Digital Investigation 29 (2019) 147–158

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

A multilayered semantic framework for integrated forensic acquisition on social media

Humaira Arshad [a, *], Aman Jantan [a], Gan Keng Hoon [a], Anila Sahar Butt [b]

[a] School of Computer Science, Universiti Sains Malaysia, Penang, 11800, Malaysia
[b] CSIRO, Gpo Box, 1700, Canberra, Australia

ARTICLE INFO

Article history:
Received 4 October 2018
Received in revised form
2 April 2019
Accepted 7 April 2019
Available online 11 April 2019

Keywords:
Online social network forensics
Hybrid Ontology model
Social network analysis
Social network forensic automation

ABSTRACT

In recent years, examination of the social media networks has become an integral part of investigations. Law enforcement agencies and legal practitioners frequently utilize social networks to quickly access the information related to the participants of any illicit incident. However, the forensic process needs collection and analysis of the information which is immense, heterogeneous, and spread across multiple social networks. This process is technically intricate due to heterogeneous and unstructured online social networks (OSNs). Hence, creating cognitive challenges and massive workloads for the investigators. Therefore, it is imperative to develop automated and reliable solutions to assist investigators. Capturing the forensic information in the structured form is crucial for automation, sharing, and interoperability. This paper introduces the design of a multi-layer framework; from collection to evidence analysis. The central component of this framework is a hybrid ontology approach that involves multiple ontologies to manage the unstructured data and integrate various social media data collections. This approach aims to find the evidence by automated methods that are trustworthy and therefore admissible in a court of law.

1. Introduction

In recent years, social media usage has increased rapidly throughout civilization. People tend to share their day-to-day activities, accomplishments, and sentiments on online social networks (OSNs). Hence, the data available on OSNs offer extensive knowledge about people and their inclinations. Criminals exploit the abundance of that information for orchestrating various cyber crimes such as malware distribution, fraud, harassment, cyberbullying and cyberstalking. They take aid from the online information to carry out traditional offenses like theft, kidnapping, and murder. Besides, they use the information as instruments to assess and gain access to their victims.

Also, law enforcement and intelligence agencies use these platforms to analyze the digital traces of misdeeds, left behind by offenders (Murphy and Fontecilla, 2013) (Seigfried-Spellar and Leshney, 2015). Social media evidence is used to expose the details of a crime, from the prospect of both suspect and victim. These platforms offer profile info, social contacts, photos of their day-to-day activities, and a partial view of their interactions. This information helps to reveal several aspects of a person's life such as relationships and events. Published content on social media along with associated timestamps may assist in discovering the whereabouts of a person in a specific time frame. It may also help to corroborate an alibi or might suggest some earlier or recent criminal activity. The availability of geo-tagged content and location-based services provided by OSNs like Facebook, Twitter, and Foursquare, even reveals the precise location of individuals.

In legal proceedings, attorneys use social media as evidence to defend their clients or prosecute the suspects. Despite privacy privileges of an individual, the social media content is legally accessible and admissible in courts (Christopher Paddock, 2016; Maurice Recchia, 2018; Mund, 2017). The law enforcement officers often have the privilege to access confidential information on social media profiles through warrant and subpoena. However, the defense lawyers are only limited to the public part of online data. Nevertheless, a court directive can compel social media sites to provide private data about a specific user or order the plaintiff and defendants to provide access to their social media profiles. Hence, the social media content can provide exceptional support for investigators in an inquiry, provided it is explored correctly.

Few critics raised the question about the validity and correctness of information posted on social networks (Sandra Dinora and

* Corresponding author.
E-mail address: humeraarshed@gmail.com (H. Arshad).

https://doi.org/10.1016/j.diin.2019.04.002
1742-2876/

Graciela, 2014; Viviani and Pasi, 2017). However, the issues of authenticity and validity exist for every evidence either digital or non-digital. Hence, it is a legal obligation to determine the correctness and reliability of the posted content before utilizing that info just like every other evidence in legal proceedings. Besides, social media sites record a large amount of metadata that is not directly controlled by users. Metadata also serves as a supporting means to authenticate the evidence.

The trials using social media data as evidence commonly involve cybercrimes, insurance, custody, and divorce cases. The defense and prosecution also used the proofs from OSNs in the trials for murder and fraud as observed in several publicly known cases such as the State of Louisiana v. Demontre United States v. Vayner and, Linscheid v. Natus and many others. A previous study has provided the details of several known cases and explained the role of social network evidence in detail (Arshad et al., 2019). Furthermore, the use of evidence from social networks is increasing since 2015 (GibsonDunn, 2015; John Patzakis, 2016; Patzakis, 2012).

Despite the noticeable standing of social media as evidence in legal proceedings and electronic discovery, the process of evidence collection from OSNs is neither straightforward not reliable. Technically, it is challenging to handle the substantial, dispersed, heterogeneous and unstructured content on OSNs for forensic and investigative use. Standard forensic analysis tools are limited due to the unstructured nature and heterogeneity of social networks. A tool designed for one social network cannot execute on other platforms due to the differences in structure and semantics. Currently, existing tools offer minimal functionality for collecting and organizing the forensic data from online platforms. The shortage of appropriate tools affects the efficiency and objectivity of the process and slows down the progress of examiners. As a result, the workloads for investigators and the backlog of cases are exponentially increasing (Adam Belsher, 2016; Lillis et al., 2016).

Investigators are compelled to use several tools even in a single investigation to access data from various OSNs. Still, they often skip crucial pieces of evidence or occasionally ignore entire volumes of online content, related to the investigation. Hence, resultant data sets are fragmented and unstructured which give a fractured view of events to the examiners. Therefore, despite a vast, promising and readily available set of information, the investigators cannot get enough support from the social media content. Furthermore, the substantially larger cases take several weeks of paralegal and lawyer time with the additional assistance of a forensic expert to gather and preserve the relevant evidence. Courts also acknowledge the cost, complexity and time needed to seek digital records, and only demand if they are critical to the case and if its cost and burden is justified (Edmond Burnett, 2016). Social media evidence is occasionally ignored due to high acquiring cost or presented inadequately due to insufficient collection and analysis.

The objective of this study is to define a framework that is suitable for semi-automated evidence collection and analysis of online social networks. This article is outlining a framework based on a hybrid ontology approach for integrating and managing the heterogeneous social media content into structured data. This approach consists of a set of upper and domain-specific ontologies for integrating the datasets from multiple OSN sources into a single, consistent and structured representation. Goals and contribution of this article are outlined as follows.

i. Structured Data Representation

This framework is implementing techniques to overcome the heterogeneity existing among social media content and sources to achieve a consistent and structured data representation. The structured information would allow automated analysis methods to process the data to get useful knowledge. Structured data is also needed for forensic data sharing and interoperability among tools.

ii. Integration of OSN sources

This framework is offering a hybrid ontology approach to integrating data from several online social network sources into a single, consistent and cohesive representation.

iii. Automated Analysis

This framework would support the development of suitable and automated forensic analysis methods. These methods will find the extent of relatedness among the entities and events. The analysis operators using these methods will sort and filter the data by using the correlations. They also present the deduced information through appropriate visualizations. The analysis operators can be incorporated into software tools to assist in forensic analysis and interpretation.

Section 2 of this articles provides a review of related literature. Section 3 explains the design issues that mainly describes the required features for the framework, and it also clarifies the technical difficulties involved in realizing these requirements. Section 4 explains the proposed design for the framework. Section 5 provides the implementation details and results of experimentation. Section 6 is concluding the article by outlining future work.

Related work

Forensic artifacts are recognized as a critical source of evidence on social media. Thus most of the research efforts have been focused on forensic evidence acquisition. Although, automated data collection is a crucial requirement to handle large and massive sources of digital information such as social networks. However, the heterogeneity of information found on social networks, analyzing massive data to gain knowledge, and legal requirements are other challenging aspects in the domain.

Social media forensic extraction and fusion

Initially, social media research focused on retrieving artifacts from digital devices, such as smartphones and hard disk drives. A study presented in 2010, a study indicated the potential sites of finding artifacts from the iPhone (Bader and Baggili, 2010) and another study identified the location of potential evidence traces on Android phones (Lessard and Kessler, 2010). Several subsequent studies examined the social media apps and the remnants of their data on various devices (Al Mutawa et al., 2011; 2012; Taylor et al., 2014; WalnyckyI MarringtonF Moore, 2015; Wong et al., 2013). Few studies also examined social media apps for windows based systems (Majeed et al., 2015; Wong et al., 2013). Most of the commercial tools such as CacheBack, Internet Evidence Finder (IEF) and EnCase, also followed the similar approaches to gather footprints from operating system databases and browser history (Cusack and Son, 2012).

