Chapter 8

DATA VISUALIZATION FOR SOCIAL NETWORK FORENSICS

Martin Mulazzani, Markus Huber and Edgar Weippl

Abstract With hundreds of millions of users worldwide, forensic data extraction from social networks has become an important research problem. However, forensic data collection is tightly connected to social network operators, which leads to problems related to data completeness and data compatibility. This paper discusses the important data sources and analytical methods for the forensic analysis of social networks. It shows how the data sources can be evaluated in an automated fashion without assistance from social network operators. While the proposed methods apply to the vast majority of social networks, their feasibility is demonstrated using a Facebook case study.

Keywords: Social networks, online forensics, visualization

1. Introduction

The number of users of social networks has been increasing phenomenally – Facebook alone claims to have 800 million users [13]. Meanwhile, social networks have introduced novel research challenges with regard to digital forensics. While traditional digital forensics relies on the physical acquisition of hardware [6, 7] and the use of hash values to ensure evidence integrity, this approach does not scale to social networking and cloud services and their use of distributed data centers. Due to the lack of standardized APIs for digital forensics and unified processes for service operators, *ad hoc* isolated solutions are still in widespread use. Another major challenge is the visualization of massive volumes of data [10, 18].

It is difficult to visualize social networking data in a manner that facilitates the answering of common questions, especially by individuals with limited technical expertise. This occurred recently in the iPhone

`consolidated.db` file controversy. The file in question contained geolocation information that was publicized back in 2010 [17]. However, the `consolidated.db` problem received widespread attention after the April 2011 release of iPhoneTracker software [20], which helped visualize the collected data. The resulting controversy forced Apple to review and change its data collection process [8].

This paper identifies the data sources of interest in forensic investigations of social networks and shows how they can be leveraged in an automated fashion. The paper also describes several visualizations that can be generated from the data sources for use in forensic investigations. Specifically, example graphs and possible visualizations are presented, some of which could be used very effectively in social network analysis. To the best of our knowledge, this work is the first to support the forensic analysis of data from social networks without the collaboration of social network operators.

2. Background

Social network forensics has to rely on a limited set of data sources in many cases. Acquiring the hard drives of the servers is not feasible, and accessing the data directly requires the cooperation of the social network operator. Moreover, as described in the Facebook Law Enforcement Guidelines published by the Electronic Frontier Foundation [12], an investigator who files a data request with an operator may or may not receive all the relevant data. Network forensic systems such as PyFlag [9] and Xplico [22] also cannot access all the data, because they are passive in nature.

This situation not only hinders evidence collection [6], but also makes it difficult for an investigator to demonstrate that the evidence that is collected is authentic and reliable. The approach described in this paper stands out because it does not require the cooperation of the social network operator. Additionally, properties such as evidence authenticity and reliability are enhanced due to the open nature of the collection methodology and tools.

2.1 Data Acquisition

Before a social network can be analyzed, the relevant data has to be identified and acquired. While traditional digital forensic methods can be used to extract artifacts from the local web browser cache [19], many other approaches are possible at the communications layer. These range from passive network sniffing to active methods such as sniffing of unencrypted Wi-Fi traffic [5], possibly in combination with ARP spoofing on

a LAN. The recently proposed "friend in the middle attack" [16], which uses a third party extension for a social network in combination with a traditional crawler component, may also be used.

However, crawling is limited at best, primarily because metadata and accurate timestamps are not shown on web pages; these are only available using social network APIs. Although it is possible to use passive logging at the communications layer, e.g., when a judge has authorized lawful intercepts on the Internet communications of a suspect, this approach is limited because it would take considerable time to collect data, and completeness is hardly possible. Furthermore, many social networks permit data encryption at the communications layer using the HTTPS protocol, which renders passive attacks useless.

Facebook recently announced that its users can download all their profile data. However, the data provided by our method is far superior compared with the Facebook profile download option, which lacks important metadata (among other data) and is, therefore, not useful for social network forensics. In general, it is not possible for a user to download everything that is connected to his or her profile on the social network.

