Multilevel Visualization Using Enhanced Social Network Analysis with Smartphone Data

Abstract

While technology matures and becomes more productive, mobile devices can be affordable and, consequently, fully integrated in people's lives. After their unexpected bloom and acceptance, Online Social Networks are now sources of valuable information. We therefore use them for tasks varying from direct marketing to forensic analysis. We have already seen Social Network Forensics techniques focused on particular networks implementing methods that collect data from user accounts. During the forensic analysis it is common to aggregate information from different sources but, usually, this procedure causes correlation problems. Here, we present our method to correlate data gathered from various social networks in combination with smartphones creating a new form of social map of the user under investigation. In addition, we introduce a multi level graph that utilises the correlated information from the smartphone and the social networks and demonstrates in three dimensions the relevance of each contact with the suspect.

Keywords: Social Network Forensics, Smartphone Forensics, Social Network Analysis, Evidence Visualization, 3D graph, Data Correlation.

1. Introduction

Online Social Networks (O.S.N.) became an integral part of our social life. Facebook active users outreached one billion on December 2012 ("Key facts", 2013) while Twitter is catching up, with its total number of users reaching 500 million on April 2012 ("Twitter to surpass", 2012). People in developed and developing countries use smartphones to communicate and they prefer to stay online and in touch with their social circles using the embedded applications on their phones. Unfortunately, miscreants could not stay uninvolved. Incidents like cyber-bullying, insults, identity thefts and phishing are currently present both on Facebook and Twitter ("A Facebook crime", 2012; "Reports of Facebook", 2012) and generally social network crime and fraud tend to increase. As a consequence, forensic investigations include the examination of social footprints in their endeavor to find crime evidence.

The variety of O.S.N. complicates the evidence gathering procedure because analysts must use different tools for distinct social networks when they gather the data. If social footprints are discovered on more than one network, they should be correlated to make a complete profile of the suspect. The correlation process can be a tedious task because a large volume of information must be combined together. Moreover, malicious users may choose to employ different

social networks to conceal their social footprints, posing a challenge to forensic analysts to deal with a great amount of data in such investigations.
Facebook ("Downloading your info, 2013") and Twitter ("Twitter blog", 2012) provide the opportunity to download a user's data and, in addition, third-party applications ("Download and view", 2013; "Facebook report", 2013; http://www.fridayed.com) have been developed to perform the same task. Furthermore, there are few commercial tools that can help an analyst to download a forensic image from a social network profile (http://www.oxygen-forensic.com/en/).

In this paper, we propose a social forensic framework capable of combining social footprints from diverse networks, found locally on the suspects' computers (such as mailboxes, Skype chat logs, O.S.N.) and on their smartphones. We then store them under an abstract data representation. Provided with their credentials, a crawling component gathers any data available from the social network and communicates under a predefined API with the tool core, to store gathered evidence on a semantically correct manner on a graph database. Graph properties of each network are also preserved in order to correlate data from different sources effectively. In particular, Facebook friends, Skype contacts and Twitter followers point out that two or more entities of each network are related with each other with a type of relationship and phone contacts describe a more strong bond among them. Finally, the visualization of correlated data results to a broader perspective of the suspect's actions and contacts and helps analysts identify malevolent activity efficiently.

The contributions of this work can be summarized as follows:

- We present a framework capable of combining data from diverse social networks under an abstract graph representation using open source tools.
- We perform data correlation tactics with artifacts found on smartphones.
- We concatenate the social media activity and the personal communication through the phone to distinguish contacts and provide them under a common visualization paradigm.
- To the best of our knowledge there exist a research gap in the area of phone forensics linking them with social network analysis and our work presents a framework that connects the two fields.
- We demonstrate a novel method to display graphs adding layers of connectivity in 3D space allowing the graph to present the most significant contacts to the user.

The paper is organized as follows: In Section 2 we discuss relevant work on Social Network Analysis and Forensics. The architecture of our tool and its implementation as proof of concept is shown in Section 3. Section 4 demonstrates the effectiveness of the proposed method on a case study. In Section 5 we talk about the results and we evaluate the tool. Finally, Section 6 includes the conclusion and our plans for further work.

2. Related Work

Social network analysis (Wasserman & Faust, 1994) is used to find and demonstrate structural properties in a social network using statistical or theoretical methods. Social networks can be represented by graphs that connect entities (nodes) and edges (Bezerianos et al., 2010). In the case of social network forensics, entities are usually persons with various attributes such as name, telephone numbers, email address and edges are the connections between those entities. In a similar fashion, connections can also have attributes describing the type of friendship that bonds the nodes. Among a plethora of popular open source tools, Gephi (https://gephi.org) or Pajek (http://pajek.imfm.si/doku.php) calculate metrics like degree, centrality, in-degree and out-degree in order to numerically describe the network and provide information that can possibly reveal the existence of communities or cliques inside it. Various methods to visualize networks have been proposed in the literature. Pretorius & Van Wijk (2008) for example use complex techniques like node-link diagram linking with parallel histograms or vertices clustering with attributes and hybrid parallel coordinates. These methods are demanding, consist of difficult steps and require machine-learning algorithms and semantic data integration. In (Heer & Shneiderman, 2012) the authors state, through a taxonomy of interactive dynamics for visual analysis, that there should be general characteristics defining visualization tools. They should be able to filter out data so they can focus on relevant items and sort them to expose patterns. Also, they should provide the option to select items to highlight or manipulate them and be able to log analysis histories for reviewing and sharing.

Online Social Networks became sources of remarkable research revealing trends and exploring scientifically and statistically the data their users share. Lampos & Cristianini (2011) present their methodology to investigate Twitter feeds and derive events after statistical learning was applied. Rainfall and flu syndromes were investigated there. Twitter posts are also used to detect and explain the mood of the British nation after the latest economic recession (Lansdall-Welfare et al., 2012). In the area of forensics, social network analysis is critical and it has been highlighted in (Sparrow, 1991) at an early stage. The article states at a preliminary stage the links between social network analysis and criminal intelligence.

Garfinkel (2010) presents a brief summary describing the bloom of digital forensics field and notes the difficulties modern forensic investigators face, considering the plethora of data types and devices they have to deal with. He underlines that we need to develop abstract data formats to enhance digital investigations and examine, among others, the Internet footprint of individuals including their online social activity. Read et al. (2009) present their unified approach to visualize network traffic using a middleware mechanism that connects visualization tools and a layer of abstract data using xml files to achieve this kind of communication through a system called DEViSE (Read et al., 2009).