However, all the device and operating system-based collection techniques were limited in term of completeness because the devices do not store the entire record of social media interactions and updates. Moreover, most of the data is repeatedly overwritten due to the limited memory of handheld devices. Therefore, device and operating system based forensic recovery do not provide complete information; this fact is acknowledged by several authors (Chau et al., 2007; Cho and Garcia-Molina, 2002; Ding et al., 2013; Psallidas et al., 2013; Wong et al., 2014). Despite the incomplete information, device analysis is instrumental in retrieving

information such as additional profiles, passwords and deleted artifacts by the user, that may not be retrieved otherwise.

Later studies suggested the use of web crawlers for online data extraction, including complete history from media sites, to overcome the limitation of incomplete retrieval (Chau et al., 2007; Cho and Garcia-Molina, 2002; Ding et al., 2013; Psallidas et al., 2013; Wong et al., 2014). Web crawlers provide more detailed collection due to their systematic browsing in addition to finding and following subsequent hyperlinks. More importantly, they archive the data in snapshots, that provide an identical view of original web pages and are also easy to preserve. However, the crawling behavior is not appreciated by social media providers because they generate high network traffic, by sending parallel requests and therefore disrupt the regular business operations of media providers. Social media providers implement IP and application based restriction policies to reduce the requests and to avoid the unnecessary burden (Cho and Garcia-Molina, 2002). This restrictive behavior by OSN providers turned out to be a significant limiting factor in web crawling operation. Furthermore, web crawlers cannot collect dynamic content such as interactive behaviors and metadata on social networks.Metadata is also crucial for establishing the credibility and reliability of forensic collection and analysis process.

At this time, many social network providers such as Twitter, Facebook, LinkedIn, and Foursquare, are offering the official APIs (Application Programming Interfaces) to capture the content, metadata, and other rich content such as interactive behaviors (Han, 2016). APIs based collection technique helps in accessing metadata and precise timestamps, which is not possible by web crawlers. However, the application based restriction policies adopted by social media providers also restrict the number of requests made through official APIs. Few approaches emulate human behavior to avoid the restrictions imposed by providers. Huber et al. presented a hybrid and optimized approach for online data extraction from social networks (Huber et al., 2011). This approach uses a combination of an automated web crawler and social media specific APIs to enhance the performance of the collection process from social media sites. Another work presented useful visualization methods for the data collected by using Huber's hybrid approach (Mulazzani et al., 2012).

In the real world, people choose to post different information on different social platforms, such as they may prefer to mention a job promotion on LinkedIn but favor Instagram for sharing the photos of a holiday trip. A report showed that 56% of users appeared on multiple social media sites in 2016. For instance, 93% of Twitter, 95% of Instagram and 92% of Pinterest users also used Facebook (Greenwood et al., 2016). Therefore, investigators are bound to examine all the social networks used by suspects or victims; to get maximum information related to an event or crime. Turnbull and Randhawa suggested that fusing the information from multiple social media would provide a less fractured and complete view to the examiners (Turnbull and Randhawa, 2015). Although, their work was limited to window based computer forensics only.

It is reasonable to argue that by combining and investigating the data from multiple OSN platforms would offer more comprehensive information related to an inquiry. However, gathering and analyzing the scattered information from multiple platforms is a daunting task for the investigators. They are compelled to use an assortment of tools to collect data from social media, but each of these tools collects data from a single platform and saves them in distinct data sets. For example, Facepager collects data from Twitter and Facebook; AlepArchive collects from Facebook, Twitter, LinkedIn, YouTube. Similarly, X1 Social Discovery supports over ten social media platforms, but manage them into separate collections for each network. Currently, there is no practical approach available to combine the data from multiple social networks into a single and cohesive data representation. Therefore, later during analysis, it becomes nearly impossible to relate, compare or associate separate data sets. Due to the reason, it is difficult to develop an accurate sequence of the events from data related to an individual but from multiple social media sites.

Content preservation

Content preservation is another challenging aspect of social media forensic collections. Currently, investigators and lawyers are using a variety of tools to gather evidence from social networks. These tools can be separated into two categories. First, are the specialized digital forensic tools such as Encase, CacheBack, IEF and few recent software X1 Social discovery and Informatica. Second, are generic tools such as NextPoint, Aleph Archives, and WARCreate that are also used to access data from online social networks. These tools mostly use textual formats for preserving the collections. For Instance, Aleph Archives and WARCreate save data in Web ARChive (WARC) format. WARC is suitable for long-term storage, but it can not support any entity based search or manipulation because of the linear and unstructured text file format. Likewise, NextPoint saves the data as PDF, HTML, and Portable Network Graphics (PNG) files. X1 Social uses MHTML (MIME Encapsulation of Aggregate HTML Documents), WARC, CSV (comma-separated values) and HTML for preservation. It can also export the data to Concordance (Fasching et al., 2012). Though the collection tools, both generic and forensic are offering various archiving mechanisms, the resultant archives are restrictive for the subsequent investigation and analysis. Furthermore, these formats do not have any built-in support for calculating the checksums and hash function to authenticate the integrity of the preserved data.

In this work, our discussion on data formats is only focused on their appropriateness for advanced analysis in OSN forensics. Several other features are critical for suitable forensic formats such as integrity, expressiveness, exchange, interoperability, and sharing. Digital forensics is using a few advanced formats such as XIRAF and AFF4 to represent digital forensic information. However, most of the commercial tools do not support these formats due to varied reasons. XIRAF is suitable to present hierarchical information structures but lacks in flexibility. While AFF4 is extremely flexible as it utilizes RDF framework for data representation, however, it needed a suitable supporting ontology, and currently, there is no standard ontology to support information representation and exchange. Recently (Casey et al., 2017) presented an open community developed specification language for interoperability and information exchange, they named it Cyber-investigation Analysis Standard Expression (CASE). Previously, open-source Cyber Observable eXpression (CybOX) schema, Structured Threat Information eXpression (STIX), Digital Forensic Analysis eXpression (DFAX) and Unified Cyber Ontology (UCO) were developed for the same purpose (MITRE, 2014) (Casey et al., 2015). However, all of these approaches are developed explicitly for expressing and exchange information. They are not intended to define suitable data models that can be used for analysis and tool development. However, the features offered by existing forensic formats such as flexibility and structured representation as offered by AFF4 and XIRAF respectively; would provide a benchmark for developing forensic preservation formats for OSN. Likewise, the concepts of interoperability, provenance and trust classification are given in CASE are critical in automation of OSN forensics.

Data and timeline analysis approaches

A significant number of studies exist for crime detection on social media. These are focused on automated detection of

cyberbullying, cyber-harassment and malware detection (Dadvar et al., 2013; Di Capua et al., 2017; Srinandhini and Sheeba, 2015; Van Royen et al., 2014). A keyword-based cyberbullying detection method on Twitter is presented by (Hon and Varathan, 2015). Some authors used natural language processing along with user activities and behavior to find aggressive and harassing behaviors (Chatzakou et al., 2017), while some used sentiment analysis to detect bullying on social networks (Dani et al., 2017; Nahar et al., 2012). Similarly, few methods are dedicated to tracing crime and criminal patterns on OSNs such as (Alami and Elbeqqali, 2015; Delavallade et al., 2017; Kastrati et al., 2015). However, we could not find any work related to forensic and timeline analysis of social networks. Even all the data on social media is arranged in temporal order, though its manual analysis is not feasible to discover evidence. A typical timeline may present data from a period of several years, consisting of thousands of activities that are not relevant to the current inquiry. Therefore, it is essential to devise techniques that can quickly filter the relevant data and sort it in a logical order.

All the techniques based on natural language processing and data mining are suitable for automatic detection of illicit behaviors. However, they are not suitable for forensic analysis and legal presentation due to two main reasons. First, legal presentation demands to explain the logical sequence used to obtain the results. Explaining that the order of events that lead to evidence is difficult when data mining methods are being used.These methods tend to lose the provenance of data during the preprocessing and normalization phase. Therefore, it is not feasible to associate the result with the source data (Glavic et al., 2013) (Hon and Varathan, 2015). Secondly, the natural language processing and data mining methods mostly use data processing methods based on clustering and probabilities which provide quick detection of illicit behavior with significant false positive rates. False positive is ignored as a necessary evil in automatic detection systems but not acceptable in judicial decisions. Similarly, the use of probability to indicate an illicit behavior provides the implication but not the evidence.

Semantic web techniques in social media forensics

It is observed from existing computer-based digital forensic and timeline analysis studies that approach based on semantic web methods provided better results for automation, managing heterogeneity and knowledge representation in the domain (Chabot et al., 2015; Schatz et al., 2004a, 2004b; Turnbull and Randhawa, 2015). Ontologies are the central component in semantic web methodologies for appropriate knowledge representation. They provide an explicit and formal specification of the domain. Ontology-based approaches can provide better and formal knowledge representation for the analysis process on the complex and heterogeneous datasets from social networks. Currently, very few ontologies exist that are specifically designed for social media, and none of them is detailed enough to use for integration or analysis.