Another interesting feature is Facebook Timeline, which encourages users not to delete anything from Facebook and to use it as a historical archive. This feature will greatly benefit digital forensic investigations of Facebook accounts.

3. Social Network Data Pool

While social networks vary considerably in their architecture and features, the following generic data sources are particularly useful from the point of view of forensic investigations of social networks:

- **Social Footprint:** What is the social graph of the user? Who are the user's friends?
- **Communications Pattern:** How is the network used for communicating? What methods are used? With whom is the user communicating?
- **Pictures and Videos:** What pictures and videos were uploaded by the user? Who are the other users on whose pictures the user is tagged?
- **Activity Times:** When was the user connected to the social network? When did a specific activity take place?

- **Apps:** What apps is the user using? What is their purpose? What information can be inferred in the social context?

A plethora of information about prolific users is stored by the operator. Facebook claims that more than 50% of its users use the social network on any given day, which corresponds to around 400 million users [13]. Sometimes, information is cached locally, but this is not a reliable source because it is neither complete nor stored persistently. Depending on the social network implementation, the availability of the data itself and the possibility of retrieving the data via API calls varies considerably. Nevertheless, most of this data can be extracted directly or inferred with low overhead without the collaboration of the social network operator. Once the data is available to the investigator, the full spectrum of social network analysis can be conducted [21].

Of course, the easiest way to obtain data pertaining to a social network account is with the consent of the user, who would have to provide the username and password. However, since this data can be quite voluminous, a forensic investigator may require specialized tools to view and analyze the data.

4. Social Network Visualizations

This section describes certain basic and advanced visualizations that can be constructed from the data sources identified in the previous section. These visualizations can be used very effectively in forensic investigations of social networks.

4.1 Basic Visualizations

Four basic visualizations can be constructed from social network data sources:

- **Social Interconnection Graph:** It is trivial to retrieve the list of friends of a user. In most social networks, this is public information or information that can be collected quite easily even without entering the social circle of the user of interest [4]. However, it is not trivial to cluster the friends themselves. For example, to find out who is connected with whom; or, if a specific contact is in the cluster of workplace friends; or, if there is a direct friend relationship.

  Our approach uses a feature of the Facebook API, which allows an application to query if two users are connected. The friends of a user may be clustered into different groups, e.g., people from work, school, family and more, because the members of the groups

are more likely to know each other. This yields a undirected graph $G =< V, E >$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of friends of a user and $E = \{(v_x, v_y), \ldots\}$ is the set of edges that connects two nodes that are friends in the social network. An example graph is shown in Section 5. Clusters of highly connected nodes represent friends who know many of each other's friends.

- **Social Interaction Graph:** In many investigations, it is important to gather information about users who communicate with each other. The communications include wall posts, direct messages, group communications and reactions to public announcements. User communications can be represented as a directed graph $G =< V, E >$ where the nodes $V = \{v_1, v_2, \ldots, v_n\}$ correspond to friends and each directed edge (v_x, v_y) in E has a weight that is incremented for every message sent from v_x to v_y. Note that different forms of communication are not distinguished when assigning the weight. Allowing investigators to add custom weights to specific forms of communication is a topic of future work.

  An example social interaction graph generated by our software is shown in Section 5. The graph enables an investigator to easily identify the top communication partners at first sight, and correlate them with other information such as phone records.

- **Complete Timeline:** The timeline of activities is of increasing importance as mobile clients on smartphones enable social network users to be online all the time. In many cases, the activities of a user as well as the activities of his/her friends can be extracted. Proper visualization of the activity times of users can be very helpful to an investigator. This includes multiple data layers, such as the activities of the user, activities of friends, group activities and reactions to events by friends, which can simplify analysis. Also, a zoomable timeline for visualizing time ranges of importance can be very useful to an investigator. This is especially effective because a social network user and his/her friends could be involved in more than 500 events on a given day.

- **Location Visualization:** The increasing use of geotagging and location applications needs to be reflected in forensic investigations. Applications such as foursquare [14] and Facebook Places are employed in social networks to capture and share geolocation information. Digital cameras and smartphones automatically geotag pictures with the exact locations where the pictures were taken.

