Data Sources for Advancing Cyber Forensics: What the Social World Has to Offer

Ibrahim Baggili and Frank Breitinger
University of New Haven Cyber Forensics Research and Education Lab (www.UNHcFREG.com)
Department of Electrical & Computer Engineering and Computer Science
ibaggili@newhaven.edu, freitinger@newhaven.edu

Abstract

Cyber Forensics is fairly new as a scientific discipline and deals with the acquisition, authentication and analysis of digital evidence. One of the biggest challenges in this domain has thus far been real data sources that are available for experimentation. Only a few data sources exist at the time writing of this paper. The authors in this paper deliberate how social media data sources may impact future directions in cyber forensics, and describe how these data sources may be used as new digital forensic artifacts in future investigations. The authors also deliberate how the scientific community may leverage publically accessible social media data to advance the state of the art in Cyber Forensics.

Introduction

Years ago, most crimes had evidence that pertained to the physical world. Nowadays, digital evidence has become of paramount importance. Subsequently, forensic sciences extended their scope to include digital evidence, thus, a new domain was born – Cyber Forensics (CF)[1]. A major challenge in this field is coping with vast amounts of data during investigations now that the trend is that everything has become digital. For instance, books, photos, letters and Long Playing records (LPs) turned into e-books, digital photos, e-mails and mp3s. Additionally, we now have smartphones providing Internet access virtually everywhere, which in turn increased the daily usage of social media like Facebook or Twitter.

CF is a discipline that only started gaining notoriety in the scientific community over the last decade. The field has been ameliorating with only a small subset of computer scientists and institutions pushing the envelope in the domain. Although it is a relatively new field, the challenges and opportunities changed dramatically during that time.

Traditionally, the domain is viewed as a subset of Information Assurance (IA), and deals with Incidence Response (IR) as shown in Figure 1. Therefore, since the domain has focused on IR – it has traditionally been thought of as a post-mortem field – where evidence is collected only after an incident has occurred. In other words, the domain has dealt with the acquisition, authentication and analysis of digital evidence extracted from systems only after the incident has occurred (Casey 2011).

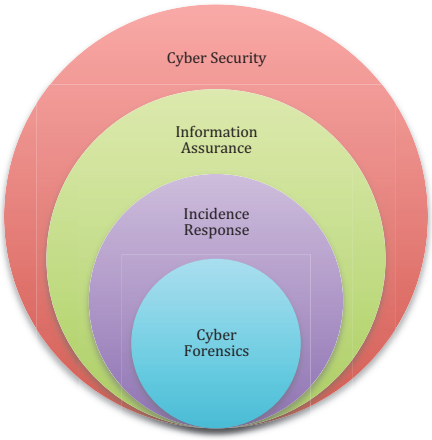

Figure 1. Domain Overview.

To date, most of the research efforts in CF have focused on ways of extracting data that may become weighty evidence in the court of law. However, ways of improving the digital forensics process has left the scientific community in awe due to the lack of data sources that are necessary towards advancing the domain – which (Garfinkel et al. 2009) attempted to answer with the digital corpora project.

Advancing the state of the art in the CF domain strongly depends on novel methods and algorithms that can help in

[1] The authors of this paper use the term Cyber Forensics to encompass a body of knowledge from various sub-disciplines such as a) Computer Forensics b) Network Forensics and c) Small Scale Digital Device Forensics d) Memory Forensics (Brinson, Robinson, and Rogers 2006);(Harril and Mislan 2007)

the identification of evidence during a case to speed up the forensics process.

Some scientific efforts have been pursued such as digital forensics triage in order to improve the overall digital forensics process. However, the field has some challenges to overcome in order to improve the state-of-the-art, which we discuss in the sections that follow.

The traditional cyber forensic process

As articulated, typically, in a CF investigation, data is acquired, then authenticated, and finally analyzed, as shown in Figure 2. In a traditional computer forensics investigation – a computer's hard drive is connected to a hardware write-blocker, and then forensically imaged (cloned) in order to create a bit-stream authentic copy of the hard drive. A hash value is computed for both the disk and the forensic copy, and the hashes are compared. If the hash values coincide, then the image is deemed authentic and is accepted because it has maintained its integrity. After this initial step, the analysis of the media commences to locate digital evidence relevant to the case at hand.

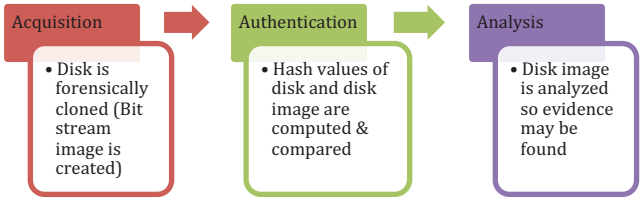

Figure 2. Traditional Computer Forensics Process.

In order to handle the vast amounts of data (e.g., common HDDs have 1 TB or even more), the forensic image is loaded in different forensic tools to analyze the data on the disk. More precisely, in a typical investigation, a forensic examiner has to inspect hundreds of thousands of files. Automatic ways of mitigating the amount of files that need to be examined have been devised in the community by known hash databases (Kent et al. 2006), and approximate matching techniques (Breitinger et al. 2014);(Breitinger et al. 2013). One can only imagine the amount of data that needs to be analyzed to locate weighty and relevant digital evidence needed to solve a case.

In most cases, disk images are created by corporate investigators, law enforcement officers or in many instances individuals that work for three-letter agencies. These disk images are typically then wiped after a case is completed – or stored for a short while in a secure facility depending on the policies of the organization conducting the investigation.

In digital investigations, current research focuses on technologies, algorithms or ideas that support the existing manual and labor-intensive forensic process. Since scientists do not have access to disk images from real cases it poses an interesting question to our scientific community: How can we learn from our past when we do not have real, accessible data to learn from?

Major challenges in CF

We present below a list of some of the major challenges in CF – but we would like to call the reader's attention to the first one:

1. The lack of real data sources
2. The young and ever changing nature of the field
3. The dependency on tools
4. The lack of published error rates for the various widely used digital forensics tools
5. The lack of basic research in this domain (cite our paper)
6. The lack of agreed upon standards and processes
7. The limitation of the hardware standards being used during the acquisition of data
8. The volatility of the evidence – such as RAM
9. The continuous change in technology
10. The use of anti forensics techniques and tools
11. The lack of a common body of knowledge

Although there are a number of issues outlined in various papers (Garfinkel 2010);(Rogers and Seigfreid 2004); (Ruan et al. 2013);(Baggili et al. 2013) - we will discuss one in particular that we believe the social domain can offer in improving the CF domain – which is the lack of real data sources to be leveraged for future research.

The lack of real data sources

Missing real data sources is a serious problem across different areas in computer science. Most agencies, vendors, providers (have to) keep their data secure and private. One cannot ignore the issue that a training set is needed in machine learning – and an appropriate training set has to come from real cases in CF. It is difficult to formulate and test any novel techniques and ideas on data that is fictitious - and although the Real Data Corpus (RDP) (Garfinkel et al. 2009) is useful for studying natural occurring phenomena on disks, it is not a corpus of real crimes that have occurred with their investigative outcomes – it is real data that exists on disks by random people.

We therefore contend in this paper that this is an issue that is quite difficult to solve, and only a few known research projects have attempted to learn from past real data to advance the state of the domain, or to improve the area of what some call push-button forensics, where a system tries to automatically analyze the data on a forensic image with little to no investigator interaction (McClelland, and

Marturana 2014);(Marturana et al. 2011); (Saleem, Baggili, and Popov 2014).

Social: What can it offer?

In the following sections – we explore what social domain has to offer to the CF domain. In the first section we discuss social media as a new source of digital forensic artifacts. In the second section, we explore the opportunities that social media has to offer in terms of public data sources for future research in CF.

Social media: New digital forensic artifacts

With the rise of social media applications on a multitude of platforms comes the potential for these applications to leave behind digital forensic artifacts that may be integral to an investigation. For example, research has shown how to extract Facebook chat logs from disks (Al Mutawa et al. 2011) and the vast amount of digital artifacts mobile social applications leave behind such as usernames, passwords, chat messages, posts, friends, location data and pictures just to name a few (Al Mutawa, Baggili, and Marrington, 2012);(Bader and Baggili 2010).

Digital forensic artifacts that could be extracted from social media applications are critical sources of digital evidence. For instance, in (Al Mutawa, Baggili, and Marrington 2012), the authors forensically examined the mobile applications Facebook, Twitter and MySpace on Blackberries, iPhones and Androids. Their results indicated that they were able to extract user and friend data, picture URLs, timestamps, comments posted, usernames and passwords in clear text (for MySpace only), photos uploaded, pictures viewed, posted tweets, device used to tweet and other digital forensic artifacts.

The artifacts we outline in this section are limited to data that may be recovered from a personal device during an investigation. Notwithstanding, the social world has more to offer in terms of public, close-to-real-time data. We will explore this opportunity in the following section.

Social media: New public data sources

The other non-intuitive source of data for CF research the social world has to offer is publically available. Publically available social media posts that include data such as geolocation, unstructured text and multimedia files are of critical importance for advancing the CF domain. We present in Table 1 some ideas of how these publically accessible data sources may be leveraged during a CF investigation.

Central to the notion of these data sources is the fact that they can be looked at as natural data sources. They are typically not fictitiously created, and they occur naturally in the world we live in. Also, central to these data sources is their accessibility when compared to real disk images as outlined in our earlier discussion. Lastly, another underpinning notion to all the data sources mentioned in Table 1 is that they can be mined in close-to-real-time. This means that investigators may now have the ability to monitor a suspect's actions in close-to-real-time as opposed to the long wait after the fact in order to investigate an incident.

Table 1. Social Media Data Sources for CF.

Social Data Source	Applicability to CF	Computing Discipline
Text Posts	Finding out who wrote a message (Author attribution)	Computational Linguistics
Friends / Groups	Finding out your network of friends	Social Network Analysis
Images	Facial recognition Object recognition	Image Processing
Text Posts	Personality profiling	Psychology + Computational Linguistics
Geolocation Data	Finding out the location of a suspect	Geographic Information Systems + Programming
Demographic information (Age / Gender / Books / Movies etc.)	Cyber profiling	Psychological / Criminal Profiling
Text Posts	Deception detection	Computational Linguistics
Text Posts (Self-reported actions)	Cyber incident spread Items purchased Current location Current activity Current mood Etc.	Computational Linguistics + Algorithms
Videos	Facial recognition Object recognition	Image Processing
Dates & Times	Event reconstruction	
Likes	Cyber profiling	Algorithms + Psychology
Language Used	Cyber profiling	Algorithms + Psychology

Discussion & Conclusion

We presented only a snapshot of the usage of social media data for CF purposes. We see a future in which CF embraces the idea of real-time intelligence, and not just post-mortem data. The notion of using social networks for solving crimes and catching criminals is of critical importance to the CF domain.

Furthermore, we are no longer limited to data sources that are difficult to obtain when using the traditional digital forensics investigative model. We now have access to data sources that are rich, new, and created by humans. The impact of these data sources should be explored in the scientific community in a multidisciplinary manner.

It is important to note that recently, a specific National Science Foundation (NSF) solicitation # NSF 15-005 was released to fund a small number of Early Concept Grants for Exploratory Research (EAGERs) under the Secure and Trustworthy Cyberspace (SaTC) program to encourage researchers to study how social data sources may be used in cyber security research. This is a strong indication for the need by the scientific community to explore research opportunities in which social media and cyber security intersect.

We would like to conclude our paper by encouraging the scientific community to leverage social media data sources in order to advance the state of the art in CF, and to collaborate in this domain. We believe that this is an untapped area of research that can foster collaborations and novel areas of discovery.

References

Casey, E. (2011). *Digital evidence and computer crime: forensic science, computers and the internet*. Academic press.

Brinson, A., Robinson, A., & Rogers, M. (2006). A cyber forensics ontology: Creating a new approach to studying cyber forensics. *digital investigation*, *3*, 37-43.

Harrill, D. C., & Mislan, R. P. (2007). A small scale digital device forensics ontology. *Small Scale Digital Device Forensics Journal*, *1*(1), 242.

Garfinkel, S., Farrell, P., Roussev, V., & Dinolt, G. (2009). Bringing science to digital forensics with standardized forensic corpora. *digital investigation*, *6*, S2-S11.

Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). Guide to integrating forensic techniques into incident response. *NIST Special Publication*, 800-86.

Breitinger, F., Guttman, B., McCarrin, M., & Roussev, V. (2014). Approximate matching: definition and terminology. *NIST Publication*.

Breitinger, F., Liu, H., Winter C., Baier, H., Rybalchenko A., Steinebach, M. (2013). Towards a process model for hash functions in digital forensics . *5th International Conference on Digital Forensics & Cyber Crime (ICDF2C).*

Garfinkel, S. L. (2010). Digital forensics research: The next 10 years. *Digital Investigation*, *7*, S64-S73.

Rogers, M. K., & Seigfried, K. (2004). The future of computer forensics: a needs analysis survey. *Computers & Security*, *23*(1), 12-16.

Ruan, K., Carthy, J., Kechadi, T., & Baggili, I. (2013). Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results. *Digital Investigation*, *10*(1), 34-43.

Baggili, I., BaAbdallah, A., Al-Safi, D., & Marrington, A. (2013). Research Trends in Digital Forensic Science: An Empirical Analysis of Published Research. In *Digital Forensics and Cyber Crime* (pp. 144-157). Springer Berlin Heidelberg.

McClelland, D., & Marturana, F. (2014, June). A Digital Forensics Triage methodology based on feature manipulation techniques. In *Communications Workshops (ICC), 2014 IEEE International Conference on* (pp. 676-681). IEEE.

Marturana, F., Me, G., Berte, R., & Tacconi, S. (2011, November). A quantitative approach to Triaging in Mobile Forensics. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on* (pp. 582-588). IEEE.

Saleem, S., Baggili, I., & Poppv, O. (2014). Quantifying Relevance of Mobile Digital Evidence As They Relate to Case Types: A Survey and a Guide for Best Practices. *Journal of Digital Forensics, Security and Law*, *9*(3), 19-50.

Al Mutawa, N., Al Awadhi, I., Baggili, I., & Marrington, A. (2011, December). Forensic artifacts of Facebook's instant messaging service. In *Internet Technology and Secured Transactions (ICITST), 2011 International Conference for* (pp. 771-776). IEEE.

Al Mutawa, N., Baggili, I., & Marrington, A. (2012). Forensic analysis of social networking applications on mobile devices. *Digital Investigation*, *9*, S24-S33.

Bader, M., & Baggili, I. (2010). iPhone 3GS forensics: logical analysis using apple itunes backup utility. *Small scale digital device forensics journal*, *4*(1), 1-15.