

# Forensic Analysis of Social Media Data : Research Challenges and Directions

Muhammad Firdaus (201956717)

Department of Information Security, Graduate School

Pukyong National University

e-mail: mfirdaus@pukyong.ac.kr

**Summary** The challenge to analyze data from social media sources not only for businesses and organizations but also for Law Enforcement Agencies. Social media offers various avenues for the collection and use of its data as evidence within a digital forensics investigation. The trail of digital information on social media, if explored correctly, can offer remarkable support in the criminal investigation. Social media evidence must be collected using careful, correct procedures and in a manner that ensures its integrity. Hence, social media evidence must be collected by a legally and scientifically appropriate forensic process and also coincide with the privacy rights of individuals. Following the legal process is a challenging task for legal practitioners and investigators to be able to carry out effective investigations and gather valid evidence efficiently. This paper explains the existing conditions of evidence acquisition, admissibility, and jurisdiction mechanisms in forensic social media. It also illustrates the direct challenges in gathering, analyzing, presenting, and validating evidence from social media data in the law enforcement process.

## 1. Introduction

Social media use has significantly increased in recent years [1]. Generally, the term social media is used to refer all the communication channels used for community-based interaction, collaboration, and content-sharing. These tools have affected the ways that individuals share and experience life. In July 2019, the total number of active social media users were 3.534 billion [2]. It means they are actively engaged in sharing their everyday activities on social media sites.

Social media evidence is a new frontier in digital forensics [3]. Since the use of social media is constantly increasing, data is being generated exponentially. And also, the exponential growth of social media has facilitated development of many serious cybercrime and malicious activities. Cybercriminals are constantly changing their strategies to target rapidly growing social media users. The misuse of social media in mobile devices may allow cybercriminals to utilize these services for malicious purposes such as spreading malicious codes, and obtaining and disseminating

confidential information [4]. According to an official survey conducted by The Office for National Statistics [5], there were an estimated 3.6 million cases of fraud and two million computer misuse offenses in a year. Although there is a variety of reasons for conducting cybercrimes, the motivation is often for financial gain. The fundamental issue associated with cybercrime consists of damage to reputation, monetary loss, in addition to impacts on the confidentiality, integrity, and availability of data [6].

Social media can also be used to facilitate malicious tasks. Criminal behavior and evidence of crimes can be found throughout social media. The data found on these websites and applications can imply intent when used in a criminal case. Many crimes can be somehow connected to social media in today's digital age. At earlier times, most crimes left breadcrumbs of evidence in the real world. Nowadays, through the interactive social media platforms, offenders engage in illicit practices such as fraud, cyber stalking, cyber bullying etc. [7]. Terrorists exploit social media to

reach audiences for potential recruits, disseminate messages and organize strategic operations.

Law enforcement agencies (LEAs) wish to take advantage of these information sources for the sake of security [7]. Social media can be used as a means of surveillance. By using digital forensics tools and techniques to review the information captured on social media, inferences can be made about a subject or an event [1]. For monitoring and analyzing criminal related activity in social media networks, one major question is to identify the most influential profiles, known also as "key player" discover. LEAs are also interested in answering the so called six W's: Who, What, When, Where, Why and How [7]. These questions are fundamental and are traditionally raised during criminal investigations.

This paper puts forward a review of the current state of research in the social media forensics. Section 2 explains the role of social media as evidence in legal proceedings and provide a review of current practices and tools used for the collection, preservation, and analysis of forensic data. This section also offers a discussion on privacy, jurisdiction and admissibility issues related to evidence collection, preservation, and presentation matters. Section 3 would provide a discussion of the limitations of current research and the challenging factors involved in the domain.

## 2. Evidence Collection

Social media evidence must be collected by a legally and scientifically appropriate forensic process and also coincide with the privacy rights of individuals. Following the legal process is a challenging task for legal practitioners and investigators to be able to carry out effective investigations and gather valid evidence efficiently.