Most social networks remove metadata from pictures prior to storage [2], but this may change in the future.

4.2 Advanced Visualizations

This section describes advanced visualizations that can be constructed from social network data sources. The following three advanced visualizations are useful to investigators and should become standard tools in social network forensics:

- **Event Tracking:** Investigations of viruses and other malicious applications that propagate via social networks require the identification of who or what started the series of events. Tracking such events is not straightforward, but having a collection of social network footprints of various users can provide insight into dissemination characteristics, propagation tactics and other analytic information.

- **Timeline Matching:** In a highly centralized system such as a social network, an investigator has the benefit of consistent timestamps provided by the social network implementation. Social network operators often run their own NTP infrastructure and keep their clocks consistent across thousands of servers. The timestamps can be used to match the timelines of different users, and to create an exact timeline for an entire cluster of friends or even a larger group. This feature, which was recently proposed for NTFS [11], is of importance to social networks as well as cloud computing systems.

- **Differential Snapshots:** The forensic image of a user profile can vary considerably depending on when it was taken. Therefore, an investigator would welcome the ability to visualize the social network data of a given user as well as the differences between previous images of the same user.

5. Experimental Results

We implemented the methodology outlined in [15] to collect data from Facebook, currently the world's largest social network. The data included all the social connections, direct communications, pictures and more. The data acquisition took approximately 20 minutes per account, which, in our opinion, is a reasonable amount of time for data collection. We parsed the output and generated the graphs presented in this section.

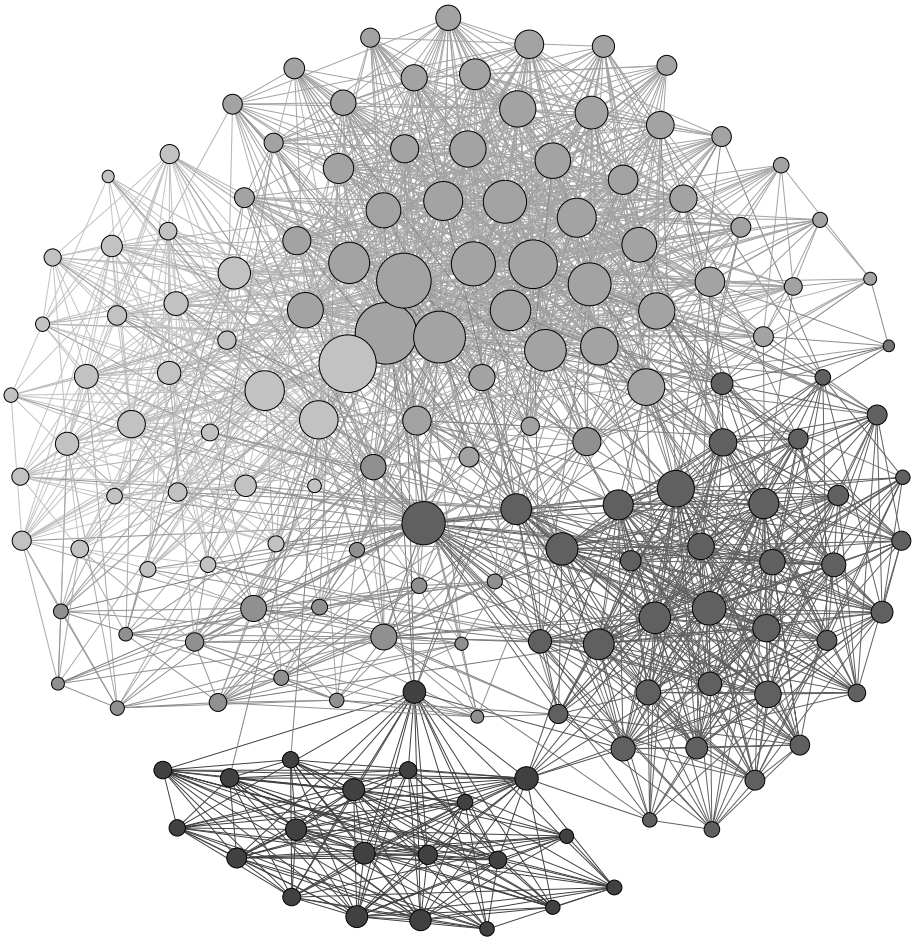

Figure 1. Anonymized social interconnection graph.