SIOC (Semantically-Interlinked Online Communities) is a generic and open-standard ontology for expressing, explicitly and implicitly, the information on social networks. SIOC is a part of a project started in 2004 was regularly updated and published (Breslin et al., 2005); in 2005 (Bojars et al., 2008), in 2008, and (Breslin et al., 2009) in 2009. It is a robust but general-purpose ontology meant for Linking Semantically-Enabled Online Community Sites; however, it is not suitable for forensic and event reconstruction. Another ontology named “SC-Ont” is published in 2016 (Kalemi and Yildirim-Yayilgan, 2016), the authors claim that this ontology is developed for forensics and crime solving on social networks. Although it does not provide any details of how this ontology would help in crime solving other than cataloging online digital evidence collection. Furthermore, this ontology is also not developed for forensic or automation purpose,hence, lack the required level of detail.

Requirements and design issues

The framework presented in this article aims to provide a suitable solution to integrate, manage and analyze the forensic data from multiple and heterogeneous social networks. It would allow the automation of the forensic process to assist the investigators during an investigation for the collection and interpretation of collected data. This work identified a few requirements based on the limitations and gaps found among the existing approaches. These features are critical for the development of the intended solution. This work also identified the related issues and challenges by a critical review of the existing literature, that is restricting the development of desired features. These problems must be addressed to design an appropriate and acceptable solution for automated OSN forensics and achieving the goals that are outlined in section 1. The requirements and design issues are outlined as follows.

Completeness and flexibility

The efficiency and accuracy of the reliable results, produced by analysis, depend upon the completeness and flexibility of the underlying collection process. If crucial information is ignored in the collection process, the resultant data will be incomplete. The missing data would affect the deduction process, and the end results might not be reliable. Likewise, if the underlying structure of preserved data is not flexible enough to allow the execution of basic queries on data; it would be impossible to test various theories for investigators. A strict or linear data structure may not allow the examiners to observe the logical order of events and hidden correlations among the data.

Current, approaches have provided technically suitable solutions for automated data collection. However, they lack in integrity management and preservation needed for appropriate forensic collection. Currently, the formats that are used for preservation are textual, unstructured and linear in structure. No specific format is designed to preserve the complex and heterogeneous OSN datasets. Therefore, they do not allow any complex query execution or an extensive analysis of the data.

Automated forensic analysis

Data collection and analysis are time intensive and multi-dimensional phases in digital forensics (Casey and Rose, 2010). The analysis phase interprets the factual information gathered in the collection phase. The interpretation involves the integration and correlation of extracted artifacts; to know the linkage such as who interacted with whom and to find the order of the events. This linkage and sequence would lead to attribution to a person.

As mentioned already, that forensic tools are lacking sophisticated analysis features due to the underlying data collection and storage formats. Even simple entity-based searching, sorting and filtering of data is not fully supported. Therefore the analysis abilities are still limited to keyword search mainly. Keyword search is an essential and efficient way to find the required information quickly, from the massive bulk of digital data. However, only the keyword search is insufficient and inefficient in processing the complex and large datasets for investigation and evidence collection.

Due to the insufficiency of keyword-based analysis, it is imperative to develop analysis methods that help to correlate separate, seemingly irrelevant pieces of information together to

arrive at some plausible conclusion by judgment and deduction process. The conclusive fragments of information can be presented as evidence in the legal proceedings. These analysis methods should be able to quickly filter and sort the data based on different parameters relevant to the investigation. The searching and sorting parameters will vary from case to case.

Provenance management

Provenance denotes to the origin and history of an object; it explains the conditions in which the data was found, preserved and then processed to its present state. In forensic analysis, it is essential to manage the provenance of data regarding people, entities, and activities involved in producing related data objects. Otherwise, the results produced by the methods, which cannot maintain and provide the provenance of data like data mining techniques, would be rejected in legal proceedings. The use of scientifically unproven digital forensic evidence in legal proceedings has always been criticized, and evidence is not acceptable in court if it cannot be verified through scientific evaluation (Arshad et al., 2018). Due to this reason, the ability to recover reliable information that describes the origin and step by step evolution of data is necessary for scientific evaluation and legal validation.

Therefore, the data mining and natural language processing techniques used for identifying crime patterns and criminal profiles are not suitable for social network forensics as explained earlier in section 2.3. Thus, the analysis operators needed in OSN forensics must be logically explainable in a court of law. Additionally, they must maintain the provenance of data to link the end results with the original data.

Combining OSN forensic sources

Technically, integrating the data from multiple platforms is not a trivial task, mostly because of heterogeneity. Ontologies are commonly used for formal domain modeling and integration of heterogeneous data. However, information can be represented at different levels of formality. Generally, two types of ontologies are used; the upper level and domain-specific ontologies. The upper-level ontologies contain generic terms that are shared by related domains and provides an abstract view of the domain. These ontologies are used for semantic interoperability and semantic data integration. While the domain-specific ontology represents the concepts within a particular domain along with their relationships and properties that are related to a particular data source; upper ontologies are preferred for an appropriate representation of the local schema in domain-specific applications such as data analysis.

Social networks differ from each other in structure, services, and business rules and may be accessed differently. They exhibit semantic, syntactic and schematic heterogeneity. Therefore, in this framework both types of ontologies are needed; an upper-level ontology for integration of OSN sources and domain-specific ontologies are required to represent data sources for individual OSN sources, such as a local ontology is needed to represent each Twitter or Facebook.

Multilayered semantic framework

In this work, the fundamental features of a semi-automated OSN forensic approach are outlined. The proposed framework would allow to collect, manage, structure and analyze the forensic data from social networks. The fully automated analysis is not feasible due to the diversity of cases and legal constraints. Hence, decision making is delegated to the investigator. This approach is integrating many components; some of them are proposed as innovative solutions and addressing the research issues, such as ontologies and ontological framework. Although few components have technical concerns, these issues are solved in existing work/literature, and we integrated them into our approach by considering their correctness and suitability. For instance, the issues related to the automated collection of social media data and metadata are already solved by (Huber, 2012) and integrated into the current approach.

There are two essential aims of this approach. The primary objective is to present a largescale, structured consistent and formal knowledge representation for social media forensic data integration. This objective will be accomplished through a knowledge model and encoding the information into ontologies. The second aim is to perform reasoning over the resulting data set and derive higher abstractions of data. The implementation of analysis operators and query translation will demonstrate the feasibility of automation on the integrated data.

Managing OSN heterogeneities

The first and foremost problem in the integration and management of OSN content is to deal with the heterogeneity. Each social media is not quite the same as others in structure and model. They exhibit semantic, schematic, and syntactic heterogeneity besides access heterogeneity. Semantic heterogeneity stands for the dissimilarities in interpreting terms or meaning of data on each network. Schematic heterogeneity demonstrates the differences in structure and model of each social media. Similarly, syntactic heterogeneity shows the variations in data formats. Access heterogeneity refers to the variations in access methods for each OSN such as web interface or apps.

Fig. 1 summarizes the types of heterogeneity involved in social media platforms and summarize the suggested techniques to overcome the variations. This approach is using a set of parsers to manage the syntactic and access heterogeneity of the content. A set of ontologies that exhibit varying levels of details are used to manage the semantic and schematic heterogeneity. The details of the ontology model are given in section 4.2.2.

The utilization of shared ontologies across multiple systems creates a consistent language, as similar terms are used to characterize the analogous data. This feature is useful to manage the dissimilarities in divergent systems. The incompatible system utilizes diverse terms to refer to the same concept or utilize identical

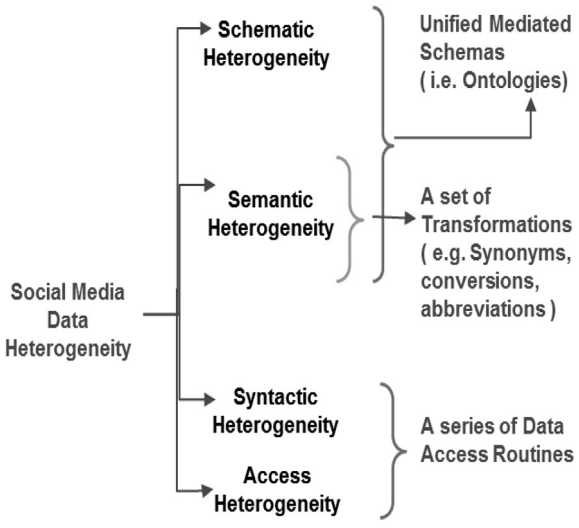

Fig. 1. Proposed Methods For managing OSN Heterogeneities.