### 2.1 Role and use of social media evidence in legal proceedings

Social media evidence provides an unlimited source of information about a potential suspect's or victim's profile that can be mined in

close-to-real-time. The contacts, messages, geolocation data, photos and generally their activities are offered in chronological order [7]. The metadata (information accompanying by content) and network data hold the adequate potential to assist in criminal investigations and to authenticate the evidence from online social networks (OSNs).

Presently, it is a legal requirement in a substantial number of serious crime investigations to seize and examine the digital devices, of victims and suspects. The data on these devices help to find traces of crime or history of digital activities, performed by the user. In general Arshad et al. [7] show the aspects of the role of social media evidence in Fig.1.

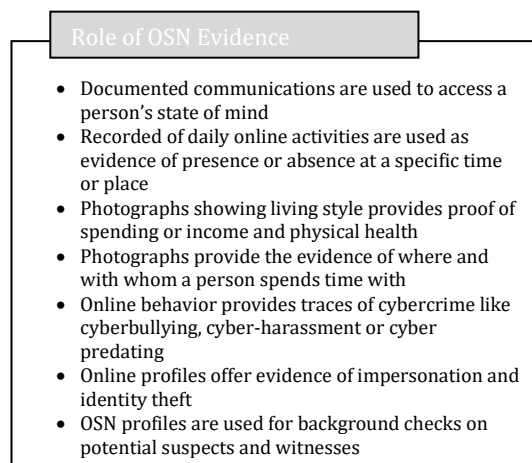


Fig. 1. Role of OSN content in Legal Proceedings [7].

The use of social media as evidence is quite common in criminal cases. Several criminal cases are now routinely investigated, prosecuted and defended through social media as evidence. Prosecution and defense lawyers equally utilize the information from OSNs in legal proceedings. However, defense lawyers face more hurdles to seek a subpoena to social media companies for accessing protected social media data.

The progression in digital communication is offering exceptional and diverse opportunities for the individuals. Tragically, the advancement is also providing endless provisions to the criminals

to commit offenses. They are using these online platforms to assess and access their targets. A brief outline regarding the application of social media evidence in legal trials is given Fig. 2.

## 2.2 Authentication for jurisdictional and privacy issues

Authentication of electronic evidence, especially social media, poses remarkable issues since anybody can make a fake profile and disguise under someone else's name. Besides, it is possible to manipulate the contents of another's data by getting the username and password. There are two basic criteria for admissibility of social media as evidence. First, it is necessary to authenticate the authorship of the evidence. Second, it is essential to provide the proof of authenticity and integrity of the material being presented to the court.

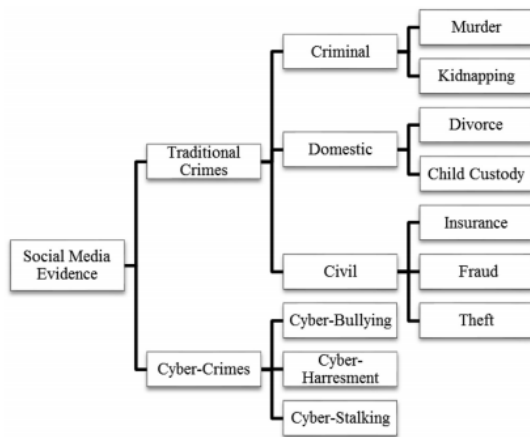


Fig. 2. Usage of social media evidence in legal cases [7]

A part of authentication also deals with the issues of preservation and chain of custody of the evidence being presented to the court. Courts rejected the proofs in simple printouts and screenshots because they can be tempered. Traditional methods of extraction and preservation of forensic data are not suitable for social media forensics. These authentication requirements demand advanced tools, which are particularly adapted for collecting, searching, indexing, preserving, and authenticating social media evidence.

