Social Network Forensics:

Tapping the Data Pool of Social Networks

Martin Mulazzani, Markus Huber and Edgar Weippl

Abstract

With hundreds of millions social network users worldwide, forensic data extraction on social networks has become an important research problem. The forensic data collection is however tightly connected to the service operator, which leads to data completeness and presentation format issues. Online social networks imply that all user communication is stored entirely at the service operator, without direct access for investigators. In this paper we identify important data sources and analytical methods for automated forensic analysis on social network user data. We furthermore show how these data sources can be evaluated in an automated fashion, and without the need of collaboration from the social network operator. While our proposed methods apply to the great majority of social networks, we show the feasibility of our approach on basis of Facebook.

Keywords: Online forensics, social networks, visualization

1 Introduction

With the increasing usage of social networks and the emerge of cloud computing, digital forensics faces novel research problems and challenges. The number of users of these services increases steadily, with e.g. Facebook currently claiming to have 800 million users [4]. While traditional forensics relies on the physical acquisition of hardware [13, 14] and the usage of hashsums to ensure evidence reliability, this approach does not scale to cloud services and their use of distributed datacenters. With the lack of standardized forensics APIs as well as unified processes for service operators, isolated solutions are still in widespread use. Another important aspect of forensics is the proper visualization of data [22, 16] due to the vast amount of available data.

It is furthermore hard to visualize gathered social networking data in a way that can answer common questions of interest on a first sight, so that people without technical background can understand it. This has been shown for example in the case of the *consolidated.db* from Apple's iPhone: the file contained geolocation information which has been already outlined in 2010 [21]. However, the *consolidated.db* problem got widespread attention with the release of the iPhoneTracker software [6] in April 2011, which visualized the collected data. Due to the iPhoneTracker software, Apple finally had to review and change their data collection process[1].

In this paper we identify data sources of interest for forensic examinations on social networks, and how they can be leveraged in an automated fashion. We furthermore identify graphs of interest that can be generated from these data sources and can answer many possible questions of a forensic examiner on first sight. To the best of our knowledge there has not yet been any work on forensic analysis of data from social networks without the collaboration of the social network operator. We show example graphs and present possible visualizations, which can and should be used for social network analysis and will be released under an open source license.

The rest of the paper is organized as follows: Section 2 gives an overview of social networks, and how they are already used for conducting and solving crimes. Section 3 shows what data can be extracted from social networks for social network forensic investigation, while Section 4 explains the data and how it can be visualized. We show the feasibility of our approach in Section 5 and conclude in Section 6.

2 Background

Social network forensics has to rely on a limited set of data sources in many cases. Acquiring the server's hard drives is not feasible, and leveraging the service operator's data directly requires the service operator's cooperation. If at all, the investigator can submit requests to the operator and may or

may not receive all the relevant data (e.g. written in the Facebook law enforcement guidelines published by the EFF [3]). This is in clear contradiction to the guidelines for evidence collection [13], as the investigator is unable to show that the evidence is authentic, complete, and reliable. Network forensic frameworks like PyFlag [15] and Xplico [8] simply cannot see or access all the data, as they are solely passive. Our approach, on the other hand, does not require the cooperation of the social network operator, and can ensure these properties due to the open nature of our collection methodology and tools.

2.1 Data Acquisition

Before being able to analyze social network data the data has to be gathered and acquired. While traditional forensic methods can be used to extract artifacts from local webbrowser cache [2], numerous other ways are possible on the communication layer. These range from passive sniffing on the network to active attacks like sniffing on unencrypted Wifis [20] or in combination with ARP spoofing on LANs. The recently proposed *friend in the middle attack* [19], which uses a third party extension for the social network in combination with a traditional crawler component, could be used as well. Crawling however is limited, as metadata and accurate timestamps are not shown on webpages. They are only available by using the social network APIs, which extend the available data of the webinterface. Even though it would be possible to use passive logging on the communication layer, for

example in cases where a judge ordered lawful interception on the Internet connection of a suspect, this approach is limited as well as it would take a tremendous amount of time for collecting information, and completeness is hardly possible. Furthermore, many social networks offer the possibility to encrypt data on the communication layer by using HTTPS, rendering passive attacks useless.