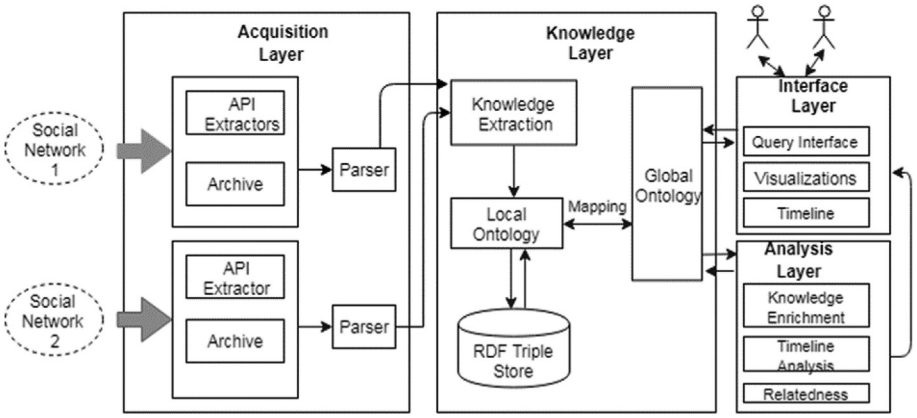

Fig. 2. The architecture of Multilayer Semantic framework for OSN forensics.

terms to explain distinct things or may indicate a different level of detail.The unified terms explicitly defined by ontologies create a cohesive language across multiple domains; that allows the information exchange and integration.

Architecture of multi-layer semantic model

An overview of the proposed framework is given in Fig. 2. It consists of the following four layers with the central concept of implementing the knowledge model through ontologies (see Fig. 3 and 5).

1. Acquisition layer explains an appropriate evidence collection approach.
2. Knowledge layer that proposed an ontological model to manage and integrate data from multiple social network sources.
3. Analysis layer describes the formulation of automated analysis operators based on the ontological schema to analyze the content extracted from OSNs.
4. The interface layer is for the presentation of deduced knowledge and evidence through appropriate techniques and visualization that will increase the comprehension of complex data.

Acquisition layer

The existing practices and methodologies have indicated limitations in a few areas of forensic collection and data management on social media. As a matter of fundamental importance, this work proposes an amendment in the pre-requisites for online social media collection. We insist on the complete collection of OSN content and all the accompanying data or metadata in context with original data. It is essential to collect all the component of data in addition to visible textual or multimedia contents such as activity data that shows online user activity or network aspects. We argue that the analysis of metadata can also yield useful information for an investigation. Additionally, any analysis of metadata can be related to the content for evidence presentation and authentication. This requirement is necessary to envelop all the vital aspects of the forensic collection on OSNs for further advance processing.

Furthermore, we advocate the use of an iterative and incremental approach for forensic extraction on social networks. Unlike standalone devices, it is impossible to preserve everything of interest from OSNs at once; due to their immense size, connectivity and shared ownership. Therefore, it is highly likely that the data collected in the first step may reveal new knowledge that indicates the involvement of new actors or suggest the presence of relevant evidence somewhere else on social networks. In that scenario, the investigators can expand the boundary of the forensic collection on OSN.

Data sources and extraction. In OSN forensic the data available from online platforms is the primary source of data. In this framework, we suggested the utilization of platform-specific APIs and crawling behavior for the collection of data from OSNs. Most of the prominent OSNs offer official APIs for the application developers, such as Facebook, Twitter, LinkedIn, and YouTube. These APIs are adequate to gather complete information and metadata from social networks. Specific parsers are used to extract data from various sources. The parsers are also responsible for transforming the individual data values to the relevant entities in the ontologies.

In addition to online content, social media data archives is another source of available data. Many social networks such as Facebook and Twitter offer archive option for data. This information is downloaded only by the user himself. However, in some cases, this information is provided to investigators by the users on court order or accessed by user consent. This information source is significant in the way that it offers slightly more data than collected by other methods such as login IP addresses and private data (i.e., direct messages, chat).

The provenance management highly depends on the completeness and extensiveness of the underlying data acquisition layer, as well as on the capability of a knowledge layer to appropriately catalog the data and provenance information. Data acquisition layer is specifically designed to collect all the available information accompanying an artifact, including metadata, that is a

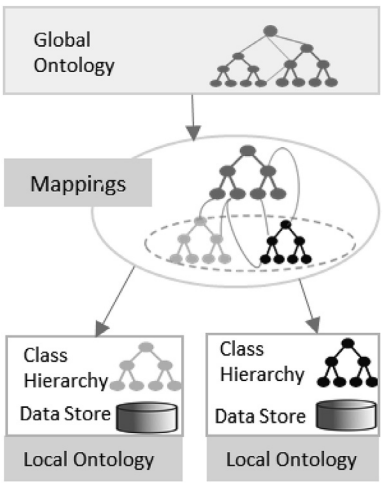

Fig. 3. Hybrid ontology model.

necessary requirement for provenance management.

Data normalization. Next step after the data collection, is to process the collected data and make it suitable for the knowledge layer. The data collected from various social networks and data sources are heterogeneous in structure and syntax — for instance, each source store data and time in different formats. This layer is also responsible for translating the data into the format, used by the upper layer. Therefore, every source needs a specific parser for translating the data. Each parser is used to read the relevant data source and translate it according to the knowledge model by identifying the suitable entities, events, and relationship. However, this layer is required to prevent any loss of data and maintain provenance in the process. The datasets collected from OSNs such as Twitter and Facebook are translated and normalized according to the respective local ontologies. The normalized data is then stored in RDF data stores by using suitable parsers.

The parsers in this layer will ensure the translation of every data item including metadata, to the right attribute, entity or relationship in the integrated knowledge layer.

Knowledge layer

The knowledge layer is mainly responsible for handling the semantic and schematic heterogeneity through unified mediated schema or ontologies. Ontologies and data stores permit a more enhanced form of storage than the typical databases. They allow for appending new and divergent properties to the existing objects. Additionally, ontological representation provides a common language and a foundation for reasoning. Fundamentally, ontologies allow a natural encoding of associations and entities with attributes. This representation can characterize the data as graphs with objects as nodes and properties as links; thus suitable to present social media data for automated analysis. The information stored in the graph structure is accessible by machines for comparison and analysis.

For instance, if the investigators identify a post or image as evidence, it must be viewed with all the related data. Such as the time when the post is uploaded, the device used to post, tagged place, people who liked it, to whom it is shared, people who are tagged, who responded to the post, what is said in reply and so forth. It is essential to have all the associated data because; first, it improves the decision making for investigators, either it is conclusive or needs further examination; second, it helps later, to authenticate the evidence.

Hybrid ontology approach. This framework is implementing a hybrid ontology approach for managing the integration of information from multiple social media platforms. A similar approach was also used for the interoperability of the semantic web (Cruz and Xiao, 2003).

This hybrid approach is used to represent the information at two levels of formalization through local and global ontologies. Local ontologies have more specific and detailed information regarding the specific social network, while the global ontology has upper-level and generic concepts.

A local or domain ontology is needed for each source schema such as Facebook or Twitter by using hybrid approach. The local ontology will describe the semantics of each source. A single global ontology is required in this model that provides the generic concept of OSN forensics. Global ontology facilitates data integration without including the data stores. The local ontologies are mapped to an upper ontology instead of mapping to each other. A separate ontology and source-to-target mapping are needed for adding a new source in this model. Additionally, if an existing source is changed or evolved, it might require a few adjustments in the defined mappings.

The use of hybrid ontology approach in this model offers several advantages. First, it allows the easy addition of new data sources. Many existing OSN sources are continuously developing new features, and many more platforms are surfacing rapidly. Thus, in this approach, the latest social network can be added to the existing framework by formulating new mappings. Second, the use of local or domain-specific ontologies would allow the explicit and consistent representation of metadata through local ontology. Third, because of global conceptualization through the upper or global ontology, it provides a consistent view of the schematically-heterogeneous source schemas. Fourth, it will provide support for customized and high-level queries. The investigators can formulate a query, by using the upper-level ontology, without specific knowledge of the underlying data sources. Then the query will be reformulated for specific social media sources such as for YouTube or Twitter. The query rewriting is based on the semantic mappings between the upper and local ontologies. Fifth, an explicit vocabulary, of formal terms through an ontology would facilitate automation.

Ontology mappings. In this approach, the upper-level or global ontology provides a way to integrate the information from several disparate social media platforms. In this model, the local ontologies are mapped to global ontology instead of mapping to each other in a point to point manner. Point to point ontology mapping is not suitable for social media sources due to a large number of potential local ontologies and the amount of overlapping among each source.

The local schema defines a local perspective on the global ontology. The global ontology acts as a mediator among local schemas. Mappings will be developed for each local to global ontology, that needs mapping rules, that indicate the corresponding terms in the other ontology. The mapping process needs the use of consistent terminology among local sources. A set of transformations, such as abbreviations, synonyms will resolve the variations in the terms. A set of mapping rules will be created that relate the terms in the global ontology to the preferred terms for a local social media source.

Analysis layer

As discussed in previous sections, the analysis of such a large and varied dataset require sophisticated tools. Analysis layer is using analysis operators and query translation mechanism to check the feasibility of the underlying data integration model.

This work aims to propose some automated analysis methods. It is important to note that the operators implemented in this work are derived from other studies and that can be subject to debate. However, the main aim of the analysis layer is to demonstrate the relevance and capability of the ontological model for managing the data and therefore, support the development of automated analysis and visualization tools. The general analysis involves the summarization of overall data. For this purpose, a few generic methods based on frequency analysis and clustering algorithm are developed. These analysis methods are derived from the existing studies related to social behaviors interactions on online social networks.

A query translator is needed for the formulation and evaluation of high-level queries. In query translation, global ontology acts as a mediator and maintain an all-encompassing schema and mappings between the global and local ontologies. A user can pose a query and explain the request through general concepts and terms of global ontology. Then the mediator translates the query into subqueries by using a reformulation procedure. The rewritten queries will execute on local sources, so the mediator can collect and combine the returned results and present them as a response to the query.

Interface layer

This layer allows the investigators to interact with the knowledge model. This layer aims to present the data, filtered or sorted by the analysis layer, in a reader-friendly manner. The knowledge deduced from analysis operators is presented by using appropriate visualization techniques to increase the comprehension of data and evidence. For instance, to provide a quick view of significant themes in user conversation, the sorted topics can be presented through a word cloud. It is crucial to select and test many more visualizations for presenting behavioral and geographical patterns. Additionally, this layer provides a query interface that can be used by skilled users to carry out SPARQL queries in response to dynamic queries by the investigators. The results of generic analysis operators would also be presented through appropriate visualizations such as cluster graphs, relative or cumulative frequency histograms, time series and scatter graphs as shown in Figs. 6 and 7.

Implementation and experimentation

Implementation of the model is performed by collecting and integrating the data from two OSNs Twitter and Facebook. The social network datasets are collected from online Twitter account and Facebook archive. The datasets are partially integrated by using the ontology model and stored in RDF data stores for testing. In experiments, we are using Trivial database TDB for the storage of ontology elements as RDF dataset. TDB is a component of Apache Jena for RDF storage and retrieval through semantic query. Apache Jena Fuseki is used for SPARQL server.

Ontology design and implementation

In this model, a local ontology is used to store and arrange all the components of data in a structured way. The social media information is unstructured; it contains textual content (i.e., messages, posts, tweets), multimedia content (i.e., photos, videos), interactional behaviors (i.e., likes, shares, replies) and relationships (i.e., friends, followers). Besides, it incorporates the time-based data and metadata. Thus, it is very challenging to organizing and indexing the structurally and semantically divergent data and also maintaining the relationships among entities. The following example is used to explain the problem and model implementation.

Facebook and Twitter are social networking sites that support microblogging and multimedia sharing. However, despite the similar features they offer for social networking, they exhibit several differences in the overall business model, and their data models vary significantly. The Twitter data model is based on objects that encapsulate other objects and core attributes that describe that object. Tweets are the basic building block of all things in Twitter data. Tweet objects act as 'parent' object to several child objects. The child objects of Tweet include User, Entities, and Extended_Entities. Geo-tagged tweets also include as a 'Geo' child object for the place.

Facebook uses an open graph structure called 'social graph.' It represents the data on Facebook as a composition of nodes, edges, and fields. Things such as a user, an image, a page, or a comment are referred as nodes, while the connections between those "things" edge such as photos on a page or a user's comments connect the nodes. Fields refer to the essential attributes that describe the node or edges. Nodes are further divided into root and non-root nodes; root nodes are directly accessible while non-root nodes are accessed through relative root nodes. A typical but partial schema for both social networks is shown in Fig. 4 to explain the difference in structure.

The example presented in Fig. 4 only explains the basic concept and core attributes; it skipped other fields to avoid complexity. The schema shows structural differences such as the "place" element; it is part of tweet object on Twitter, but it is attached with "Post" and "Status" on Facebook, while status is part of "Newsfeed." This placement of the object changes the visibility of the communications on social media platforms. However, they are semantically equivalent data elements. Such as we can obtain a list of places tagged by the user from Facebook and Twitter by following different path patterns.

This approach is using "model-based schema transformation" so attributes names used here are the same as referred by respective official APIs for Facebook and Twitter. Objects and nodes are converted to RDF classes and attributes into RDF properties which are attached to the corresponding classes in local ontologies. The global ontology is representing the high-level concepts from social network forensic domain. This ontology acts as an intermediate schema that allows to incorporate new schemas (OSN sources) in the existing design and make it extendable. A partial overview of detailed integration design is given in Fig. 5.

In Fig. 5, the upper half is showing a partial view of global ontology while the lower left and right halves are showing few of Twitter and Facebook ontology concepts respectively. The dotted lines are representing the mappings among the concepts from local to upper ontology while the solid lines are showing the relationships within a single ontology. The partial ontology concepts are explained by using Visual Notation for OWL Ontologies (VOWL) (Janowicz et al.,). In an implementation, the mapping process involves the process of class merging and generalization. Related properties, classes, and relationships are merged with or extended from the upper ontology.

As indicated in Fig. 5, the user object is mapped to a subject in upper ontology, and the subject is associated with an incident that is explaining the occurrence of an illicit event. The "User" in domain schema is mapped to subject in global schema. The "Subject" in

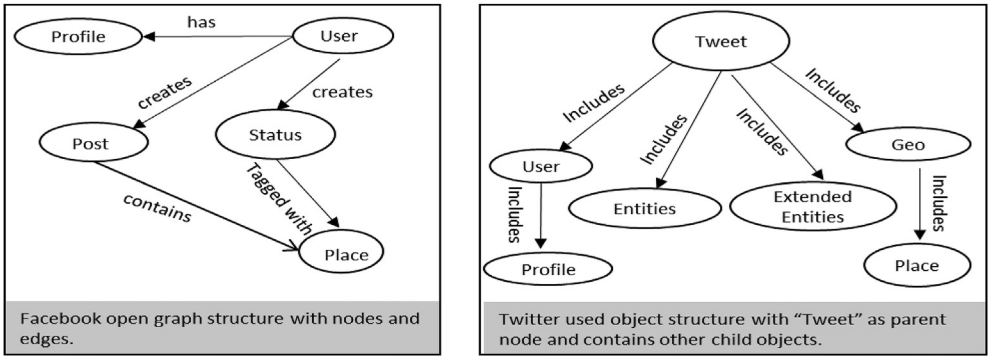

Fig. 4. Partial schema for facebook and twitter.