In criminal case, LEAs obtain social media data of a suspect from social media providers through search warrants and government subpoenas. Information provided by OSN providers to serve a summons usually contains subscriber info, dates of connection, IP addresses and so forth. Social media providers are not ready to make concessions on the privacy of their customers by providing information about them to law enforcement.

This problem gets more intricate in global-jurisdictional conflicts, where the criminal activity occurs in one jurisdiction and social media providers existed somewhere else. Due to the different laws, legal agencies struggle with jurisdiction issues for preserving and accessing data held by companies in other countries.

Moreover, some legal procedure demand that investigators need to work with the companies hosting the data and ensure the process of collection observe with the statutory requirements, a chain custody and terms of use. This collaboration seems impractical under different legal jurisdictions. Therefore, it's need to develop consistent international legal frameworks, to address global cross-jurisdictional social media evidence access.

## 2.3 Forensic acquisition of social media content

Forensic artifacts are recognized as a critical source of evidence on social media. Hence most of the research efforts are focused on forensic evidence acquisition. The requirements for forensic collection from social media are generally outlined as

1. Collecting the relevant data or content from multiple social media sites.
2. Collecting metadata with social media content.
3. Ensure the integrity of data in the forensic collection process.

A recent research, Yusoff et al. [4] has successfully acquired 24 forensic images, each from FxOS internal phone and volatile memory. They managed to recover and trace social media

account credentials, especially on Facebook and Twitter services and also managed to get exact same forensic traces and evidences when they analyzed the same service in application and mobile web platforms.

The forensic acquisition of social media data through device forensics suffers the limitation of retrieving partial data. These applications are utilized to access OSNs on mobile devices. However, these handheld devices are not designed to save an entire copy of social media on storage. Few commercial tools, i.e., CacheBack, Internet Evidence Finder (IEF) and EnCase Forensic are also used with limited success to retrieve social media forensic artifacts from browser history and databases.

However, device analysis usually discloses other valuable pieces of information such as passwords, deleted artifacts, and additional online profiles by the user. Cellebrite recently released another product called UFED Cloud Analyzer, which enables clients to utilize verification codes and passwords saved by applications to automatically login to Gmail, Google Drive, Facebook, Twitter, Dropbox, and Kik [7]. Cloud Analyzer is then ready to download messages, message history, documents and contact records as accessible.

## **2.4 Archiving social media forensic collection**

All digital forensic processes, techniques, hardware, and software are intended to ensure compliance with the three fundamental and critical principles in digital forensics: first, the evidence must have been collected without altering it; second is to demonstrate the fact the acquired data is identical to the source, and third is that examination and analysis are performed in an accountable and repeatable manner. In addition to integrity few other criteria are also outlined in literature for appropriate storage formats to ensure sound preservation. These criteria include completeness of data, scalability of data management processes and flexibility of managing embedded metadata. The resultant archive must be scalable regarding reducing the

overhead involved in sophisticated analysis, search and mining methods. This feature is crucial in the case of social media archives that are bigger and diverse due to the number and variety of included artifacts [7].

Currently, investigators use specialized tools for forensic data collection in digital forensics, such as Encase, CacheBack, IEF, and more recent tools like Informatica Enterprise Data Integration tool and X1 Social discovery. The developers of X1 Social discovery claimed that it is specially adapted for social media forensics. Some other generic tools such as Aleph Archives, NextPoint, and WAR- Create are also used by detectives to preserve data from online social networks. Aleph Archives and WARCreate utilize the web-crawling approach for data collection and save this data to the Web ARChive (WARC) format. NextPoint stores the collection as PDF, HTML, and Portable Network Graphics (PNG) files, it also exports data to Concordance and XML. X1 Social discovery, save data in the WARC and MHTML (MIME Encapsulation of Aggregate HTML Documents) format; MHTML is a web page archive format employed to merge the HTML code, and it is accompanying resources in a single document. The package can export the data to Concordance, CSV (comma-separated values) files and HTML [7]

The large volumes of electronic data also dictate the importance of complete, scalable and traceable data preservation that ensures the integrity and advanced processing of archived data. Therefore the data can be managed efficiently and effectively with the help of numerous sophisticated analytical techniques.

## **3. Research Gap and Challenges**

Recent research [7] explains the challenges that law enforcement personnel face when handling social media data for forensic investigation. The first issue the authors observe is that in a single social media investigation, some data elements are considered out of context and not taken into account. Moreover, the components of the data

that they consider important are stored separately. All fragmented and unstructured social media information, although it may seem to be of little importance, would be very useful if it was in a coherent representation and chronological order. Besides, due to the separate storage of data, data analysis is limited to keyword search which is inadequate for a forensic investigation. The second challenge in a social media forensic tool development is to deal with the heterogeneity of different social media platforms. There are also several references in the literature that emphasize the need for cyber forensic tools that support heterogeneous research in an automated way [2][8][9]

### 3.1 Heterogeneous social media data analytics process

The objective research by Nikolaidou et al. [3] is to bridge the gap between the dispersed social media data and social media forensic applications. Their study defines a unified framework to exceed the heterogeneity of different social media platforms and to homogenize data. The security-oriented framework will permit both the preservation as well as the search and the correlation while providing automated tools for data analysis and visualization. They propose framework a social media analysis process that mainly corresponds to the three steps: tracking, preparation and analysis for data from multiple social media sources. The framework is depicted in Fig. 3.

After data tracking by platform-specific APIs provided by online social networks such as Facebook, Twitter, LinkedIn, and others, or crawlers, the preparation phase follows. In this work, a new ontology is introduced to reflect the majority of today's social media perceptions. The new ontology will be called Unified Social Network Ontology (USNO) [3]. To easily adapt the proposed approach, social-media-specific plugins have been developed for some social media sites. Each social-media-specific plugin takes as input the corresponding social media data. The USNO describes in detail how social media

data must be stored in the graph database. It depicts the node labels, the node properties and the relationships that will link the nodes/entities in the graph. The data is processed according to each social media site to retrieve implicit knowledge related to the unified ontology that introduced. Then, the data is mapped to the USNO.

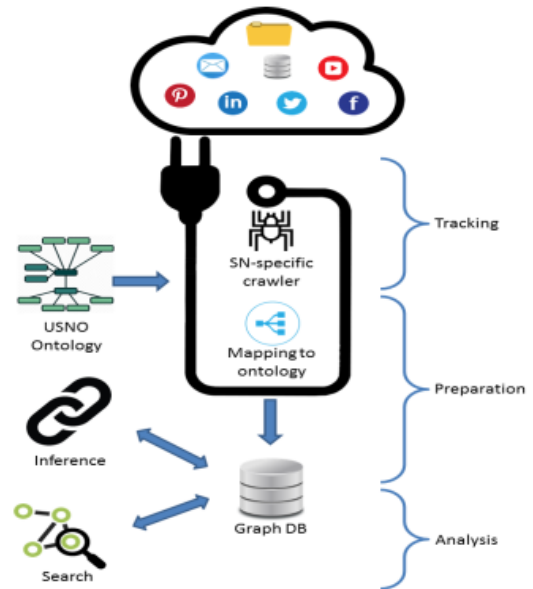


Fig. 3. Proposed framework for data preparation and analysis [3]

In the next step, the social media data is stored in the graph database. To thoroughly analyze the interaction between the profiles, an inference mechanism has been implemented to create direct links between the profiles that communicate with each other if these links are not explicitly provided. A for the analysis phase, the cypher queries has developed to retrieve information and relationships related to forensic research on profiles. Queries can take as arguments the inputs of the end-users. A number of predefined query placeholders which give insights into the network structure and reveal the most important profiles through SNA methods. Besides, LEAs will easily detect any indication that could provide the social media data. The automatic data homogenization, preservation and analysis will reduce significantly the investigation cost.

### 3.2 Unified social network ontology

The heterogeneity of social media platforms is another limiting factor in developing social media forensic tools and techniques. Social Media platforms are an autonomous and self-regulatory collection of online sites. They are designed independently, and every system differs from others in the set of services offered to the users. They differ in structure and models; they follow their data models. Every social network holds different modes of access and governed by a diverse set of rules. Besides their structure the data shared and posted on them also vary in structure and format; it includes images, videos, textual data, and applications [7].

Furthermore, people are accustomed to utilizing more than one social media platform hence the investigators must collect and examine data from multiple social media platforms even in a single investigation. Therefore, it becomes a challenging task for investigators to collect and interpret the varied data in the absence of any unified or standard tool [7].

Unified Social Network Ontology is depicted in Fig. 4. It consists of 11 classes that are associated with 20 relationships [3].

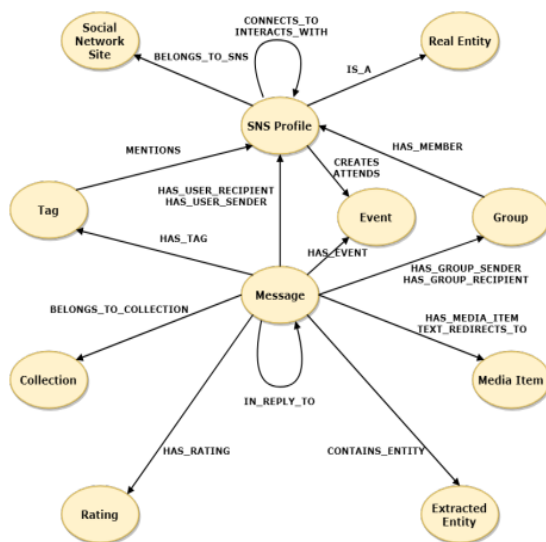


Fig. 4. Unified social network ontology [3]

In addition to the attributes listed below, each class has another property, its identification number.

1. Social Network Site: It can be any social media site like Facebook, Instagram, Stormfront etc. The properties include:
  - Platform: The name of the social network platform e.g. Facebook, Instagram, Stormfront etc.
2. SNS Profile: A social network profile. The properties include:
  - Username: A profile's username.
  - Profile title: A profile's title e.g. member, guest.
  - Email: A profile's email address.
  - Profile\_lang: The speaking language of a profile.
  - Registered date: The date a profile became member of the social network.
  - Ethnicity: The ethnicity of the profile.
  - Gender: The gender of the profile.
  - Birthday: The birthday of the profile.
3. Real Entity: A real physical entity e.g. a person, a business, a group of people. The properties include:
  - Entity type: The real physical person(s) who manage the profile e.g. person, business, group of people
4. Message: Refers either to the inbox or post. The properties include:
  - Msg Type: The type of the message e.g. post or inbox.
  - Content: The full content of the message.
  - Msg lang: The language in which the message was written.
  - Msg security: The security level of the message e.g. public or private.
  - Date sent: The date and time of sharing the message.
  - Num views: The number of views of the message.
5. MediaItem: A media item can have different content formats such as embedded image, embedded video, YouTube video, hyperlink

with clicking text or image and it is attached to a message.