While Facebook announced recently that users are now able to download all their profile data, the data provided by our method is far superior compared to the Facebook profile download option which lacks e.g., important metadata and is thus not useful for 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 recently announced is Facebook Timeline, which encourages users to never delete anything from the social network, and to use it as an historic archive. This surely is interesting for forensic examinations, as the user is less likely to delete data.

3 Social Network Data Pool

While social networks vary in features and architecture, we identify the following generic data sources to be of interest in forensic examinations on social networks:

- The social footprint: What is the social graph of the user, with whom

is he or she connected ("friend")?

- Communications pattern: How is the network used for communicating, what method is used, and with whom is the user communicating?
- Pictures and videos: What pictures and videos were uploaded by the user, on which other peoples pictures is he or she tagged?
- Times of activity: When is a specific user connected to the social network, when exactly did a specific activity of interest took place?
- Apps: What apps is the user using, what is their purpose, and what information can be inferred in the social context?

All this information cannot be found on a suspect's hard drive, as it is solely stored at the social network operator. Especially for people that use the social network on a daily base a plethora of information is stored at the social network operator. Facebook claims that more than 50% of its users use it any given day, which would be something around 400 million users [4]. Sometimes information is cached locally, but this is not a reliable source of information as it is neither complete nor stored persistently. Depending on the implementation of the social network, the availability of data itself and the possibility to retrieve the data via API calls can vary among different social networks. However, most of this data can be extracted either 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 [23]. The easiest way for obtaining the data is of course with the consent of the user, who can provide username and password. While the data can be easily analyzed manually afterwards to answer specific questions, the humongous amount of data requires automated tools for an forensic examiner to see the full picture.

4 Visualizations

Based on the social network data pool we define the following graphs of interest and visualizations for social network forensics.

4.1 Basic Visualizations

Social Interconnection Graph: It is trivial to retrieve the list of friends from social networks. In most social networks this is public information, or can be easily collected [12] even without entering the social circle of the account under investigation. However, it is not trivial to cluster these friends, namely to find out who is connected with whom: is a specific contact part of the cluster of work-friends, or are they directly related?

In our approach we use a feature of the Facebook API which allows an application to query if two users are connected. This allows our software to cluster the friends of a user into different groups e.g., people from work, school, family and more, as the members of the groups are much more likely to know each other. The graph can be represented as an undirected graph

$G =< V, E >$ where $V = v_1, v_2, ...v_n$ is the set of friends of a user, and $E = (v_x, v_y), ...$ is the set of edges that connects two nodes in case they are friends on the social network. Highly connected nodes have a high degree, representing well connected friends that know most of the suspect's friends as well. An example graph will be shown in Section 5.

Social Interaction Graph: For many investigations it is of importance to find out who communicated with whom. Various ways of communication are possible among users, like wall posts, direct messages, group communication or following public announcements. Communication can be represented as a directed graph $G =< V, E >$ where the nodes $V = v_1, v_2, ...v_n$ are all the friends while the edges $E = (v_x, v_y), ...$ are directed and the weight of (v_x, v_y) is incremented for every message sent from v_x to v_y. In this paper, however, we do not distinct between the different forms: direct messages, wall posts, and tags are treated equally as these are the most direct form of communication. We leave finding different metrics for future work which would allow the investigator to add custom weights to specific communication forms. An investigator can easily identify the top communication partners on a first sight, and compare them with e.g., obtained phone records. An example graph generated by our tools will be shown in Section 5 as well.

Complete Timeline: With social network users being online 24/7 by using mobile clients on smartphones, the timeline becomes of increasing im-

portance. Not only activity of the user itself can be extracted, but also the activity of all his or her friends. Often the times of activity can be seen easily, if properly visualized. To make analysis feasible it is necessary to use respectively allow different data layers: activity of the user, the friends, group activities, reactions on events from friends, and so forth. It is also crucial that the timeline is zoomable, to visualize time ranges of importance - a single day can have easily more than 500 events for a given profile and all his or her friends.

Location Visualization: Geotagging and location applications are novel features with increasing usage that needs to be reflected in forensic examinations. With foursquare [5] and Facebook Places, just to name a few, the geolocation information stored in social networks is growing steadily. While our toolset is not yet available to extract geodata, we believe that this will become more and more of an issue. Digital cameras as well as smartphones automatically geotag pictures taken with the exact location. Up till now, most social networks remove metadata during the transformation for picture storage [10], but this might change in the future.

4.2 Advanced Visualizations & Information Inference

While the features discussed so far are rather straightforward, we believe that the following list of advanced features and components should become

standard tools in forensics:

Event tracking: For viral scammers and other malicious applications that use the social network for propagation it might be of interest who or what started such a series of events. Tracking such events is not straightforward, but with a collection of social network footprints of various users these events can be easily dissected. This would allow insight into dissemination characteristics and propagation tactics of scammers, as well as advanced analytical capabilities.

Timeline matching: In highly centralized systems such as online social networks, an investigator has the benefit of consistent timestamps as they are provided by the social network. The operators often run their won NTP infrastructure, and keep the clocks consistent across thousands of servers. This can then be used to match timelines of different profiles, and eventually create an exact timeline for a complete cluster of friends or even bigger. While this has been proposed recently for the NTFS file system [17], we believe that this will be of importance for social networks and cloud computing as well.

Differential Snapshots: Once a forensic image of a user profile is collected, at a later point in time the image might look completely different. Therefore, the forensic framework must provide the functionality to not only

visualize the social network data of a user, but also the functionality to visualize differences with previous images of the same user.

5 Results

We implemented the data collection methodology outlined in [18] to collect data from Facebook, currently the biggest social network service. We then parsed the output and generated the graphs as seen below. Data acquisition takes approximately 20 minutes per account, which is in our opinion a reasonable amount of time for data collection. The data currently includes all the social connections, direct communication, pictures and much, much more.

5.1 Results Visualization

For the social interconnection graph we used a feature from the Facebook API that allows an application to query if two specific users are connected. We iteratively tested if the first friend is in connection with the $n-1$ friends of the tested profile, then tested for the second the remaining $n-2$, and so forth. We then extracted the social interconnection graph and plotted it with Gephi [9], an open source graph visualization tool, using the Fruchterman-Reingold algorithm [11]. The nodes that seem to be from the same cluster are colored accordingly without manual intervention, which makes cluster analysis for a forensic examiner very convenient. An example of a social in-

terconnection graph from one of the authors can be seen in Figure 1. Please note that the names of the profiles have been replaced by a random subset of the list of computer scientists on Wikipedia [7]. As the replaced set is random, it obviously does not show real connetions between the computer scientists.

With the data from Facebook it becomes possible to create different social interaction graphs. In our implementation we created different graphs for different forms of interaction, while they could be easily integrated to a complete social interaction graph. An example for a social interaction graph based on tags in pictures on Facebook can be seen in Figure 2. The graph is created using the following steps: (1) starting from an account under investigation, all pictures from all friends are collected, and searched for tagged people. (2) People that are tagged in pictures, and not in the list of friends, are ignored. (3) If the tagged person is in the friend list as well, an edge is added between the two nodes pointing from the profile that uploaded the picture to the profile that is tagged, or the weight increases by one if the edge already exists. The edges are directed and weightened. An investigator can find with this graph persons that have a tight social connection.

Another form of social interaction graph can be created using direct messages: instead of using picture tags, an edge is added between two nodes if the profile under investigation exchanged messages with the other profile.