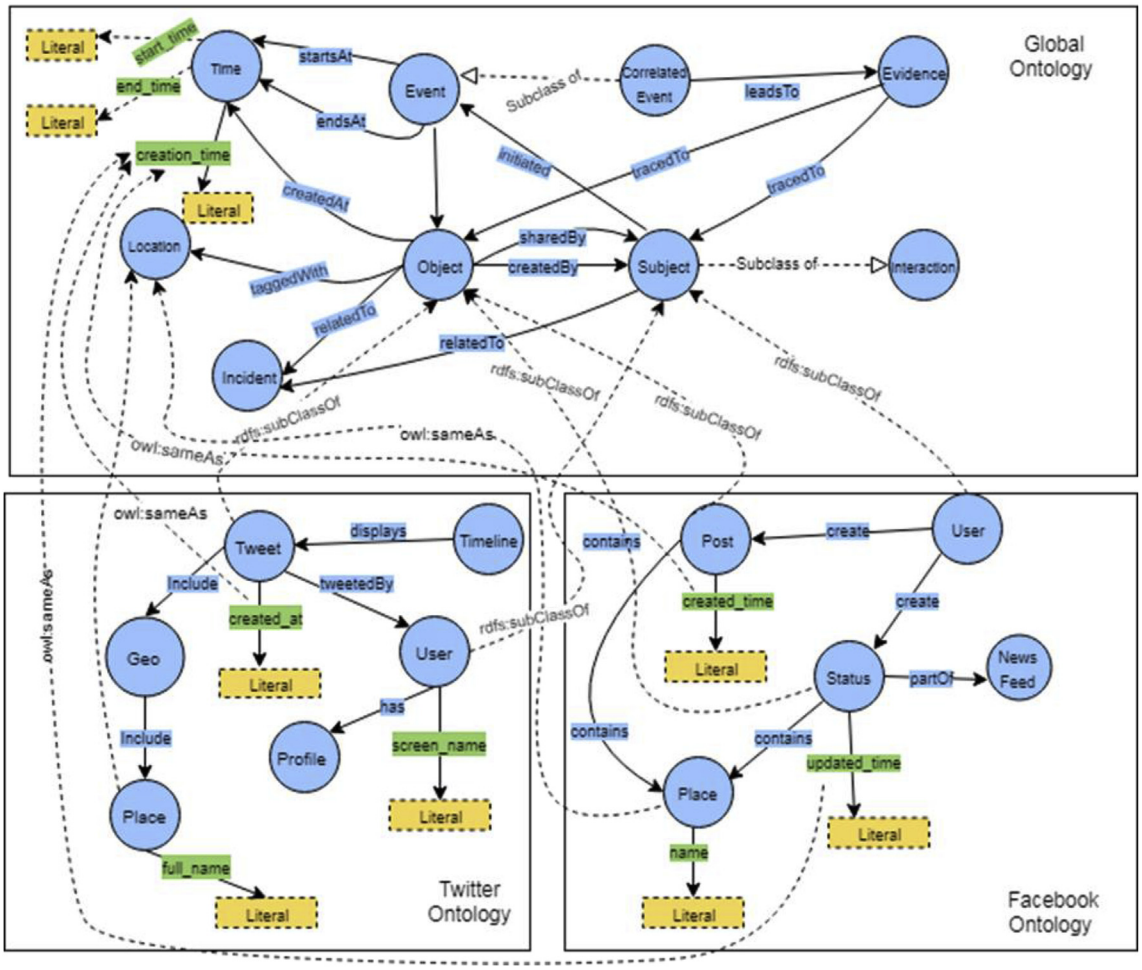

Fig. 5. Concept demonstration for Hybrid Ontology Model.

upper ontology is explaining the role of that user in an investigation, the related data sources and his involvement in an incident. The ontology structures given Fig. 5 are just for demonstrating the core concept of OSN data integration and omitting implementation details to avoid the complexity. The examples here are showing only a few data and object properties and no inverse object properties related to the entities.

A typical object from upper and domain ontology is presented as follows by instance 1 and 2.They are showing the concept of "User" and "Subject" from local twitter ontology and upper ontology respectively. They are demonstrating the implementation concept of ontology objects, and they are presented by using Turtle serialization. The user individual in the local schema is presenting all the attributes with data values from the Twitter schema.

Instance 1. User Concept on Twitter.	Instance 2: Subject from Upper ontology
#User	#victim
: User an owl: Class;	: victim rdf: type owl: Class; rdfs:subClassOf:Subject.
An individual of class User.	***An individual of type Subject***.
:Eve a owl:NamedIndividual,:User;	:eve rdf:type owl:NamedIndividual, :Victim;
:name "Eve";	:type "User";
:user_id "25073877"	:subject_id "User_CB_012XX";
:profile data	:incident
:created_at "Sat Jan 06 04:32:08	:incident_date "Sun 16 Dec 2018";
+0000 2018";	:incident_id "CB_12XXX";
:favourites_count "100";	:incident_name "Cyberbullying";
:followers_count "50";	:datasource 1
:friends_count "20";	:ds_name "Twitter",
:statutses_count: "20";	:ds_type "Online",
:geo_enabled "true";	:user_id "25073877"
:contributors_enabled "false";	:datasource 2
:has_extended_profile "false";	:ds_name "Facebook",
:id_str "25073877";	:ds_type "Archive",
:location "XYZ";	:user_id "454548751578"
…	….
#	#

Query translation

Investigators may ask a high-level abstract query such as they want to find the timestamps of all the content posted by a user recently, on various social media; used by that person. The query mentioned above is first written for global ontology by using unified terminologies. Then this query is reformulated for Facebook and Twitter ontologies separately to compute the results from distinct sources. Eventually, the mediator will combine the results from separate sources and present them in an integrated manner; by using the consistent vocabulary of global ontology.

based on the parameters provided by the investigators to describe the incident. The parameters will be the subject, object, period or location relevant to the incident. The correlation operators will allow the investigators to have a quick and overall view of the events and the links among them. The identification of correlated events will be performed by using the following criteria; First, the degree of relation among events due to a common subject, second due to shared objects. Third, temporal proximity; four, geographical proximity due to a physical location or IP address.

Interaction graphs are based on the concepts that only a few individuals among a social graph are actively communicating. This

```
SELECT ? time_stamps WHERE
{
?user rdf:type snfo: subject.
?user snfo:name "Alice".
?objects snfo:isCreatedBy ?User.
?objects snfo:creation_time ?time_stamp.
}
```

Query (A) . Timestamps for all the objects created by a user “Alice.”

```
SELECT ? time_stamps WHERE
{
?user rdf:type twitter: User
?user twitter:name "Alice".
?tweets twitter:isTweetedBy ?User.
?tweets twitter:created_at ?time_stamp.
}
```

Query (B). Timestamps for all the Tweets created by a user “Alice.”

An example of a query to find the timestamps of the objects created by a user “Alice ” is given in “Query A” that is written for upper ontology named Social Network Forensic Ontology (snfo). “Query B” provides the translation of that query on the Twitter data set and ontology (twitter).

Analysis operators

Social media data is already arranged in a timeline; hence the primary analysis tool in this work will be a timeline analysis method, based on finding the correlation between two social events. The correlation represents a link that shows a casual or specific relationship among events that are related to the incident under investigation. The identification of such relationships will be

idea is derived from interaction distributions and explains in (Wilson et al., 2009). Analysis operators that produce interaction graphs and temporal patterns are formulated and tested on the dataset collected from Twitter. An example of the interaction graph is given in Fig. 6. The inner circle represents the subject, and outer circles represent the social contacts from OSN for that person. The contacts that used to interact more with the subject are nearer and more substantial in the visualization in comparison to the others with less communication. Likewise, the thicker connecting lines are showing more density of communications between the two persons.

Typically, the social network usage patterns are also unique among individuals, and they tend to maintain these patterns (Aledavood et al., 2018) (Aledavood et al., 2015) (Randler et al.,

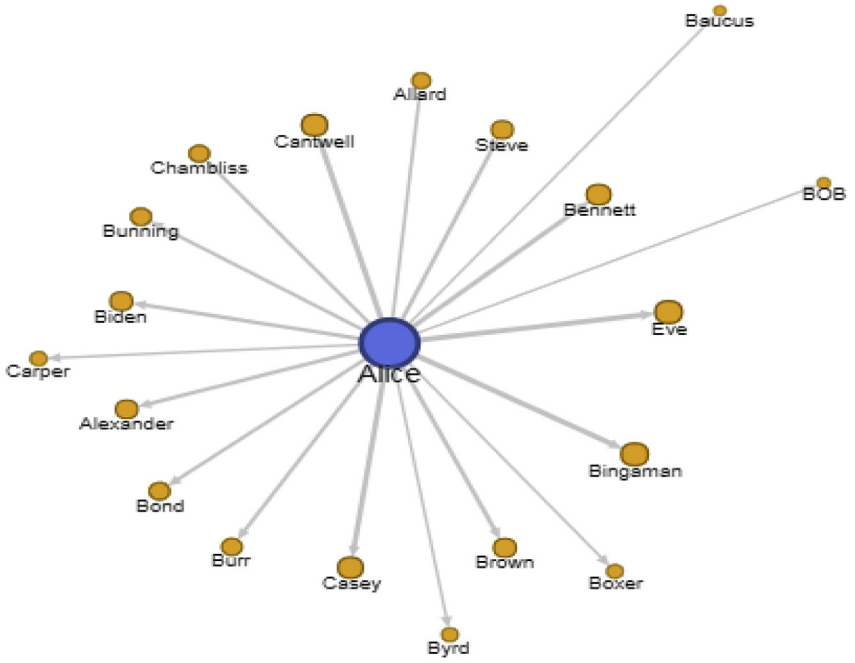

Fig. 6. Directed and weighted Interaction Graph from Subject to Contacts.

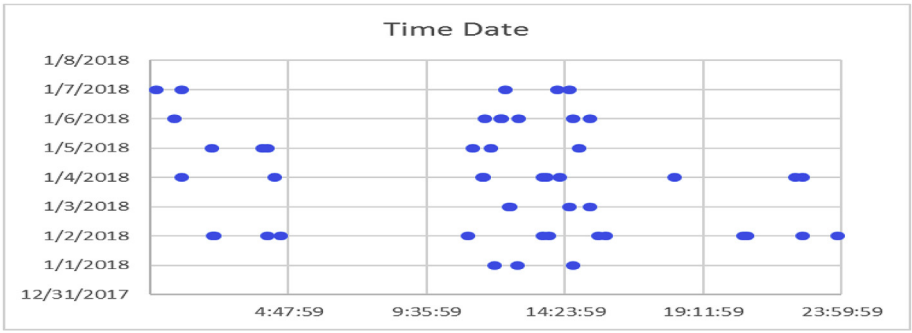

Fig. 7. Temporal activity pattern by subject.