Our research is closely related to work performed by Huber et al. (2011). Authors implemented a method for collecting data from Facebook accounts, called social snapshots. According to their evaluation, a snapshot of a Facebook account required about 15 minutes average. Crawled data are visualized on a graph representation in order to infer clusters of accounts. Considering the way tools usually present their data, most of them utilize graph theory and 2D graphs. The usability of 3D social graphs though is mentioned in (Mattar & Pfeiffer, 2011). The article shows a user study performing 3D visualizations of social graphs and stresses that these representations can be utilized effectively and are, finally, preferred by users. In their tests the authors assess the ability of participants to efficiently recognize if user A is a friend of user B or distinguish their mutual friends in a social network. They demonstrate that participants acted faster when asked to accomplish these tasks under a 3D visualization environment.

Herman in his survey about information visualization has demonstrated the evolution of graph visualization and navigation (Herman, 2000). The survey reveals a trend to use 3D visualization techniques to present graphs (radial algorithm, cone trees and hyperbolic views). Munzner had already presented the H3 layout algorithm to draw large graphs on 3D hyperbolic space (Munzner, 1997). The implementation successfully laid out structures of over 20,000 nodes. During the same period Parker et al. (1998) advocate that 3D visualization has a number of advantages compared to 2D graphs, concluding that the graphs provide distinctive information when presented in three dimensions (Parker et al., 1998). Scientists also discuss the cognitive benefits of 3D visual environments in recent research projects (Van der Land et al., 2013; Reda et al., 2013).

The growing smartphone market ("Global Smartphones Market", 2013) in combination with the increased use of social network applications indicates that valuable evidence of one's social interactions can be derived by smartphone forensics. Mutawa et al. (2012) conducted forensic analysis on social networking applications using a variety of smartphone devices and analyzing evidence that Facebook, Twitter and MySpace applications leave on smartphones. The forensic analysis indicates that iPhone and Android devices store a significant amount of social interactions of the user. We furthermore know (Andriotis et al., 2012) that it is possible to conduct forensic analysis using open source tools in Android phones without any financial cost. This fact in conjunction with ratings showing the dominance of the Android operating system in smartphone markets (http://www.idc.com/home.jsp), are critical factors to convince researchers experiment with phones or other devices running the specific operating system.

Authors in (Gessiou et al., 2011) extracted social structures from metadata in various documents on the Web (Microsoft Office documents and PDF files). They managed to associate users collaborating on documents with Twitter accounts. Thus, an adversary could use them for malicious purposes (e.g. spam mail, social engineering attacks on organizations). Moreover, according to (Mao et al., 2011), users may post various kind of sensitive information on Twitter (for example tweets about vacation plans) that can be exploited by criminals. However, there are not enough sources in the literature presenting combinational research in the two fields of smartphone forensics and social network analysis. We therefore present our methodology to combine data gathered from online social networks and smartphones of the suspect, in order to perform a critical forensic analysis and we also create a 3D social graph to depict the associations of the suspect with other users in a comprehensive manner. Here we use Facebook, Twitter and Skype as our reference networks. The specific social networks were chosen because of their popularity, but the techniques described below can be applied to other networks, like LinkedIn. Moreover, we can expand the information collection range to include, for example, email accounts.

3. Implementation and algorithm

This section presents the architecture of our social forensic tool and provides a detailed analysis of its components. Our goal is the implementation of a tool, capable of constructing the 'social image' of a suspect through his or her interactions with diverse social networks. The main assumption we make is that the forensic analyst has acquired valid credentials for the respective accounts or the person under investigation cooperated with authorities and provided usernames and passwords. Such credentials can also be found (unencrypted) in some cases in the internal storage of older versions of Android smartphones (Andriotis et al., 2012). Data from Online Social Networks, such as Facebook or Twitter, include social interactions carried out either on suspects' computers (like logs from Skype chats, mailboxes etc.) or on their smartphones. Moreover, the combination of such data can unveil interesting findings like visited locations and friends that were present on particular events.

The social forensic tool has been designed to be modular and its architecture is depicted in Figure 1. It consists of a small core, responsible of coordinating interactions among plugins. The core is also responsible of initiating the crawling mechanism for a given social network by calling a specific plugin. Moreover it provides a data storage interface to plugins using a MySQL database. Communication between plugins and the core is carried out through a predefined "Plugin API". Using the standard API methods (provided by the social media for application developers), the core can manage plugins transparently. Each plugin is responsible to fetch any data available on a specific social network including information about the account under investigation as well as any available interaction with social network entities (friends, groups or events participated, photos uploaded or tagged). The plugins handle the authorization step and any other rate limiting issues might occur especially when dealing with Twitter. The modular design of the tool allows plugins for social networks to be separately developed. Hence, information from other social networks can be integrated very quickly.

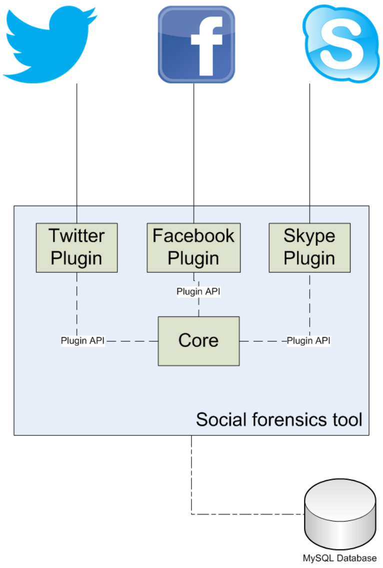
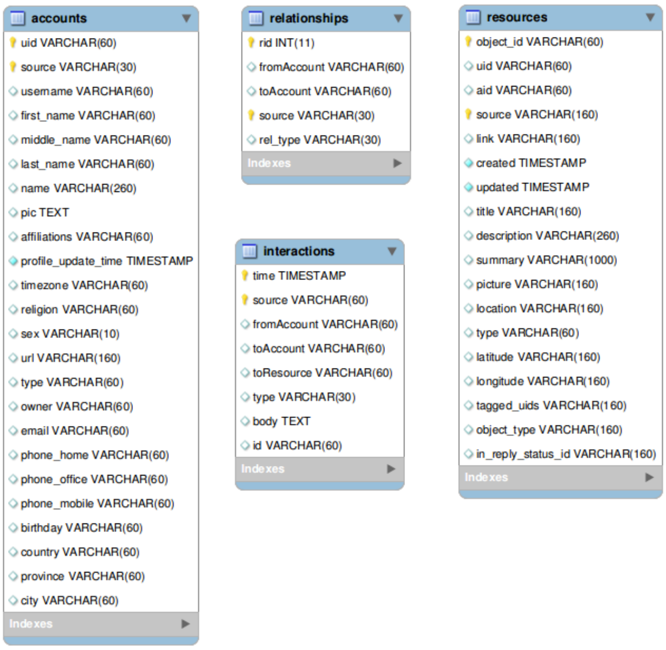