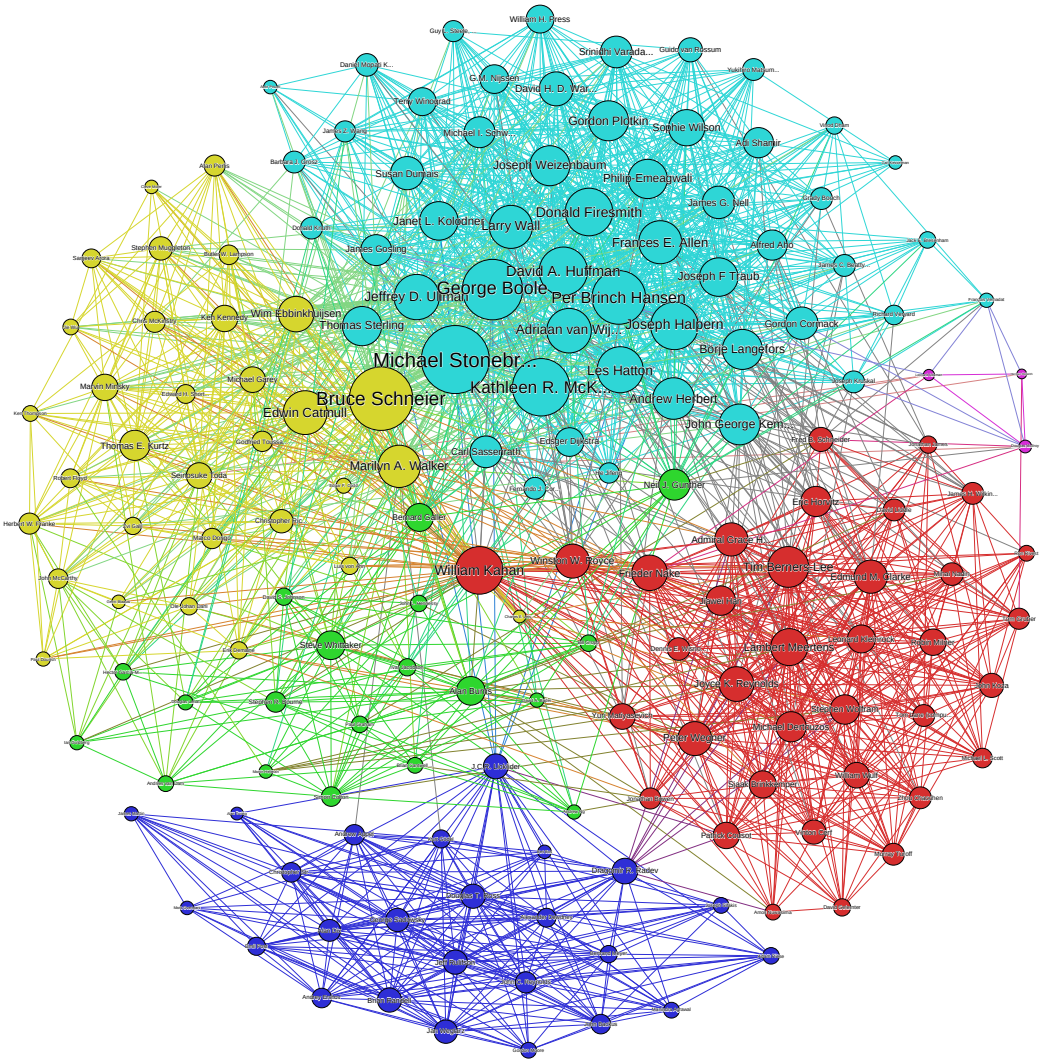

Figure 1: Anonymized Social Interconnection Graph

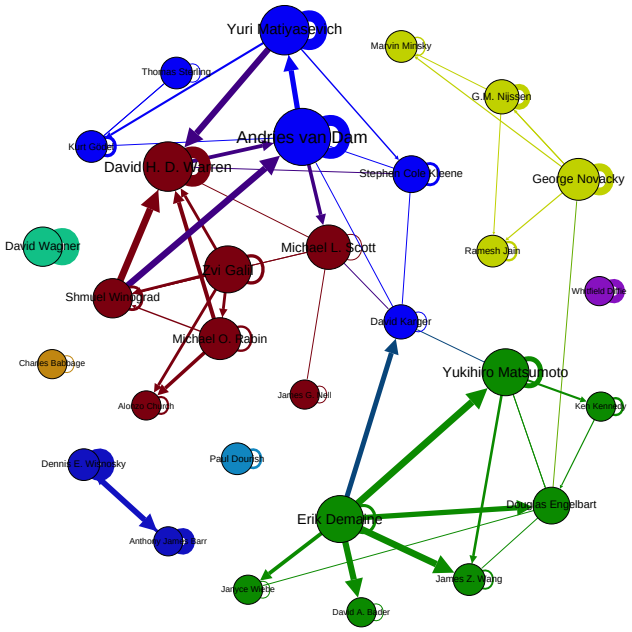

Figure 2: Anonymized Social Interaction Graph using Picture Tags

Intuitively, an edge pointing to a node represents a message sent to that profile. Again the edges are weightened, for the number of messages sent. An example for a social interaction graph using direct message communication can be seen in Figure 3.