2016). Hence, by inspecting these rhythms closely, they may help to compare the identity of an anonymous user with a known user and thus to find the identity. An analysis operator to observe the temporal pattern shows the temporal activity pattern of an individual as shown in Fig. 7. The scatter graph is showing the more frequent times of online activity observed for a user from one month of the data. The respective operator collected the timestamps of every online interaction by a specific subject from extracted data and plotted the intervals with higher activity level.

The analysis operators used in this approach will be more focused on summarizing and filtering the data by using domain semantics and proximity. These operators also intend to present the data through visualizations to increase the understanding of analyzed data. Investigators can formulate and test the hypothesis and explore the data with the assistance of these operators. They can choose the final evidence with their expert knowledge.

Conclusion and future work

A deliberate and elaborate approach is needed to respond to the requirements of the social network forensic domain, that is expanding swiftly. Current literature and commercial tools show that social media forensic collection and analysis are limited in abilities. Hence, a systematic approach is required to manage the issues of heterogeneity and large volumes of data by supporting appropriate automation. Otherwise, without an efficient and precise approach, the process remains unreliable, complicated and time consuming for law enforcement and legal practitioners.

This article has presented a new systematic approach to managing the most intimidating challenges of social media forensics. It presented a hybrid-ontology approach for collecting and integrating the forensic data from multiple social networks. This methodology would allow the integration of multiple social media sources with accuracy and precision. In the future, we aim to implement and test the feasibility of using advanced automated operators on the suggested approach. Artificial Intelligence and machine learning paradigms are also valuable in this field (Abiodun et al., 2018).

Nevertheless, this approach is designed with the provenance management concept. Hence it focused on the collection and management of all the relevant data and metadata. Furthermore, it focused on maintaining transparency in automated analysis procedures to avoid the limitations of data mining approaches. However, it does not provide an explicit methodology for maintaining the provenance. In the future, we intend to propose a provenance model for this framework by using some generic and standard provenance management approaches such as PROV ontology or CASE. However, it is observed by the PROV ontology is not flexible enough to manage the input, outputs, and resultant actions. While CASE seems more appropriate for social media forensics because it is explicitly representing the relationships between entities, also, it is offering the needed flexibility to manage input or output and actions in cyber investigations. We intend to align the global ontology with CASE for managing the provenance records; it may also provide interoperability with other tools in the domain.

Acknowledgement

This study is partially funded by Universiti Sains Malaysia RUI grant, Account No. [1001/PKOMP/8014017].

References

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A.E., Arshad, H., 2018. State-of-the-art in artificial neural network applications: a survey. Heliyon 4 (11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938.

Adam Belsher, n.d., 2016. The Growth of Digital Evidence Backlogs and Making Them a Thing of the Past - Magnet Forensics Inc [WWW Document]. accessed 1.20.17. https://www.magnetforensics.com/blog/growth-digital-evidence-backlogs-making-thing-past/.

Al Mutawa, N., Al Awadhi, I., Baggili, I., Marrington, A., 2011. Forensic artifacts of Facebook's instant messaging service. In: 6th International Conference on Internet Technology and Secured Transactions, pp. 771–776.

Al Mutawa, N., Baggili, I., Marrington, A., 2012. Forensic analysis of social networking applications on mobile devices. Digit. Invest. 9. https://doi.org/10.1016/j.diin.2012.05.007.

Alami, S., Elbeqqali, O., 2015. Cybercrime profiling: text mining techniques to detect and predict criminal activities in microblog posts. In: 2015 10th International Conference on Intelligent Systems: Theories and Applications, SITA 2015. IEEE, pp. 1–5. https://doi.org/10.1109/SITA.2015.7358435.

Aledavood, T., Lehmann, S., Saramäki, J., 2015. On the digital daily cycles of individuals. Front. Physiol. 3. https://doi.org/10.3389/fphy.2015.00073.

Aledavood, T., Lehmann, S., Saramäki, J., 2018. Social network differences of chronotypes identified from mobile phone data. EPJ Data Sci 7. https://doi.org/10.1140/epjds/s13688-018-0174-4.

Arshad, H., Jantan, A. Bin, Abiodun, O.I., 2018. Digital forensics: review of issues in scientific validation of digital evidence. J. Inf. Process. Syst. 14 (2), 126–138.

Arshad, H., Jantan, A., Omolara, E., 2019. Evidence collection and forensics on social networks: research challenges and directions. Digit. Invest. 28, 126–138. https://doi.org/10.1016/j.diin.2019.02.001.

Bader, M., Baggili, I., 2010. iPhone 3GS Forensics : logical analysis using apple iTunes backup utility. Small scale digit. Device Forensics J 4, 1–15.

Bojars, U., Breslin, J.G., Peristeras, V., Tummarello, G., Decker, S., 2008. Interlinking the social web with semantics. IEEE Intell. Syst. 23, 29–40. https://doi.org/10.1109/MIS.2008.50.

Breslin, J.G., Harth, A., Bojars, U., Decker, S., 2005. Towards semantically-interlinked online Communities. Semant. Web Res. Appl 500–514. https://doi.org/10.1007/11431053_34.

Breslin, J., Bojars, U., Passant, A., Fernandez, S., Decker, S., 2009. Sioc: content exchange and semantic interoperability between social networks. W3C Work. Futur. Soc. Netw. 15–16.

Burnett, Edmond, 2016. eDiscovery Costs: Social Media Evidence - WebPreserver.Com [WWW Document]. webpreserver.Com accessed 2.6.17. https://webpreserver.com/ediscovery-costs/.

Casey, E., Rose, C.W., 2010. Forensic analysis. In: Handbook of Digital Forensics and Investigation. Elsevier, pp. 21–62. https://doi.org/10.1016/B978-0-12-374267-4.00002-1.

Casey, E., Back, G., Barnum, S., 2015. Leveraging CybOX™ to standardize representation and exchange of digital forensic information. Digit. Invest. 12,