- Mediaitem type: The type of the media item e.g. embedded image, embedded video, YouTube video, hyperlink with clicking text or image.
  - Mediaitem url: The url of the media item.
  - Mediaitem text: The alternate text of an image or the content of a clicking text.
  - Filename: The filename of the media item.
  - Lng\_lat: A text representation of the coordinate data as latitude and longitude e.g. "470999 1234300".
6. Collection: A collection contains messages with common elements e.g. topic, hashtag.
- Collection type: The collection's type e.g. topic, library, shared folder, hashtag.
  - Collection name: The collection's name.
7. Group: A group provides a space to communicate about shared interests with certain people.
- Group title: Title of the group.
  - Description: Description of the group.
  - Group security: The security level of the group e.g. public, closed or secret.
8. Tag: A link to a profile along with a media item.
- Bounding box: A text representation of 2 longitudes and 2 latitudes coordinates of a tagged profile in a media item e.g. "6.73462 6.75652 -53.57835 -53.56289".
9. Event : An event lets a profile to organize and respond to gatherings in the real world.
- Event title: The event's title.
  - Event location: An abstract region of space (e.g. a geospatial point or region) where the event will take place.
  - Event security: The security level of the event e.g. private or public.
  - Event timestamp: The time and date of the event
10. Rating: A measurement of how good or popular a message is.
- Num like: The score that depicts the positive rating of a message.
  - Num dislike: The score that depicts the negative rating of a message

11. Extracted\_Entity: An entity that is extracted from a message after performing named-entity recognition. It can be: location, person or organization.

- Extracted entity type: The extracted entity's type e.g. location, person or organization
- Extracted entity name: The extracted entity's name. e.g. for type "location", the name could be "London".

## 4. Conclusions

Social media evidence has the potential to be a powerful asset in any investigation using digital forensics. Since many crimes involve social media, the need for legalities and ethics when handling this evidence is very critical. Therefore, it is essential to develop innovative and better ways to associate and present the information to the investigators so that they can comprehend and better utilize that information. Heterogeneity across social media is a substantial problem to overcome the consistent and effective development of social media forensic tools. The social media analysis tools could be highly useful if provided with a visual representation of extracted and investigated data in addition to provenance management. The Unified Social Network Ontology (USNO) [3] has been introduced as part of a framework for tracking, preparing and analyzing data. The ontology covers the majority of today's social media perceptions.

In addition, Machine learning techniques can be applied to data classification, organization, and analysis. Big Data methods can also assist in managing and processing massive data volumes on social media. More importantly, an improvement in social media forensic extraction and preservation is also required.

## REFERENCES

- [1] Powell, A., & Haynes, C. (2020). Social Media Data in Digital Forensics Investigations. In *Digital Forensic Education* (pp. 281-303). Springer, Cham.
- [2] Kemp, Simon. (2019). 'Digital 2019: Q3 Global Digital Statshot'. Available at: <https://datareportal.com/reports/>

digital-2019-q3-global-digital-statshot (Accessed: 12<sup>th</sup> December 2019)

- [3] Nikolaidou, A., Lazaridis, M., Semertzidis, T., Axenopoulos, A., & Daras, P. Forensic Analysis of Heterogeneous Social Media Data.
- [4] Yusoff, M. N., Dehghantanha, A., & Mahmud, R. (2017). Forensic investigation of social media and instant messaging services in Firefox OS: Facebook, Twitter, Google+, Telegram, OpenWapp, and Line as case studies. In *Contemporary Digital Forensic Investigations Of Cloud And Mobile Applications* (pp. 41-62). Syngress.
- [5] BBC. (2017). 'Cybercrime and fraud scale revealed in annual figures'. Available at: <https://www.bbc.co.uk/news/uk-38675683> (Accessed: 12<sup>th</sup> December 2019)
- [6] Montasari, R., & Hill, R. (2019, January). Next-Generation Digital Forensics: Challenges and Future Paradigms. In *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)* (pp. 205-212). IEEE.
- [7] Arshad, H., Jantan, A., & Omolara, E. (2019). Evidence collection and forensics on social networks: Research challenges and directions. *Digital Investigation*.
- [8] Soltani, S., & Seno, S. A. H. (2017, October). A survey on digital evidence collection and analysis. In *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 247-253). IEEE.
- [9] Caviglione, L., Wendzel, S., & Mazurczyk, W. (2017). The future of digital forensics: Challenges and the road ahead. *IEEE Security & Privacy*, 15(6), 12-17.