Figure 1: The architecture of the social forensic tool

Each social network uses its own database schema, which differs from others. Moreover, some notions and relationships between entities being present on one social network do not exist in another. For instance, a friend on Facebook is an undirected connection. Recently, Facebook added the "follower" attribute providing a type of directed connection. Similarly, a follower on Twitter is directed because someone can follow a Twitter user without being followed by the same account. These implications impose a challenge, in order to combine data from diverse sources under the same database, preserving the semantic correctness of data. However, despite their differences, social networks exhibit common structures that can be presented in a graph.

Using the aforementioned observation, we followed a more abstract approach concerning the database design and management. Our database schema (figure 2) is constructed as a graph representation. Consequently, each data from different sources must be finally interpreted as a graph component. Each node of the graph represents the notion of account or the entity of a social network. For example, an entity can be a Facebook profile or page, a Twitter or a Skype account. Edges represent relationships among existing accounts on the network (Facebook friends, Twitter followers). Due to diversity of relationships exhibited on social networks, each relationship type can be specified during creation from the plugin author. An example of relation types we used is: *isFriend*, *isGroupMember*. The relationship between two accounts consists of distinct actions such as comments, mentions, photo tags or chats. All these are interactions that two entities can have within their relationship and are logged from each plugin. Finally, social networks provide structured information for particular resources. These resources can be geolocated tweets or posts, photo albums or file transfers. Our tool provides the potential to store and use this valuable information to provide a general view of the suspect's actions.

In more detail, our database schema consists of three tables: accounts, relationships and interactions respectively. A fourth table, called resources has been incorporated to the schema to retain any structured data provided (geolocated tweets, status messages, file transfer logs etc.) We used common fields found on Facebook and Twitter for our database attributes. Three plugins have been developed, for Facebook, Twitter and Skype respectively.



accounts
uid VARCHAR(60)
source VARCHAR(30)
username VARCHAR(60)
first_name VARCHAR(60)
middle_name VARCHAR(60)
last_name VARCHAR(60)
name VARCHAR(260)
pic TEXT
affiliations VARCHAR(60)
profile_update_time TIMESTAMP
timezone VARCHAR(60)
religion VARCHAR(60)
sex VARCHAR(10)
url VARCHAR(160)
type VARCHAR(60)
owner VARCHAR(60)
email VARCHAR(60)
phone_home VARCHAR(60)
phone_office VARCHAR(60)
phone_mobile VARCHAR(60)
birthday VARCHAR(60)
country VARCHAR(60)
province VARCHAR(60)
city VARCHAR(60)
Indexes

relationships
rid INT(11)
fromAccount VARCHAR(60)
toAccount VARCHAR(60)
source VARCHAR(30)
rel_type VARCHAR(30)
Indexes

interactions
time TIMESTAMP
source VARCHAR(60)
fromAccount VARCHAR(60)
toAccount VARCHAR(60)
toResource VARCHAR(60)
type VARCHAR(30)
body TEXT
id VARCHAR(60)
Indexes

resources
object_id VARCHAR(60)
uid VARCHAR(60)
aid VARCHAR(60)
source VARCHAR(160)
link VARCHAR(160)
created TIMESTAMP
updated TIMESTAMP
title VARCHAR(160)
description VARCHAR(260)
summary VARCHAR(1000)
picture VARCHAR(160)
location VARCHAR(160)
type VARCHAR(60)
latitude VARCHAR(160)
longitude VARCHAR(160)
tagged_uids VARCHAR(160)
object_type VARCHAR(160)
in_reply_status_id VARCHAR(160)
Indexes

Figure 2: The database schema of the crawling component.

Facebook

Facebook introduced a REST API, to facilitate the development of rich content applications for Facebook users by third-party developers. In particular, it provides the Graph API ("Graph API", 2013) for accessing entities on Facebook social graph. It also utilizes the Facebook Query Language (FQL) to perform more advanced queries with syntax similar to SQL. To avoid abuse from malicious applications, Facebook uses the OAuth2.0 (http://oauth.net/2/) protocol to control the access of an application to a user's resources. Hence, each application is bundled with a number of permissions (read-only or read-writable) to resources that the application is authorized to access.

We employed fbconsole ("Facebook/fbconsole", 2013), a Python wrapper for Facebook API calls to perform requests and fetch data from Facebook. During the authorization stage, we request permission to handle any read-only resource, to prevent tampering of data during information gathering. Assuming we are provided with the correct credentials of an individual we have access to a variety of resources like friends, profile information, photo albums, direct messages (Facebook chat logs), likes, Wall posts, notifications and other actions of the account. Any available information on the individual profile or on friends' profiles is safely downloaded. We therefore have a broader scope of actions, as well as information for other users that are inaccessible otherwise (for instance some Facebook friends allow access on specific resources such as photo albums, to a limited scope of users). Our crawling mechanism has the ability to fetch data recursively at greater depth.

Twitter

A similar strategy is followed for Twitter. Like Facebook, Twitter incorporates its own REST API ("Documentation, Twitter Developers", 2013) for developing custom applications. It also uses OAuth2.0 for authorizing application access to a user's profile. A wide variety of Python wrappers exist for Twitter API and we chose *tweepy* (http://tweepy.github.io) for our crawling mechanism due to its simplicity. Twitter implements a rate limiting mechanism on requests performed through its REST API. This would limit us to 350 requests per hour and may result to the hypothesis that the gathered information would not be forensically sound. To overcome this obstacle, we requested resources that are private (such as direct messages) at the beginning. Moreover, during tweet collection from friends and followers, accounts with small amount of tweets are crawled first. When rate limiting occurs, the crawler pauses and sleeps for a pre-defined period of time.