S102–S110. https://doi.org/10.1016/j.diin.2015.01.014.
Casey, E., Barnum, S., Griffith, R., Snyder, J., van Beek, H., Nelson, A., 2017. Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language. Digit. Invest. 22, 14–45. https://doi.org/10.1016/j.diin.2017.08.002.
Chabot, Y., Bertaux, A., Nicolle, C., Kechadi, T., 2015. An ontology-based approach for the reconstruction and analysis of digital incidents timelines. Digit. Invest. 15, 83–100. https://doi.org/10.1016/j.diin.2015.07.005.
Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., 2017. Mean birds: detecting aggression and bullying on twitter. Proc. 2017 ACM Web Sci. Conf. - WebSci '17 13–22. https://doi.org/10.1145/3091478.3091487.
Chau, D.H., Pandit, S., Wang, S., Faloutsos, C., 2007. Parallel crawling for online social networks. In: Proceedings of the 16th International Conference on World Wide Web - WWW '07. ACM Press, New York, New York, USA, p. 1283. https://doi.org/10.1145/1242572.1242809.
Cho, J., Garcia-Molina, H., 2002. Parallel crawlers. In: WWW Proceedings of the 11th International Conference on World Wide Web, pp. 124–135. https://doi.org/10.1145/511463.511464.
Cruz, I.F., Xiao, H., 2003. Using a layered approach for interoperability on the semantic Web. In: Proceedings of the 7th International Conference on Properties and Applications of Dielectric Materials (Cat. No.03CH37417. IEEE Comput. Soc, pp. 221–231. https://doi.org/10.1109/WISE.2003.1254485.
Cusack, B., Son, J., 2012. Evidence examination tools for social networks. In: Proceedings of the 10th Austrailian Digital Forensics Conference. Novotel Langley Hotel, pp. 33–40. https://doi.org/10.4225/75/57b3afc1fb861.
Dadvar, M., Trieschnigg, D., Ordelman, R., De Jong, F., 2013. Improving cyberbullying detection with user context. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 693–696. https://doi.org/10.1007/978-3-642-36973-5_62.
Dani, H., Li, J., Liu, H., 2017. Sentiment informed cyberbullying detection in social media. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
Delavallade, T., Bertrand, P., Thouvenot, V., 2017. Extracting future crime indicators from social media. In: Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime. Springer International Publishing, Cham, pp. 167–198. https://doi.org/10.1007/978-3-319-52703-1_8.
Di Capua, M., Di Nardo, E., Petrosino, A., 2017. Unsupervised cyber bullying detection in social networks. In: Proceedings - International Conference on Pattern Recognition. IEEE, pp. 432–437. https://doi.org/10.1109/ICPR.2016.7899672.
Ding, C., Chen, Y., Fu, X., 2013. Crowd crawling: towards collaborative data collection for large-scale online social networks. In: First ACM Conference on Online Social Networks. ACM Press, New York, New York, USA, pp. 183–188. https://doi.org/10.1145/2512938.2512958.
Fasching, D., Kaliner, S., Karel, T., July 2012. Social media data preservation tools and best practices. Law J. Newsletters 29 (3). LJN's Legal Tech Newsletter. https://aleph-archives.com/pub/Social%20Media%20Data%20Preservation%20-%20Tools%20and%20Best%20Practices.pdf.
GibsonDunn, 2015. 2015 MID-YEAR E-DISCOVERY UPDATE Progress on Some Fronts, but Significant Dangers Remain, and New Dangers Emerge.
Glavic, B., Siddique, J., Andritsos, P., Miller, R.J., 2013. Provenance for data mining. Proceedings of the 5th USENIX Conference on Theory and Practice of Provenance, p. 5.
Greenwood, S., Perrin, A., Duggan, M., 2016. Demographics of Social Media Users in 2016 | Pew Research Center [WWW Document] accessed 11.28.17. http://www.pewinternet.org/2016/11/11/social-media-update-2016/.
Han, F., 2016. Cloud based forensics framework for social networks and A case study on reasoning links between nodes. J. Futur. Gener. Commun 9, 23–34. https://doi.org/10.14257/ijfgcn.2016.9.1.03.
Hon, L.C., Varathan, K.D., 2015. Cyberbullying detection system on twitter. Int. J. Inf. Syst. Eng. 1.
Huber, M., 2012. Social snapshot framework: crime investigation on online social networks. ERCIM News 90, 28.
Huber, M., Mulazzani, M., Leithner, M., Schrittwieser, S., Wondracek, G., Weippl, E., 2011. Social snapshots: digital forensics for online social networks. In: Proceedings of the 27th Annual Computer Security Applications Conference. ACM Press, New York, New York, USA, pp. 113–122. https://doi.org/10.1145/2076732.2076748.
Janowicz, K., Lohmann, S., Negru, S., Haag, F., Ertl, T., n.d. Visualizing Ontologies with VOWL.
Kalemi, E., Yildirim-Yayilgan, S., 2016. Ontologies for social media digital evidence. Int. J. Comput. Electr. Autom. Control Inf. Eng. 10, 335–340.
Kastrati, Z., Imran, A.S., Yildirim-Yayilgan, S., Dalipi, F., 2015. Analysis of online social networks posts to investigate suspects using SEMCON. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Cham, pp. 148–157. https://doi.org/10.1007/978-3-319-20367-6_16.
Lessard, J., Kessler, G.C., 2010. Android Forensics : simplifying cell phone examinations. Small scale digit. Device Forensics J 4, 1–12. https://doi.org/10.1.1.185.698.
Lillis, D., Becker, B., O'Sullivan, T., Scanlon, M., 2016. Current challenges and future research areas for digital forensic investigation. Proc. 11th Annu. ADFSL Conf. Digit. Forensics, Secur. Law (CDFSL 2016) 9–20. https://doi.org/10.13140/RG.2.2.34898.76489.
Majeed, A., Zia, H., Imran, R., Saleem, S., 2015. Forensic analysis of three social media apps in windows 10. In: 2015 12th International Conference on High-Capacity Optical Networks and Enabling/Emerging Technologies (HONET). IEEE, pp. 1–5. https://doi.org/10.1109/HONET.2015.7395419.
MITRE, 2014. Cyber Observable eXpression — CybOX ™ A Structured Language for Cyber Observables 2.
Mulazzani, M., Huber, M., Weippl, E., 2012. Social network Forensics : tapping the data pool of social networks. In: Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics.
Mund, B., 2017. Social media searches and the reasonable expectation of privacy. Yale JL Tech 19, 238–238.
Murphy, J., Fontecilla, A., 2013. Social media evidence in government investigations and criminal proceedings: a frontier of new legal issues. Rich. JL Tech. XIX, 1–30.
Nahar, V., Unankard, S., Li, X., Pang, C., 2012. Sentiment analysis for effective detection of cyber bullying. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 767–774. https://doi.org/10.1007/978-3-642-29253-8_75.
Paddock, Christopher, 2016. The Irony of Privacy Settings: Can Lawyers Use Social Media Posts in a Court of Law? - Intella Blog.
Patzakis, J., 2012. Published Cases Involving Social Media Evidence [WWW Document]. eDiscovery Law Tech Blog, p. 689 (accessed 2.7.17. https://articles.forensicfocus.com/2012/04/16/689-published-cases-involving-social-media-evidence-with-full-case-listing/.
Patzakis, John, 2016. Hundreds of Thousands of Legal Cases Estimated to Address Social Media in 2016 [WWW Document]. x1discovery.Com. https://blog.x1discovery.com/2016/08/31/hundreds-of-thousands-of-legal-cases-estimated-to-address-social-media-in-2016/.
Psallidas, F., Ntoulas, A., Delis, A., 2013. SocWeb: efficient monitoring of social network activities. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 118–136. https://doi.org/10.1007/978-3-642-41154-0_9.
Randler, C., Wolfgang, L., Matt, K., Demirhan, E., Horzum, M.B., Beşoluk, Ş., 2016. Smartphone addiction proneness in relation to sleep and morningness–eveningness in German adolescents. J. Behav. Addict. 5, 465–473. https://doi.org/10.1556/2006.5.2016.056.
Recchia, Maurice, 2018. Court of Appeals Declares Facebook "Private Data" and Other Social Media Subject to Discovery | New York Law Journal. www.law.com.
Sandra Dinora, O.-J., Graciela, V.-Á., 2014. Validity and reliability in the assessment of the vulnerability of social networks. Ing. Invest. Tecnol. 15, 585–592. https://doi.org/10.1016/S1405-7743(14)70656-0.
Schatz, B., Mohay, G., Clark, A., 2004a. Generalising event forensics across multiple domains. In: Australian Computer Network and Information Forensics Conference, pp. 1–9.
Schatz, B., Mohay, G., Clark, A., 2004b. Rich event representation for computer forensics. Asia Pacific Ind. Eng. Manag. Syst. APIEMS 2004 1–200416.
Seigfried-Spellar, K.C., Leshney, S.C., 2015. The intersection between social media, crime, and digital forensics: #WhoDunIt?. In: Digital Forensics: Threatscape and Best Practices. Elsevier, pp. 59–67. https://doi.org/10.1016/B978-0-12-804526-8.00004-6.
Srinandhini, B., Sheeba, J.I., 2015. Online social network bullying detection using intelligence techniques. In: Procedia Computer Science. Elsevier, pp. 485–492. In: https://doi.org/10.1016/j.procs.2015.03.085.
Taylor, M., Haggerty, J., Gresty, D., Almond, P., Berry, T., 2014. Forensic investigation of social networking applications. Netw. Secur. 2014 (11), 9–16. https://doi.org/10.1016/S1353-4858(14)70112-6.
Turnbull, B., Randhawa, S., 2015. Automated event and social network extraction from digital evidence sources with ontological mapping. Digit. Invest. 13, 94–106. https://doi.org/10.1016/j.diin.2015.04.004.
Van Royen, K., Poels, K., Daelemans, W., Vandebosch, H., 2014. Automatic monitoring of cyberbullying on social networking sites: from technological feasibility to desirability. Telematics Inf. 32, 89–97. https://doi.org/10.1016/j.tele.2014.04.002.
Viviani, M., Pasi, G., 2017. Credibility in social media: opinions, news, and health information—a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov 7 (5), e1209. https://doi.org/10.1002/widm.1209.
Walnycky, D., Baggili, I., Marrington, A., Breitinger, F., Moore, J., 2015. Network and device forensic analysis of android social-messaging applications. Digit. Invest. 14, 77–84.
Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P.N., Zhao, B.Y., 2009. User interactions in social networks and their implications. In: Proceedings of the Fourth ACM European Conference on Computer Systems - EuroSys '09, p. 205. https://doi.org/10.1145/1519065.1519089.
Wong, K., Researcher, S., Lai, A.C.T., Yeung, J.C.K., Lee, W.L., 2013. Facebook forensics. J. Infect. Dis. 208, NP. https://doi.org/10.1093/infdis/jis918.
Wong, C.-I., Wong, K.-Y., Ng, K.-W., Fan, W., Yeung, K.-H., de Luis Gonzaga Gomes, R., 2014. Design of a crawler for online social networks analysis. WSEAS Trans. Commun. 13, 263–274.