5.1 Visualization

The creation of the social interconnection graph employed a feature from the Facebook API that allows an application to query if two users are connected. To create the social interconnection graph, we tested if the first friend in a user profile is in a connection with the $n-1$ friends of the profile, then tested for the second friend with the remaining $n-2$ friends, and so on. Figure 1 shows an example social interconnection graph, which was constructed using the profile of one of the authors of this paper. The graph was plotted with Gephi [1], an open source graph visualization tool, using the Fruchterman-Reingold algorithm [3]. The nodes that appear to be from the same cluster are colored automatically,

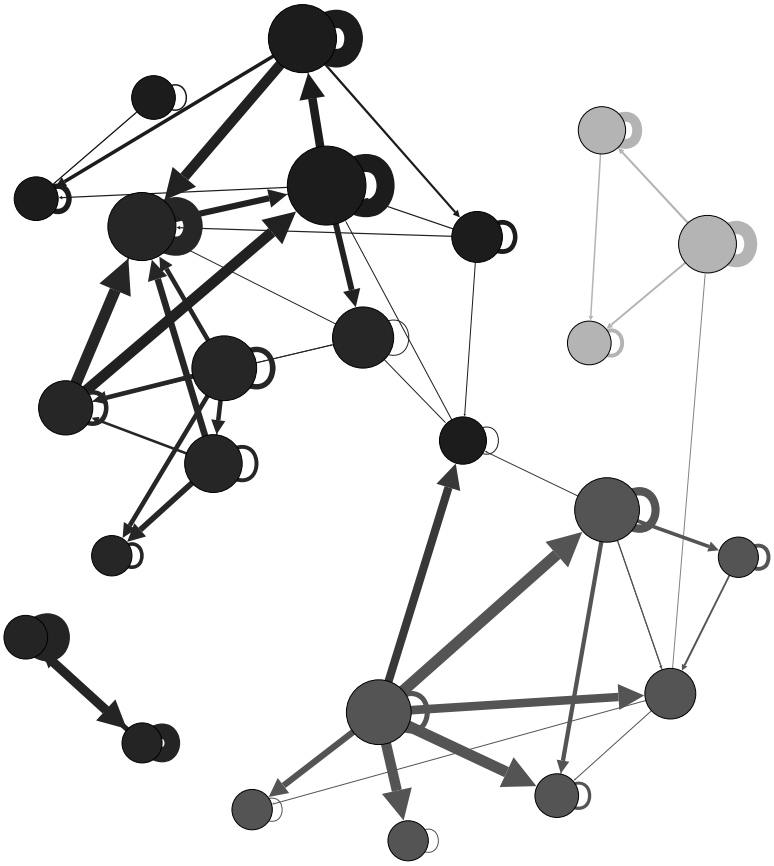

Figure 2. Anonymized social interaction graph derived from picture tags.

which simplifies cluster analysis. Note that the connections are hypothetical because the real user names are replaced by random names of computer scientists listed in Wikipedia.

Several different social interaction graphs may be created from Facebook data. Our software makes it possible to create separate graphs for different types of interaction, which could be subsequently integrated to produce a complete social interaction graph.

Figure 2 shows an example social interaction graph derived from picture tags in Facebook. The graph is created using the following three steps: (i) for a user account of interest, all the pictures of all the friends are collected and searched for tagged users; (ii) a user who is tagged in a picture but is not in the list of friends is ignored; (iii) if a tagged user is in the list of friends, then an edge is created between the two nodes (users) pointing from the user who uploaded the picture to the tagged

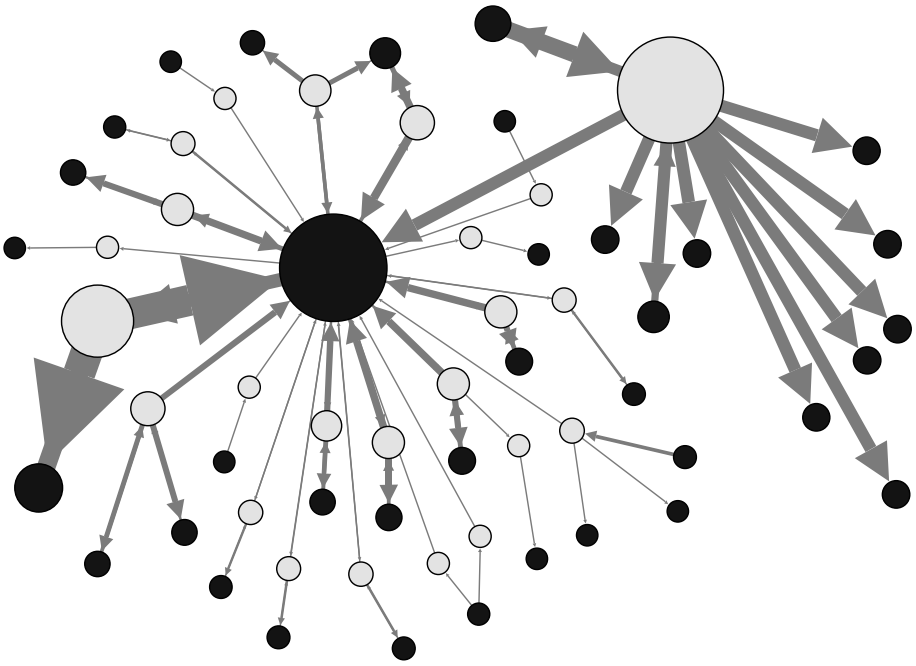

Figure 3. Social interaction graph derived from direct messages.

user, or the weight of the edge is increased if the edge already exists. Since the edges are directed and weighted, the graph can be used to identify users with "tight" social connections.

Another type of social interaction graph can be created based on direct messages (Figure 3). In this case, an edge is added between two nodes (users) if the user of interest has exchanged messages with another user. The edges are weighted according to the number of messages sent by the users.

Social network data can also be used in the automated creation of timelines. While this feature is currently under development, an example timeline of events in Facebook over a 24-hour period is shown in Figure 4.

5.2 Evaluation

The proposed data collection and visualization methods are novel, and are easily integrated with existing social network analysis methods. However, a key drawback is that data collection is hardly reproducible – the graphs and timelines are expected to change considerably over time because of the highly dynamic aspects of social networks. In a single day,

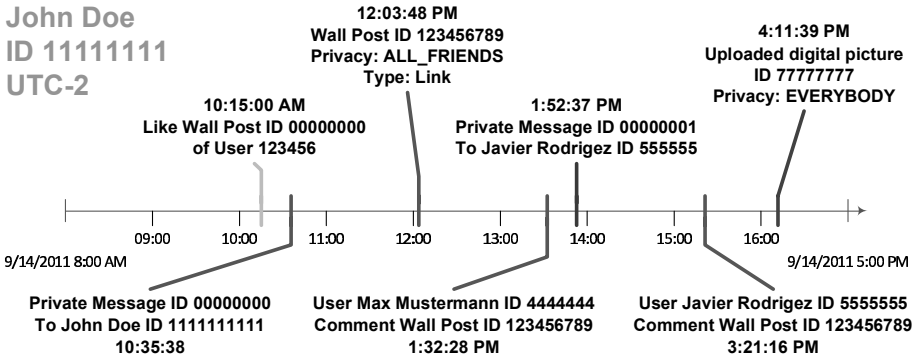

Figure 4. Anonymized timeline for a 24-hour period.

a user could change his/her social interconnection graph considerably, and could even attempt to hide communications traffic.

Other data, such as user IP addresses as provided by the Facebook NeoPrint [12], are only available to social network operators, and are not accessible with our methods (using an automated web browser or an API). Furthermore, it is difficult, if not impossible, to guarantee data completeness – once data is deleted by the user, it can only be recovered by the social network operator, so this data is not available for analysis without operator cooperation. Nevertheless, much social network data is static in nature and our methods are applicable and, indeed, very useful in forensic investigations.

6. Conclusions

Social network forensics is rapidly becoming an important component of digital forensic investigations. This paper has identified several valuable data sources and has shown how they can be leveraged in forensic investigations, even without the cooperation of social network operators. The proposed methods are fully automatable and support cluster analysis as well as timeline visualization. The proof-of-concept software system for creating social interconnection and social interaction graphs from Facebook data demonstrates the feasibility and utility of the methods in social network forensic investigations.

Our future work will implement additional visualizations and an automated report generation feature. Also, it will increase the number of social networks supported by the software system, which will be released as open source in the near future.

References

[1] M. Bastian, S. Heymann and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, pp. 361–362, 2009.

[2] D. Beaver, S. Kumar, H. Li, J. Sobel and P. Vajgel, Finding a needle in Haystack: Facebook's photo storage, *Proceedings of the Ninth USENIX Conference on Operating Systems Design and Implementation*, 2010.

[3] V. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008(10), 2008.

[4] J. Bonneau, J. Anderson, R. Anderson and F. Stajano, Eight friends are enough: Social graph approximation via public listings, *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pp. 13–18, 2009.

[5] E. Butler, Firesheep (`codebutler.com/firesheep`), 2011.

[6] D. Brezinski and T. Killalea, RFC 3227: Guidelines for Evidence Collection and Archiving (`www.faqs.org/rfcs/rfc3227.html`), 2002.

[7] B. Carrier, *File System Forensic Analysis*, Pearson, Upper Saddle River, New Jersey, 2005.

[8] B. Chen, Apple promises fix for location-gathering "bug" on iPhone, *Wired* (`www.wired.com/gadgetlab/2011/04/iphone-location-bug`), April 27, 2011.

[9] M. Cohen, PyFlag – An advanced network forensic framework, *Digital Investigation*, vol. 5(S), pp. S112–S120, 2008.

[10] G. Conti, *Security Data Visualization: Graphical Techniques for Network Analysis*, No Starch Press, San Francisco, California, 2007.

[11] X. Ding and H. Zou, Time based data forensic and cross-reference analysis, *Proceedings of the ACM Symposium on Applied Computing*, pp. 185–190, 2011.

[12] Facebook, Facebook Law Enforcement Guidelines, Menlo Park, California (`www.eff.org/sites/default/files/filenode/social_network/Facebook2010_SN_LEG-DOJ.PDF`), 2010.

[13] Facebook, Facebook Statistics, Menlo Park, California (`www.facebook.com/press/info.php?statistics`).

[14] Foursquare Labs, foursquare, New York (`foursquare.com`).

[15] M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek and E. Weippl, Social snapshots: Digital forensics for online social networks, *Proceedings of the Twenty-Seventh Annual Computer Security Applications Conference*, pp. 113–122, 2011.

[16] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler and S. Goluch, Friend-in-the-middle attacks: Exploiting social networking sites for spam, *IEEE Internet Computing*, vol. 15(3), pp. 28–34, 2011.

[17] S. Morrissey, *iOS Forensic Analysis*, Apress, New York, 2010.

[18] S. Teelink and R. Erbacher, Improving the computer forensic analysis process through visualization, *Communications of the ACM*, vol. 49(2), pp. 71–75, 2006.

[19] Trustedsignal – Blog, Facebook Artifact Parser version .02 (`trustedsignal.com/code/fbartiparse.py`), 2011.

[20] P. Warden, iPhone Tracker (`petewarden.github.com/iPhoneTracker`).

[21] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, United Kingdom, 1994.

[22] Xplico, Network Forensic Analysis Tool (`www.xplico.org`).