Our data collection method also allows automated creation of timelines. While this is currently under development, an example for a timeline of events on Facebook over a 24-hour period can be seen in Figure 4.

5.2 Threats to Validity

Whilst our method is novel and can be easily used in addition to already existing and deployed social network analysis methods (i.e., subpoena requests

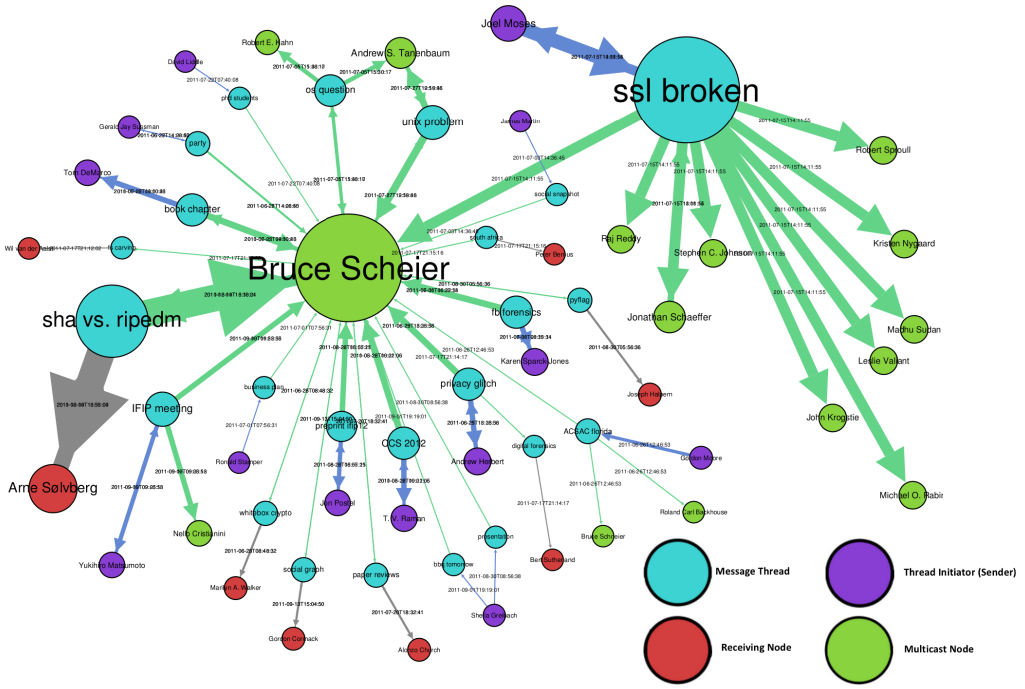

Figure 3: Social interaction graph using direct messages

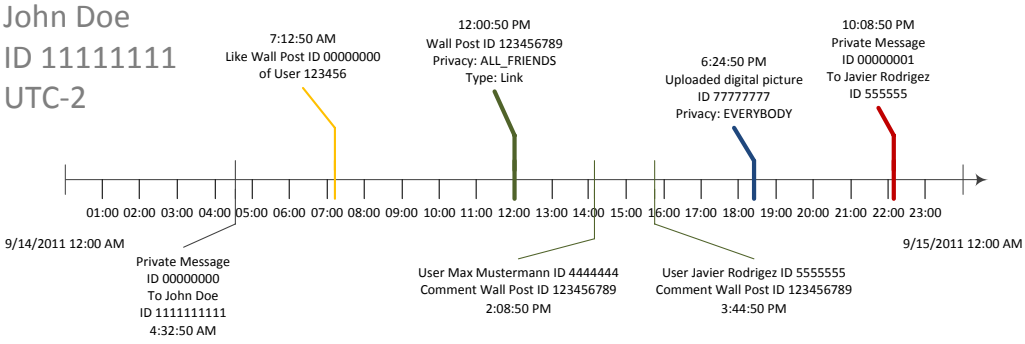

Figure 4: Anonymized Example Timeline for 24-hour period

to the social network operator) they introduce new challenges at the same time as they solve others. One of the most obvious drawbacks is that the data collection is hardly reproduceable: the timelines and graphs generated will look differently for multiple runs as the social network is very dynamic in nature, and the amount of data rather big. Within a single day a user could change his social interconnection graph to large extend, and could try to hide his or her communication in covert traffic. Some other data, like login IP addresses as provided by the Facebook NeoPrint [3], are only available to the operator of the network and not accessible with our method, neither with an automated webbrowser or an API. Furthermore it is not easily possible to guarantee completeness: once data is deleted by the user it can only be undeleted from the social network operator, and is thus not available for our analysis. We believe, however, that most of the data items are rather static in nature, and that our methods are applicable and auxiliary for forensic investigations.

5.3 Future Work

While some of the features are already implemented in our toolkit, we plan on implementing the missing visualizations as well as the automated report generation soon. We would also like to extend the number of supported social networks. It would be very interesting to analyze how static social graphs are, which is so far beyond the scope of this paper. This information could be used to confirm the validity of our methods even further.

6 Conclusion

Social networks and the cloud computing paradigm will undoubtedly change the way forensics examinations are done in the near future. In this paper we identified valuable data sources in social networks, how they can be leveraged for analysis, and discussed to what extend this is possible without the collaboration of the social network operator. Compared to traditional social analysis this can be done fully automated and allows cluster analysis as well as different timeline visualizations. We implemented a proof-of-concept application for creating social interconnection and social interaction graphs (with whom is a user connected, with whom does the user interact often) based on Facebook to show the feasibility of our approach, which we will release as open source software.