Skype

Skype records various information like chat logs, file transfers, and call stats on an SQLite database called main.db, found locally, under pre-defined locations on each system. Many applications such as SkypeLogView ("Skype Logs Reader/Viewer", 2012) and Skype History Viewer ("SkypeHistoryViewer", 2013) have been released in order to present data extracted from the main.db database. Usually Skype syncs logs between two devices running the application. Thus, logs from a remote device can be transferred to the suspect's local computer and vice-versa ("Is chat history", 2012). In the database main.db, the most interesting piece of information is stored under tables named *Conversations*, *Accounts*, *Contacts*, *CallMembers*, *Chats*, *Transfers* and *Messages*. These tables contain detailed information about Skype accounts that have logged in from this computer, chat logs, file transfers between Skype contacts and other call information. The Skype plugin is responsible for finding the location of main.db and interprets its contents to the graph database of our tool, to preserve graph semantics. It checks for evidence locally and does not require any kind of authorization. Data are fetched using Python's standard library module for SQLite databases ("sqlite3", 2013).

Privacy and Ethical Issues

At this point we should stress that the proposed scheme relies on information retrieval using methods provided by the social networks. We handle sensitive personal information and the methodology of our techniques should adhere to the privacy and security policies of the specific networks. Also, the forensic analyst should act under the jurisdiction of the local authorities and law. The policy of each social network differs but there is a common characteristic for all. Any attempt to gain and store sensitive data from the users requires written permission by the specific companies. This means that the person who investigates the accounts should have a written permit to do so. Also, the use of the tool should be governed by a privacy policy that informs which data will be utilized. (“Facebook Platform Policies”, 2013; “Developer Rules of the Road”, 2013; “Skype Privacy Policy”, 2013).

Smartphone data

Smartphones running Android usually do not have applications like Facebook or Twitter preinstalled. This fact tends to change by the introduction of the Facebook Home interface and the new smartphones that launched in the market having the application preinstalled. Apart from that, the already existing applications in the Google Play Store (Facebook, Twitter, Skype) are very popular and it is possible to be found in any phone (“Nearly 75 Percent”, 2013).

During a forensic research the authorities seize all the devices a suspect uses. With the assumption in mind that the individual under investigation uses an Android smartphone we proceed to the forensic analysis of the device. In this section and in this particular paper our reference phone is a Samsung Galaxy Y running version 2.3.5. Authorized commercial tools usually perform the data acquisition but a feasible and forensically accepted method to extract data with open source tools can be found in (Andriotis et al., 2012). Very briefly, the forensic examiner can use the Android SDK, a USB cable, a Linux computer (development machine) and a variety of Linux commands to perform physical acquisition of the internal memory of the phone. We are basically interested in the data partition. Hence, the analyst should connect the phone with the development machine and using the *adb* tool, open a shell, type *mount* to see the file system, utilize the *dd* command to extract the data partition and finally *pull* the image from the SD card to the development machine. Then it will be possible to *mount* the image and perform the analysis safely on the Linux computer, leaving the device intact. A drawback to the method is that the phone should have Super User privileges, but we can bypass this disadvantage using the boot loader of the operating system.

When the analyst mounts the image to a folder of the computer, it is feasible to retain data from the applications. In our case, we are looking for information left in Facebook, Twitter and Skype applications. The corresponding folders and the databases we are looking for in the data partition are listed below:

- data/com.facebook.katana/databases/fb.db
- data/com.twitter.android/databases/*twitter_id*.db
- data/com.skype.raider/files/*username*/main.db

(Note: *twitter_id* is the user id assigned by Twitter and *username* is the username of the Skype account of the suspect.)

In more details, in the first database we are collecting information from the *connections*, *friends_data* and *search_results* tables, from the second database we gather data from *users*, *user_metadata*, *statuses*, *messages, lists* tables and in the last database we focus our interest on the *Contacts*, *CallMembers*, *Conversations*, *Calls*, *Voicemails*, *Messages* and *Chats* tables. The aforementioned databases provide the opportunity to the forensic analyst to perform an investigation on the activity of the user and also juxtapose data found after the execution of our crawler and on the device. This activity will enhance the integrity of data because we correlate information gathered from two different devices but refer to the same account. In addition, if the device does not automatically sync with the applications, we will have the chance to see what might have happened in the suspect's social media life using those databases as a historic reference in contradiction with the current crawled data.

Besides the gathering of the particular information we need to investigate more databases for our project. We are looking at the following places:

- data/com.android.providers.contacts/databases/contacts2.db
- data/com.android.providers.telephony/databases/mmssms.db
- data/com.google.android.gsf/databases/talk.db

The databases above keep information about the contacts stored on the device and the messages that were exchanged via sms, mms and chat, respectively. We therefore take into account specific tables as seen in the following Table 1.

Database	*Table*
contacts2.db	accounts
	data
	raw_contacts
	speed_dial
	status_updates
mmssms.db	pdu
	sms
	threads
talk.db	contacts
	messages

Table 1: Information gathered from smartphone.

The concept behind the collection of these data is that we are able to see who are the people that the suspects contact more via their smartphones. This feature will distinguish the 'electronic' friends from the real contacts and will give a better view of the users' interactions. We call these types of contacts the *'closest'* friends.

Details of the crawled data

We are using the gexf form of data because we intend to visualize our information with open source tools like Gephi (Bastian et al., 2009). While reading the documentation provided in the official web page of Gephi, we concluded that gexf files are more appropriate to store data associated to a node – entity ("Supported Graph Formats", 2012) because they are able to recognize nodes position, color and size attributes. For complete reference of the gexf file format the specifications are listed in a separated source ("GEXF File Format", 2009). Basically, the xml files consist of nodes and edges representations. A node may have attributes (position, geolocation, time series) and the connections between them are specified within the edges tags. The following snippet presents a node and an edge of a graph file our crawler produced (anonymized).

```
<?xml version="1.0" encoding="utf-8"?>
<gexf xmlns:ns0="http://www.gexf.net/1.1draft/viz" version="1.1" xmlns="http://www.gexf.net/1.1draft"
xmlns:viz="http://www.gexf.net/1.1draft/viz" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance">
<graphdefaultedgetype="undirected" mode="static">
<attributes class="edge" mode="static">
<attribute id="21" title="networkx_key" type="integer" />
<attribute id="22" title="int_type" type="string" />
<attribute id="23" title="name" type="string" />
</attributes>
<attributes class="node" mode="static">
<attribute id="0" title="province" type="string" />
<attribute id="1" title="username" type="string" />
<attribute id="2" title="last_name" type="string" />
…
<attribute id="20" title="email" type="string" />
</attributes>
<nodes>
<node id="Facebook-xxxxxxxxxx" label="">
<ns0:color b="152" g="89" r="59" />
<attvalues>
<attvalue for="0" value="" />
…
<attvalue for="20" value="" />
</attvalues>
</node>
</nodes>
<edges>
<edge id="0" source="Facebook-xxxxxxxxx" target="Facebook-xxxxxxxxxx" weight="aaaaaaaa">
<attvalues>
<attvalue for="21" value="0" />
<attvalue for="22" value="message" />
</attvalues>
</edge>
</edges>
</graph>
</gexf>
```

Correlation of data and visualization

Gathered data are just various pieces of a puzzle that may come together. The identification of common entities on diverse social networks is crucial for the forensic procedure. For example, assume that a suspect has two accounts on respective social networks, Facebook and Skype. The suspect is chatting on Skype with his/her abettor about details of an upcoming event. After the incident, the suspect is posting a message on the abettor's profile on Facebook. Even if a forensic analyst has acquired both pieces of information, they won't be sufficient if not combined. A check on common attributes of entities from diverse networks can enhance the correlation of two entities.

A common attribute is the name an individual uses on a social network. According to Facebook's Name Policy ("Facebook's Name Policy", 2013) users are obliged to provide their real name during profile registration. Violation of this policy is susceptible of future account lockouts. ("Facebook account lockout", 2013). Despite the fake profiles on Facebook ("Facebook has 83 million fake profiles", 2013) we observed that users tend to retain their real name on Skype. Unfortunately, the name is not an attribute that distinguishes uniquely a person. Many individuals have the same first and last name (for example John Smith). Moreover, crooks would not reveal their real identity or they would probably perform malicious activities using fake profiles. Hence, we can also use the date of birth as another attribute. As users tend to provide real information on social networks during registration, it is probable that the date of birth is preserved on different social networks. We use the aforementioned attributes as a pilot method to correlate accounts on different social networks. Correlation takes place after crawling has finished. For each Facebook account, our tool searches for common names and date of births on Skype and Twitter. If a match is found, a correlation edge is assigned between accounts, containing a weight.

As already mentioned, the output of our tool offers visualization capability of the crawled data as a graph. Graph representation helps analysts distinguish core entities (usually accounts) that the suspect has interacted with. Our database design allows exporting data to a graph format data (gexf), but they can be also exported to other formats. This feature provides the opportunity to analysts to use the visualization tools they prefer. Our tool is responsible to generate the xml file and the visualization tools will handle the presentation of data. For each relationship between two entities we count the number of the respective interactions and use it as a weight of the relationship. The number of interactions is used to calculate edge thickness. Figure 3(a) is a sample output of our relationship representation on a graph. Two accounts belonging to the same individual are used on Facebook and Skype consisted of 563 and 95 friends, respectively. Nodes from different social networks have different colors. Thus, nodes belonging to Facebook are colored purple and nodes belonging to Skype are colored light blue. Edge thickness indicates that some nodes have more interactions than others. Green connections indicate that the correlation mechanism successfully identified nodes from Facebook and Skype that belong to the same individual. Changing small parts in the source code we can alter the way the visualization looks. Figure 3(b) shows the visualization using standard colours: red, blue, green.

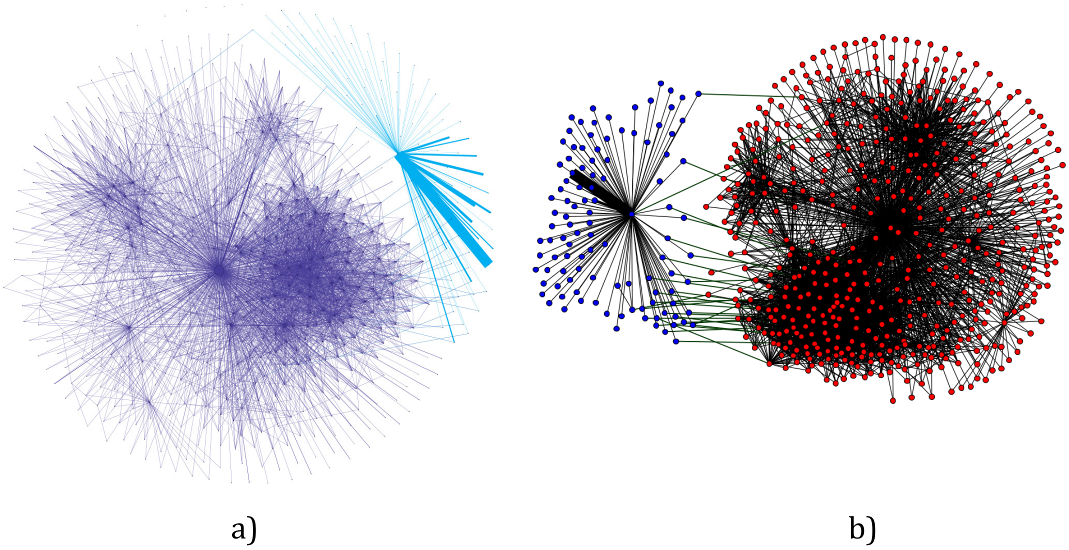
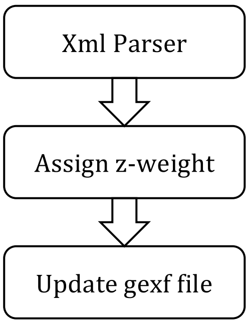


Figure 3: a) A sample graph output of social forensics tool from Facebook (dark blue), Skype (light blue). Twitter data do not exist here. On the Right (b) red color is for Facebook and blue for Skype.

Figure 2 shows how our tool visualizes the crawled data in 2D space and how the correlation of the accounts would be achieved. This type of visualization is indeed quite trivial and produces messy graphs. Our tool not only aggregates information taken by several social media but it also creates a new form of 3D visualization, by taking into account the suspect's *closest* friends and correlating information found on the smartphone device. The concept behind this proposal is simple. If the suspect interacts with specific people through the smartphone, it can be an indication of a closest relationship. We particularly investigate cases where the individual sends or receives short messages, talks or chats with contacts stored in the smartphone list. Hence, we use the email address, the telephone number and finally the name of people located on the smartphone databases. It is possible the person under investigation (subject) to sync data using the applications provided for their smartphones. The applications we are using in this paper (Facebook, Twitter, Skype) provide this choice to the smartphone user. Thus, our aim to find people that interact with our suspect would be easier because all the correlation effort will be automatically done by the applications. However, if the sync choice is not applicable, the use of Table 1 will provide the type of information we want to achieve our goal.

