop on Artificial Intelligence for Cyber Security at the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 195–202, 2016.

[31] C. Tassone, B. Martini and K. Choo, Forensic visualization: Survey and future research directions, in *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, K. Choo and A. Dehghantanha (Eds.), Elsevier, Cambridge, Massachusetts, pp. 163–184, 2017.

[32] S. Vandeven, Forensic Images: For Your Viewing Pleasure, InfoSec Reading Room, SANS Institute, Bethesda, Maryland, 2014.

[33] O. Vermaas, J. Simons and R. Meijer, Open computer forensic architecture as a way to process terabytes of forensic disk images, in *Open Source Software for Digital Forensics*, E. Huebner and S. Zanero (Eds.), Springer, Boston, Massachusetts, pp. 45–67, 2010.