References

[1] Apple Promises Fix for Location-Gathering 'Bug' on iPhone. Online at `http://www.wired.com/gadgetlab/2011/04/iphone-location-bug/`.

[2] Facebook Artifact Parser, retrieved September 15th, 2011. Online at `http://trustedsignal.com/code/fbartiparse.py`.

[3] Facebook Law Enforcement Guidelines, retrieved September 15th, 2011. Online at `http://www.eff.org/files/filenode/social_network/`

Facebook2010_SN_LEG-DOJ.PDF.

[4] Facebook Statistics, retrieved September 7th, 2011. Online at http://www.facebook.com/press/info.php?statistics.

[5] foursquare. Online at https://foursquare.com/.

[6] iPhoneTracker, retrieved September 7th, 2011. Online at http://petewarden.github.com/iPhoneTracker/.

[7] List of computer scientists - Wikipedia, the free encyclopedia, retrieved September 15th, 2011. Online at http://en.wikipedia.org/wiki/List_of_computer_scientists.

[8] Xplico - Internet Traffic Decoder. Network Forensic Analysis Tool. Online at http://www.xplico.org.

[9] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, pages 361–362, 2009.

[10] D. Beaver, S. Kumar, H. Li, J. Sobel, and P. Vajgel. Finding a needle in haystack: Facebook's photo storage. In *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, pages 1–8. USENIX Association, 2010.

[11] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

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

[13] D. Brezinski and T. Killalea. Rfc 3227: Guidelines for evidence collection and archiving, 2002. Online at `www.faqs.org/rfcs/rfc3227.html`.

[14] B. Carrier. *File System Forensic Analysis*. Addison-Wesley Professional, 2005.

[15] M. Cohen. Pyflag-an advanced network forensic framework. *digital investigation*, 5:S112–S120, 2008.

[16] G. Conti. *Security Data Visualization: Graphical Techniques for Network Analysis*. No Starch Pr, 2007.

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

[18] M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek, and E. Weippl. Social snapshots: Digital forensics for online social net-

works. *Proceedings of the 27th Annual Computer Security Applications Conference (ACSAC)*, 2011.

[19] 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: Special Issue on Security and Privacy in Social Networks*, 5 2011. Pre Print.

[20] J. Marques. Firefox extensions. Online at `http://codebutler.com/firesheep`.

[21] S. Morrissey. *iOS Forensic Analysis*. Apress, 2010.

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

[23] S. Wasserman. *Social network analysis: Methods and applications*. Cambridge university press, 1994.