When the classification of the contacts is completed (closest or not friends) the entity–account is assigned a *z-weight* to distinguish persons found on the smartphone of our subject. Z-weight is a value representing the z-axis in a 3D xyz space. The suspect is positioned at the front of our system (z = 0 by default), the closest friends' coordinates have a specific z-weight = Z (it is adjustable) and the other accounts are assigned a z-weight = 2Z. Figure 4 describes the flow and the pseudo code for this procedure.



```
for node in nodes:
    if node == closest:
        z_weight = Z
    elif node == not_closest:
        z_weight = 2*Z
    else:
        z_weight = 0
    f.write('<viz:position  z="z_weight"/>')
```

Figure 4: Enhancing the crawler output in 3D space.

Now our data can be seen in the 3D space and the closest friends (smartphone contacts) are indeed closer to our subject. The other entities of the graph appear to be more distant, lying on another layer. Thus, we now have two layers of proximity: the closest friends (found on the smartphone) and the social media friends. In order to achieve the particular feature with Gephi, we have to use the OpenOrd layout (Martin et al., 2011) or a similar plugin. (We also tested our files with the new algorithm Force Atlas 3D but the outcome was not encouraging. Thus, the recommended algorithm is OpenOrd).

The following figure (figure 5) conceptualizes the proposed visualization scheme. The closest friends are presented in a shorter distance than those entities found on social media.

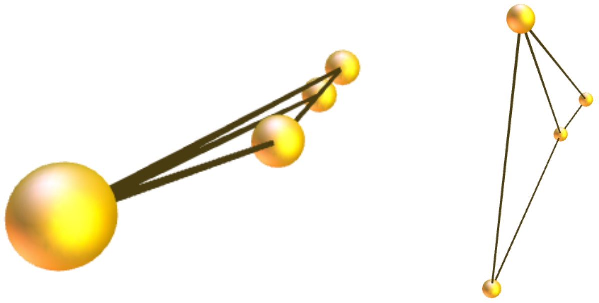


Figure 5: Visualization of the proposed scheme.

4. Case study

We created a scenario of a typical interconnection of an individual acquiring three accounts for Facebook, Twitter and Skype (FSuspect, TSuspect, SSuspect, respectively). Facebook connections will be represented as F01, F02, … , Twitter contacts (followers, following) as T01, T02, … and similarly for Skype we will write S01, S02, … . In this particular study we will show only the Twitter interactions of people that are followers of the suspect and the suspect also follows them (i.e. T01), because here we investigate unidirectional graphs and only this type of Twitter interaction can be assumed as unidirectional. The person under investigation uses the same credentials to log into Facebook and

Twitter and a different password set to log into Skype. There are 8 Facebook friends, 1 Twitter follower, 3 Skype contacts and the user follows 5 Twitter accounts (one of them is the follower T01). A mapping of the social life is provided in Table 2.

Facebook	*Twitter*	*Skype*
F01	T01	S01
F02	T02	S02
F03	T03	
F04	T04	
F05	T05	
F06		
F07		
F08		
Same credentials		Other

Table 2: Social media connection mapping for case study subject.

The suspect of our case study uses primarily Facebook and communicates most with F02, F03, F04, F07 and F08. Our tool crawled the accounts and produced the following result (Figure 6).

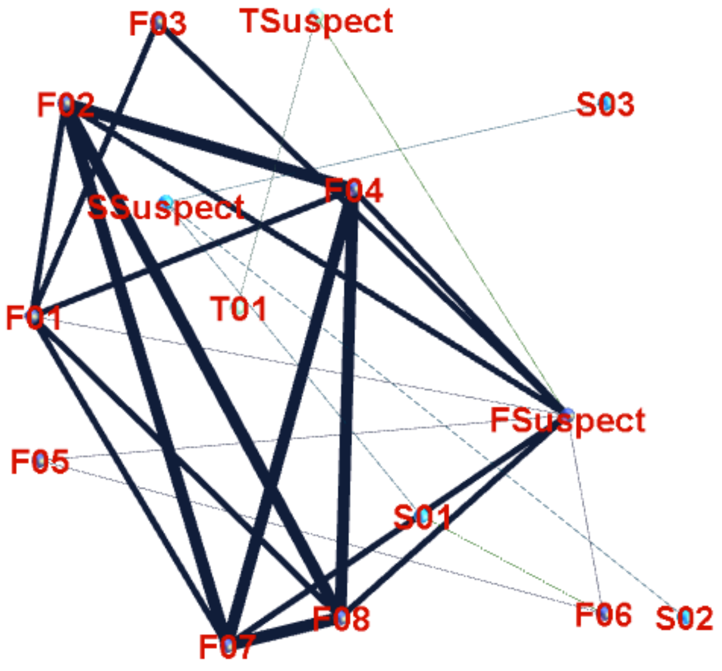


Figure 6: The social activity of the suspect in our case study.

In Figure 6 we can see that our tool produced an interesting graph, linking the FSuspect and TSuspect account (as expected because they use the same credentials) but we can also see that S01 and F06 are also connected (because they have the same name). The latter connection provides an indication that maybe these two accounts belong to the same person. Also we can see that F02, F04, F07 and F08 are linked with stronger bonds and they are interacting more regularly than the other users. The specific observation leads to the conclusion that these four people constitute a community or a clique in the social map of the suspect. However, the crawler was not able to link directly the SSuspect account with the other accounts of the suspect.

The following Figures (7 - 8) present our novel perspective for evidence visualization combining the data correlation from the suspect's smartphone with the crawled data from the social media. Figure 7 demonstrates a part of the graph produced after applying the OpenOrd layout (using Gephi) to our data. FSuspect now appears at the foreground and its closest friends (F04, F05, F07) are physically closer to the node FSuspect. TSuspect and SSuspect are also at the foreground and the three suspect accounts are now connected because they were found on the smartphone.

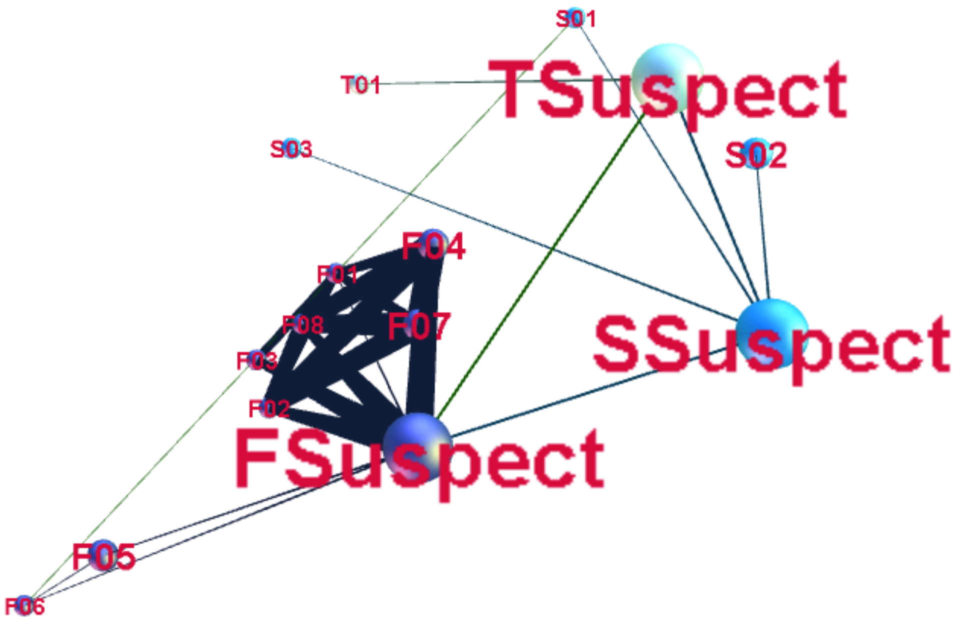
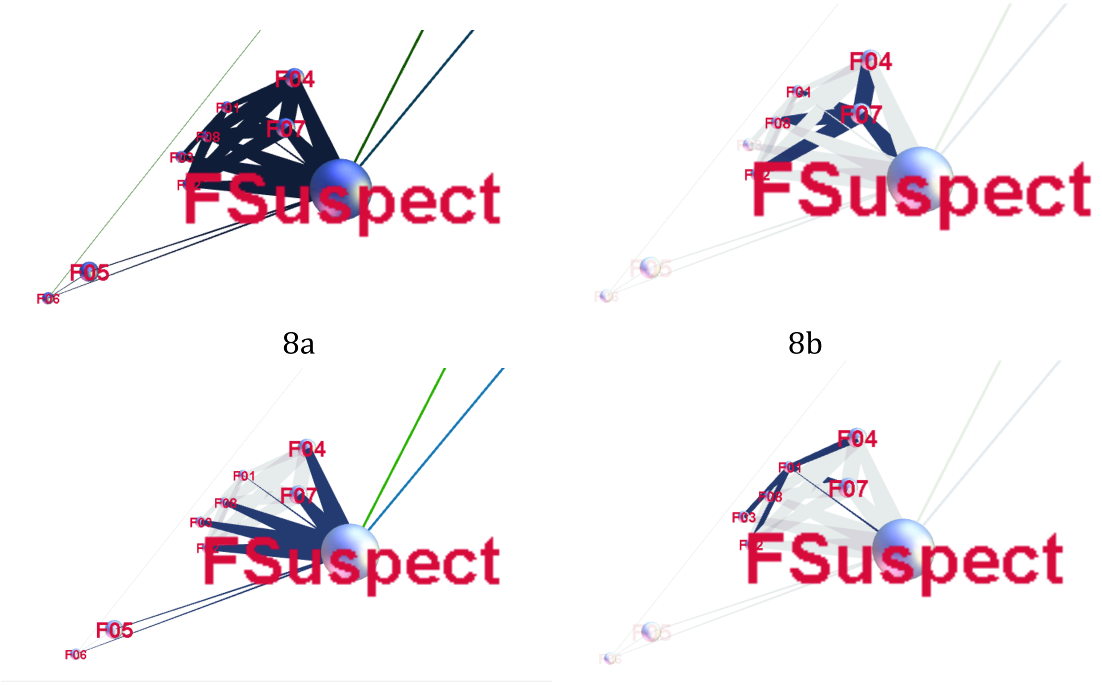


Figure 7: The proposed visualization scheme after the correlation of data from the suspect's smartphone (case study).

The next figures focus on the Facebook activity of our case study subject. Figure 8a might look overloaded with thick edges but shows a cluster of the suspect's social media activity (Facebook). This type of visualization will show the Facebook friends (8b) and the most active connections indicated by the edge thickness. As already mentioned the closest connections are located on the second layer of the z-axis (F04, F05, F07) and they are now more distinctive than the other Facebook accounts (8c). Finally (8d), We can easily conclude that the suspect is linked with accounts such as F04, which appears to be distant but it is heavily connected with two of the closest friends.



8a

8b

8c

8d

Figure 8: Focusing on Facebook activity (Case Study).

5. Case Study Results and Evaluation

The outlined case study brings to the surface the advantages of the proposed scheme. First, the crawled data if visualised without been processed would look like Figure 6. This mapping of evidence is enlightening because it performs an initial scanning of the data and produces correlated information that will display an overall view of the social media life of the person under investigation. After the second phase of the process (parsing and enhancing the data with the smartphone evidence) we reach another level of information aggregation. Figure 7 illustrates that it is possible to achieve a very informative framework to help the forensic analyst to work more efficiently. This task is feasible because the new format of visualization will distinguish the smartphone evidence from the crawled data but at the same time we will have the opportunity to see all the involved parts within one graph.

The depth in the visualization map now shows the level of involvement with the particular suspect. The closest layer to the suspects illustrates contacts found on their smartphones and interacted with them either by sms and chatting or by calling them. These actions reveal a closer connection with the person under investigation and separates entities from the social media friends that are now depicted in a more distant layer. Hence, instead of using three separate graphs to depict the social media activity of the suspect and another one for the smartphone data, we are now able to have an overview using a single enhanced

3D graph. We can also use the tools that programs like Gephi provide without having any particular problem with the form of our data. Clustering is also possible as we already saw in our case study. Furthermore, communities and cliques can be easily found because of the additional data we gather and their interactions are marked by the thickness of the edges that connect them.

In our case study we were able to connect the three suspect accounts with three closest entities (F04, F05, F07) and realized that there are cases that two of the closest friends of the suspect (F04 and F07) are connected with strong bonds, which is an indication that these persons might be real friends or colleagues. Observations like the former would be more difficult to occur by a simple graph. Our method adds two layers of aggregation. First, we stress interactions between the entities and then we bring a type of relativeness among them.

We conducted a series of tests in order to measure runtime performance of the crawling infrastructure. We used accounts F01, S01 and T01 for Facebook, Skype and Twitter respectively to measure crawling time of respective plugins from our crawling mechanism. We crawled each social network separately, 10 times and here we report the average values. The Facebook plugin average crawling time was 65.26 seconds, while Twitter and Skype plugins needed 26.13 and 2.03 seconds respectively. The low crawling time of Skype is a result of the evidence locality. As already mentioned Skype plugin simply translates evidence found locally on main.db. Moreover, as shown on Table 2, Facebook plugin has increased workload, as it will crawl 8 accounts. Moreover, the diversity of data (photos/albums/likes etc.) increases the burden. When we ran our mechanism for all social networks (we also did that 10 times) the average time reported was 99.21 seconds.

One can argue though that if we upscale the case study to meet more realistic scenarios, the graphs will look busier and it will affect their readability. This stands true but we believe that it is a common characteristic of all graphs. They tend to look more complicated while their nodes grow. We added some additional interactions to our case study to stress that factor. The following figure (Figure 9) depicts the interactions between our suspect with 12 Facebook accounts, 12 Twitter accounts and 5 Skype contacts. We can conclude that the best approach when we have to deal with a lot of data is to focus (using utilities the visualization tools provide) on specific clusters of the graph. The functionality and statistics those tools provide will help the forensic analysts to improve the outcome of their investigation.

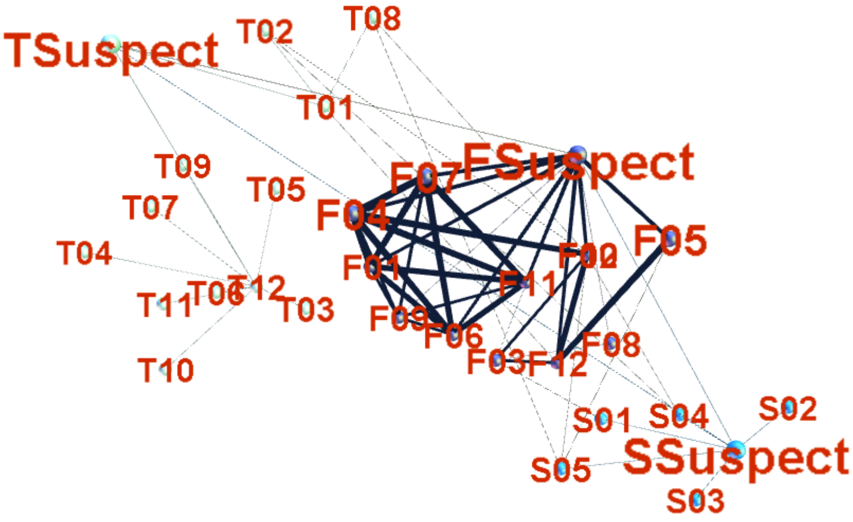


Figure 9: A larger scale social graph.

6. Conclusions and future work

In this paper we introduced an interdisciplinary model that provides the opportunity to forensic analysts to perform investigations connecting the social media activity of suspects with evidence emerged from their real life and gathered from their Android smartphones. We tried to merge Social Network Analysis and Forensics and proposed a tool that crawls specific social networks and produces a grid of data able to be used by visualization programs like Gephi. We correlate artefacts from different accounts used by the same person and aggregate smartphone evidence arose by sent or received messages, chats or telephone calls. Furthermore, we follow an abstract binding of the information to propose a novel visualization scheme that separates suspect's contacts to closest and distant friends and categorize them on separate layers of the 3D space. Thus, it is possible to find more effectively interactions among entities that were unseen otherwise.

This framework collects data from social networks and smartphones but it can integrate more modules to form a powerful tool. The social networks we presented here were chosen because of their popularity among the users but there exist other networks like LinkedIn that could reveal targeted information to the analyst such as the social circles of colleagues and co-workers of the person under investigation. The concept presented in Figure 1 is upgradable and it is built to store and support all kinds of Social Networks. Our intentions for future work include the addition of capability to depict directed interactions among the entities, like those found on the Twitter Network. Also, the thickness of interactions between two entities might produce complex and confusing graphs. Therefore, we should consider developing a formula that produces more elegant graphs that stress the importance of the interaction frequency using thickness but also takes into consideration the style of the graph. In addition, the positioning of entities can be more efficient if we develop and run a special layout for tools like Gephi.

Another interesting extension for the project would be the integration of our data with concepts like the DEViSE middleware to produce a more robust scheme that can be used by more visualization tools. This kind of abstraction on the data can be useful because it provides the opportunity to enrich the information gathered for every entity and therefore perform more efficient analysis considering for example the location of the users.

The implementation for Android devices could be a limiting factor but we believe that the methodology we proposed can be migrated to Apple's devices too, because it consists of distinct modules bound together at the end of the process. Hence, only the smartphone data part should adopt any changes for the iPhone integration. We will investigate this possibility in the future in order to expand the scope of our tool. In addition, the use of tools like Gephi allows time series graphs. The presentation of data in such a fashion would be a powerful addition to the tool. The next upgrade of the presented framework will introduce a new feature that allows the end user to add entities manually to the social contacts of the suspect. During investigation the tool should be able to handle information obtained from various sources. To establish good forensic practices, all interactions of the end user with the data should be recorder and logged at an efficient manner.

Acknowledgement

This work has been supported by the European Union's Prevention of and Fight against Crime Programme "Illegal Use of Internet" - ISEC 2010 Action Grants, grant ref. HOME/2010/ISEC/AG/INT-002 ("ForToo", 2013). This work has also been supported in part by the Faculty of Engineering's Systems Centre (University of Bristol).

References

20 Infamous Crimes Committed and Solved on Facebook [INFOGRAPHIC] (2013). Retrieved April 26, 2013, from http://mashable.com/2012/03/01/facebook-crimes/

5,000 people investigated by police for something they said on Facebook or Twitter as 'social network crime' soars 800% | Mail Online (2013). Retrieved April 26, 2013, from http://www.dailymail.co.uk/news/article-2253692/Facebook-Twitter-crime-sees-fold-increase-police-deal-5-000-cases-involving-websites.html

A Facebook crime every 40 minutes: 12,300 cases are linked to the site | Mail Online (2012). Retrieved April 26, 2013, from http://www.dailymail.co.uk/news/article-2154624/A-Facebook-crime-40-minutes-12-300-cases-linked-site.html

Andriotis, P., