IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 2, Ser. IV (Mar - Apr 2019), PP 12-18
www.iosrjournals.org

A Digital Forensics Framework for Facebook Activity Logs

Youssef Bassil
LACSC – Lebanese Association for Computational Sciences
Registered under No. 957, 2011, Beirut, Lebanon
Corresponding Author: Youssef Bassil

Abstract: *Facebook is one of the most widely used social networks with over two billion active users. According to recent surveys, five new users are created every second on Facebook, of which 3.6% are fake. Fake users are generally created for hiding people's real identity, nonetheless, they are sometimes created to commit illegal activities and cybercrimes. Facebook has lately introduced to their platform a feature called "Activity Logs". It is a tool that lists all online activities performed by a particular user on his Facebook account including posts, comments, likes, tags, friends added, connections made, location visited, and people searched for.As a result, Facebook Activity Logs can represent valuable forensics evidences as they maintain a history of online user's behavior. This paper proposes a framework for formalizing, processing, and analyzing Facebook Activity Logs in a digital forensics context. It comprises four processes: 1) an ontology which formally represents the knowledge contained in the Facebook Activity Logs domain using OWL or RDF, 2) an automated data extractor which extracts Activity Logs data into structured XML datasets, 3) a data visualization model with data mining and Social network analysis (SNA) features which discovers intelligence, patterns, and trends from the digital evidences extracted from the Facebook Activity Logs, and 5) a query language which provides abroad retrievalcapabilities for searching the acquired digital evidences. The experiments conducted demonstrated how the Vector Space Model and the Cosine Similarity metric can be used to classify Facebook users' comments as either malicious or innocent.*

Date of Submission: 03-04-2019 Date of acceptance: 18-04-2019

I. Introduction

Social networking has undeniably invaded all aspects of our everyday life, and has consequentlyestablished new ways for online communication, collaboration, and data sharing [1].As a result, countless Internet-based social communities have been developedto share and disseminate ideas, opinions, backgrounds, interests, and real-life connections [2]. In fact, social networking applications are nowadays consideredtrendy and free of charge virtual platformsthat support freedom of expression and information allowing people to experiment with who they are and to publicize their abandoned voice with other people, colleagues, and friends. Facebook is by far the most widely used social network with over two billion active users [3]. Basically, Facebook is a website that allows usersto sign-up for free accounts and connect with friends, colleagues and even with people they don't know, over the Internet. Facebook allows users to post and share their pictures, music and video clips, articles, hyperlinks, messages as well as their own thoughts and opinions with however many people they like. According to recent surveys, five new user profilesare created every second on Facebook, of which 3.6% are fake [4]. In actual fact, most of the people who create fake users do it either for fun, for hiding their real identity, or for imitating other real persons. However,some of themcreatefake accounts to commit illegal activitiesand cybercrimes such as espionage, fraud, terrorism, bullying, illicit trade, and counterfeiting [5]. For this reason, Facebookhas introduceda new feature to their platform called "Activity Logs".It is a toolthat lists all online activities performed by a particular user on his Facebook account including posts, comments, likes, tags, friends added, connections made, location visited, people searched for, and much more details from present time back to the very beginning when the account was first created.In effect, the Activity Logs can maintain a trace or history of online user's behavior on Facebook, which couldconstitute valuable forensics evidences to convictcriminal actions and malicious users.

This paper proposes a framework for formalizing, processing, and analyzing Facebook Activity Logs in a digital forensics context. It comprises four processes: 1) an ontology whose purpose is to formally represent the knowledge that exists in the Facebook Activity Logs using machine-interpretable specifications such as OWL or RDF, 2) an automated data extractor whose purpose is to crawl one's Facebook account and extract data into a structured dataset implemented using XML, 3) a data visualization model with data mining andSocial network analysis (SNA) features whose purpose is todiscover intelligence and draw conclusions from the digital evidence extracted from Facebook data, and 5) a query language whose purpose is to provide comprehensive means for searching the acquired digital evidence.

II. Related Work

To the best of our knowledge there has not yet been any work on forensics analysis of Facebook Activity Logs. Furthermore, there has not been any complete work on creating ontologies and visual models for social networks. In fact, popular and commercial forensics software such as EnCase [6] and FTK [7] do not handle Facebook activity logs nor they provide any means to process and analyze data from social networks. On the other hand, there have been several attempts in the academic domain to develop ontologies for social networks. For instance, [8] proposed the FOAF (Friend of a Friend) ontology which is a machine-readable descriptive vocabulary ontology based on RDF and OWL languages. It describes people and their relationships with each other. Every person can have an FOAF document called a social profile that describes himself and his connections to his friends. [9] presented the Atom Activity Streams ontology which is based on RDF vocabulary and is meant to represent the kind of activities that occur in social networking sites and applications such as Facebook and MySpace. It is able to model several social entities including friends, events, annotations, posts, likes, and others. [10] presented a standard user profiling ontology for social networking. Its purpose is to model the static profile of users independently of any kind of domain or application. The social information included are user's id, name, telephone, email, among other preferences. [11] proposed UPOS (User-Profile Ontology with Situation-Dependent Preferences Support) which is an OWL-based user profile ontology for addressing both static user data and context-aware data. Static data are personal data including user's name, telephone, address, email, interests, and preferences; while, context-aware data include relationships with friends, interactions and mutual activities. [12] proposed an ontology for security access control in a social networking environment based on semantic web standards which allows users of a social networking service to represent access control policies on their related information. This provides an authority model that dictates which user's policies are operative on what private resources.In a related context, [13] proposed a set of guidelines and an investigatory digital forensics model for online social networks. The model comprises four phases: a preliminary phase in which investigators plan the strategy that will be applied in the proceeding processes; an investigation phase in which potential digital evidences are collected and stored; an analysis phase in which the acquired evidences are examined and analyzed in order to determine its value and impact; and an evaluation phase in which investigators can present their evidences through documentation and report.As for social data collection and analysis, [14] proposed a method for automating the forensics extraction and analysis of social network user data. It is based on crawling public information for a particular user and gathering important information into data sources. Afterwards, these data sources can be used to generate graphs helpful for a forensics examiner. Likewise, [15] presented a novel method for harvesting data from Facebook user pages. The approach uses a combination of web crawlers and web services to scrap user data and save them into datasets. Data include but not limited to personal data, photos, music, and posts. Basically, these two proposals lack many features, one of which is that they both do not define an ontology nor a metadata specification for the extracted social data, in addition that they do not deliver methods to analyze, search, and mine the extracted social data. A last lacking feature to mention is that these approaches do not work on Facebook Activity Logs but only on the public pages of Facebook users.

III. Proposed Solution

This paperproposes a framework for formalizing, processing, and analyzing Facebook Activity Logs in a digital forensics context. It is composed of fourmodules: 1) an ontology, 2) an automated data extractor, 3) a data visualization model with data mining features, and 5) a query language.

Ontology: It is a formal representation for the knowledge that existsin the Facebook Activity Logs. It models basic concepts in the domain and relations among them using machine-interpretable specifications such asOWLorRDF. The process of developing this ontology requires defining classes of the ontology;organizing them in a taxonomic hierarchy; defining their various properties, roles, and constraints; and describing possible interactions among them [16]. Obviously, many classes can be defined for the Facebook Activity Logs domain. Below is a preliminary list of these classes:

- Searching activities including searching for users, groups, and pages
- Befriending, Unfriending,and Friend Requests
- Posting and Commenting including personal posts, posts by others, and posts tagged in
- Referring friends
- Likes activities including liking posts, comments, photos, status, pages, groups, and favorites.
- Posting photos, videos, and music
- Tagging and tagged in
- Updating status
- Check-in details

- Updating personal data including name, email, school, job, cover picture, favorites, and others
- Updating timeline
- Following/Unfollowing activities
- Sharing links
- Apps activities
- Invitations to groups
- Groups and Events activities

As an example, Facebook's searching activities can be modeled using an RDF ontology as follows:

<rdf:Descriptionrdf:about="User Subject">
<Searching item predicate>Searched object</Searching item predicate>
<rdf:Description>
<rdf:Descriptionrdf:about="George Smith">
<rdf:users>
<user>Tony Davis</user>
<user>John Rihana</user>
</rdf:users>
<rdf:pages>
<page>Discovery Channel</page>
<page>Sci Planet</page>
</ rdf:pages>
<rdf:groups>
<group>Tech Trends</group>
<group>Tech Future</group>
</ rdf:groups>
</rdf:Description>

Automated Data Extractor: It is an automated scraper software that crawls one's Facebook account, then parses the content of his Activity Logs and extracts data into a structured dataset. The dataset can be implemented using various technologies including relational database, XML, and spreadsheet to mention a few. The scraper has to be authenticated with the Facebook account that needs to be crawled and thus a user's credentials must be obtained beforehand.Facebook has already released the Graph API for developers to ease the interaction between software applications and Facebook platform [17]. This REST-based API can be used to assist the data extractor. It would permit retrieving social interactions and events among a variety of other tasks that are valuable to forensics investigators. As a result, extracting Facebook Activity Logs can be done reliably and seamlessly via the Graph API.

Data Visualization Model: It is a graph-based modelrepresenting the Facebook Activity Logs data as a network diagram. It consists of vertices representing individual social entities within the network such as users, posts, comments, photos, etc; and edges representing the different interactions and relationships between these social entities such as friendship, commenting, posting, liking, tagging, etc. Relationships within the graph can be represented by an adjacency matrix which indicates if a relationship exists between two entities. Figure 1 depicts a Graph model displaying the interaction between different objects on Facebook.

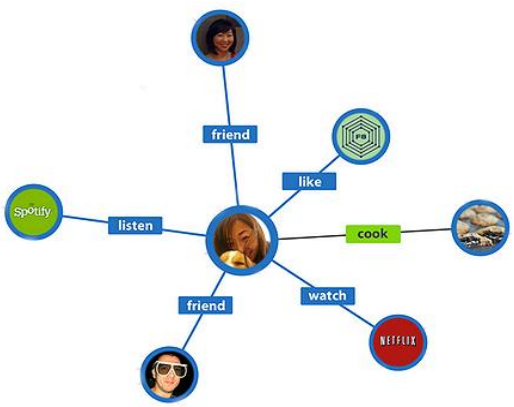

Figure 1: Graph Model for Facebook Activity Logs

The graph-based data visualization model has many benefits, one of which is conveying complex textual unstructured information into structured, intuitive, and interactive information, reinforcing human cognition, interpretation, and reasoning skills. Another benefit for the graph model is that it supports social network analysis (SNA) which can be used to examine, analyze, and measure the relationships between social entities, and explore associations between them. Fundamentally, Social Network Analysis (SNA) is the study and analysis of online social networks [18]. It entails analyzing social relationships (friends, membership, ownership, etc) and social interactions (liking, posting, searching, etc) between the different actors within the network. These networks are often depicted in a graphical network diagram, where actors are represented as vertices or nodes and interactions are represented as lines or edges. Social Network Analysis is often carried out by means of data mining models and algorithms such as Classifications, Clustering, Association Rules, and Anomaly Detection [19].

In essence, some of the data mining techniques that can be exploited in Social Network Analysis include calculating the Similarity metric which finds similarities between entities based on parameters such as nationality, gender, location, common posts, similar comments, similar status, or any other significant characteristics. It can eventually reveal how much the nodes in the network tend to cluster together i.e. share certain characteristics. As such, clustering techniques can be used to group together posts and comments sharing common wordings, or create clusters for likes pertaining to similar entities, or grouping searches and inquiries related to the same topic. Another technique that can be applied on graphs is the Clustering-Coefficient metric which is a measure of degree to which nodes in a graph tend to cluster together i.e. share certain characteristics. Both of these metrics complement each other and they often employ a term weighting scheme such as the TF-IDF [20] (Term Frequency - Inverse Document Frequency) and a similarity equation such as the Cosine metric [21] to discover closeness between various entities in the social network and cluster them into groups that share common content. For instance, Similarity and Clustering-Coefficient metrics can be used to cluster together posts and comments sharing common wordings, or create clusters for likes pertaining to similar entities, or grouping searches and inquiries related to the same topic. Formally, the Cosine metric can be calculated using the following equation:

$$sim(d_j, q) = \cos(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j || \vec{q}|} = \frac{\sum_{i=1}^{N} Wi,q \times Wi,j}{\sqrt{\sum_{i=1}^{N} W_{i,q}^2} \times \sqrt{\sum_{i=1}^{N} W_{i,j}^2}}$$

Basically, in order to determine the similarities between two documents d_1 and d_2, the cosine of the angles between d_1 and d_2 is calculated. If the two documents are alike, they will receive a cosine of 1; if they are orthogonal (sharing no common terms), they will receive a cosine of 0.

Another data mining models that can explored in Social Network Analysis are probabilistic models which have the benefit in finding statistics on network entities. Some examples may include finding how many friends a given user has, finding the closest friends, the highest referred friends, the total number of comments, the total number of tags, the most frequent activities, etc.Furthermore, the detection of anomalies is yet another data mining technique which can be used in Social Network Analysis. Inherently, Anomaly Detection is the process of identifying items and events which do conform to a certain pattern. Such pattern could be grammatical error, malicious code, malicious attack, offensive statement, or email containing threatening, violence, or aggressive messages [22]. When analyzing Facebook Activity Logs, a list of abnormal interactions/patterns (such as liking certain suspicious pages, posting certain noisy words, searching for certain malicious users) can be compiled and fed to the system which in turn can detect and pin point outliers that conform to these patterns. In fact, this can be considered as a form of supervised classification in which pattern of interests are available prior to starting identifying input data as either noisy or not.

Query Language: It provides comprehensive searching capabilities over the visualization model. It could be used to extract sub-graphs from a large network of data such as extracting a user and all its directly related comments and likes. Another use of the query language is to search the network, a sub-graph within the network, or even a particular entity for a certain keyword or substring. Another usage is finding detailed information about a certain event such as finding the check-in location of a user who commented on a suspicious post at a given time, or finding all friends of a user who posted a suspicious photo on his wall. The query language can also be used to traverse the network and find a connected sequence of events, for instance, finding the path that connects a user's comment with one of his friend's post. Implementation-wise the query language could be visual, based on GUI controls or textual with special syntax and vocabulary such as SQL or LINQ.

IV. Experiments& Results

In the experiments, we will to try to classify a Facebook user's comments using the Vector Space Model and the Cosine Similarity metric. The aim of this process is to come up with a conclusion whether or notthis user has any comments involved with drug trafficking. First, the automated data extractor processes the Activity Logs of the user under investigation and extracts all his comments into a structured dataset built using XML. Below is a sample showing the user's comments in XML format:

```
<facebook>
<user1>Tony Davis</user1>
<data>
<comments>
<comment user='Tino B'>Let's meet at 11</comment>
<comment user='John James'>Nice picture, you look awesome</comment>
<comment user='Bobby Smith'>Cheap Marijuana and Drug</comment>
<comment user='Tino B'>What a day, I am tired…</comment>
</comments >
<likes>
        …..
</likes >
<friendsRequests>
        …..
</friendsRequests>
        …..
        …..
        …..
</comments>
</data>
</facebook>
```

Second, a collection of 100 newspaper articles all pertaining to drug trafficking were compiled. Then a sample article document *d*is selected out of the collection, and all its words and terms are modeled and weighted using the *VSM* model and the *TF-IDF* scheme. Table 1 delineates the results obtained after applying the *TF-IDF* weighting scheme on document *d*.

Table 1: TF-IDFResults

Term	TF	IDF	TF*IDF
Drug	83	0.008	0.66
Trafficking	11	0.19	2.09
Cocaine	30	0.09	2.70
Heroin	26	0.12	3.12
Money	32	0.14	4.48
Cartel	10	0.51	5.10
Weed	37	0.32	11.84
Marijuana	66	1.52	100.32
…			

Subsequently, the document *d* can be represented using the *VSM* model as:

d = (*w*(*Drug*), *w*(*Trafficking*), *w*(*Cocaine*), *w*(*Heroin*), *w*(*Money*), *w*(*Cartel*), *w*(*Weed*), *w*(*Marijuana*))
d = (*0.66*, *2.09*,*2.70*, *3.12*,*4.48*, *5.10*, *11.84*, *100.32*)

Third,a user comment is selected from XML denoted by *q*="marijuanaand drugs weed". It is mainly composed of three terms and can be thus represented using *VSM* as:

q = (*w*(*marijuana*), *w*(*drug*), *w*(*weed*))
q = (*1*, *1*,*1*)

Fourth, in orderto classify comment *q*under the 'Drug Trafficking' category, *q*has to be closely related to document *d*, that is *q* and *d* should share some common terms. For this purpose, the cosine similarity metric is computed. If *q* and *d* are alike, they will receive a cosine of 1; if they are orthogonal (sharing no common terms), they will receive a cosine of 0. Below are the results of calculating the cosine metric for *q* and *d*:

$$|\vec{d_j}| = \sqrt{\sum_{i=1}^{N} w_{i,j}^2} = \sqrt{0.66^2+2.09^2+2.70^2+3.12^2+4.48^2+5.10^2+11.84^2+100.32^2}$$

$$= \sqrt{10{,}272.19} = 101.35$$

$$|\vec{q}| = \sqrt{\sum_{i=1}^{N} w_{i,q}^2} = \sqrt{1^2+1^2+1^2}$$

$$= \sqrt{3} = 1.73$$

$$\vec{d_j} \bullet \vec{q} = 100.32*1 + 0.66*1 + 11.84*1 = 112.82$$

$$\frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}||\vec{q}|} = 112.82 / 101.35 * 1.73 = 0.64$$

The above results clearly show that calculatingthe cosine metric over *q* and *d* yields a value of 0.64, and thus a strong match exists between *q* and *d*since 0.64 is closer to 1 than to 0 (the more the cosine value is close to 1, the more are the two vectors similar). As final verdict, the comment *q*="marijuana and drugs weed" is classified under the "Drug Trafficking" category.

V. Conclusions

Facebook Activity Logs is a remarkably important area that can disclose massive amount of information regarding the online social behavior of Facebook users. Forensics investigators can dig into the activity logs and find valuableinformation about a certain user including such information as his searches,posts, check-in, preferences, likes, tags, favors, and much more. All these social activities can be considered as electronic traces that maycontain incriminating evidences against convicts. The proposed solution delivers four vital modules: An ontology, an automated data extractor, a data visualization model coupled with data mining capabilities, and a query language. Their advantages can be summarized as follows:

Ontology is one of the most popular methods to formally model knowledge of a domain. It makes it possible to formalize domain concepts and the relations between them and the operations that can be performed on them in a structured way. Furthermore, ontology supports information exchange processes and standardizes the communication and cooperation between systems developed at different sites, by different researchers, and using different platforms.

Automated Data Extractor helps in automating processes making data extraction easier, faster, and more accurate. Besides, the output of the extractor is a dataset implemented using a standardtechnology. This would support information exchange between different platforms and untie data representation from data implementation.

Data Visualization Model provides an intuitive, scientific, and process-able representation of data in the form of a graph network. Its advantage is that it allows several artificial intelligence, probabilistic, and statistical computations to be performed on the various entities of the network. These types of computations allows discovering new trends and patterns from the pairwise relations and interactions of social entities and create associations between them, which assist in the examination, analysis, and decision-making of information. As for Social Network Analysis (SNA) and data mining, it provides an intuitive, scientific, and process-able representation of data in the form of a social graph network. Its advantage is that it allows several data mining techniques to be executed on the various entities of the network such as clustering and anomaly detection. Such techniques allow discovering new trends and patterns from the pairwise relations and interactions of social entities and create associations between them, which assist in the examination, analysis, and decision-making of forensics information.

Query Language provides a structured and a comprehensive way to search the social network and find valuable information about various entities and social interactions.

Acknowledgments

This research was funded by the Lebanese Association for Computational Sciences (LACSC), Beirut, Lebanon, under the "Big Data Research Project – BDRP2019".

References

[1]. Liu, H., Maes, P., Davenport, G., Unraveling the Taste Fabric of Social Networks. InternationalJournal on Semantic Web and Information Systems, vol.2, no.1, pp. 42-71, 2006.
[2]. "Key Facts", Facebook Newsroom, FacebookInc, Retrieved January, 2019.
[3]. "Facebook Reports Fourth Quarter and Full Year 2018 Results", Menlo Park, Calif, Retrieved January 30, 2019
[4]. Zephoria, The Top 20 Valuable Facebook Statistics, March 2019, Url: https://zephoria.com/top-15-valuable-facebook-statistics/
[5]. Roger McNamee, "Zucked: Waking up to the Facebook Catastrophe", HarperCollins. ISBN 9780008318994, 2019
[6]. EnCase, Guidance Software, URL: www.guidancesoftware.com/encase-forensic.htm
[7]. Forensic Toolkit (FTK), AccessData, URL: www.accessdata.com/products/digital-forensics/ftk
[8]. D. Brickley, L. Miller, The Friend of a Friend (FOAF) project, FOAF Vocabulary Specification 0.98, 2012.
[9]. Libby Miller, NoTube Project, 2010.
[10]. M.Golemati, A.Katifori, C.Vassilakis, G.Lepouras,and C.Halatsis, Creating an ontology for the user profile:Method and applications, Proceedings of the First IEEE InternationalConference on Research Challenges in Information Science, 2007.
[11]. M.Sutterer, O.Droegehorn, K. David, UPOS: User profileontology with situation-dependent preferences support, Proceedingsof the First International Conference on Advances in Computer-Human Interaction, pp.230–235, 2009.
[12]. A. Masoumzadeh, J. Joshi, OSNAC: An Ontology-Based Access Control Model for Social Networking Systems, IEEE Second International Conference on Social Computing, pp.751-759, 2010.
[13]. N. Zainudin, M. Merabti, D. Llewellyn-Jones, A Digital Forensic Investigation Model for Online Social Networking, the 11th Annual Conference on the Convergence of Telecommunications, Networking & Broadcasting, 2010.
[14]. M. Mulazzani, M. Huber, E. Weippl, Social Network Forensics: Tapping the Data Pool of Social Networks, 8th Annual International Conference on Digital Forensics, 2012
[15]. M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek, E. Weippl, Social Snapshots: Digital Forensics for Online Social Networks, in Annual Computer Security Applications Conference (ACSAC), 2011
[16]. T. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, Vol. 5, pp.199-220, 1993.
[17]. Facebook Graph API, URL: https://developers.facebook.com/docs/graph-api/, Retrieved March 2019
[18]. EvelienOtte, Ronald Rousseau, "Social network analysis: a powerful strategy, also for the information sciences", Journal of Information Science, vol. 28, no. 6, pp.441–453, 2002
[19]. Jiawei Han, MichelineKamber, "Data mining: concepts and techniques", Morgan Kaufmann, ISBN 9781558604896, 2001.
[20]. G. Salton, A. Wong, and C.S. Yang, "A vector space model for Information Retrieval", Journal of the American Society for Information Science, Vol. 18, No.11, pp. 613-620, 1975.
[21]. Jurafsky D., Martin J., "Speech and Language Processing", 2nd ed, Prentice Hall, 2008.
[22]. V. Chandola, V., A. Banerjee, V. Kumar, "Anomaly detection: A survey", ACM Computing Surveys, vol. 41, no.3, pp.1–58, 2009

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Youssef Bassil. " A Digital Forensics Frameworkfor Facebook Activity Logs" IOSR Journal of Computer Engineering (IOSR-JCE) 21.2 (2019): 12-18.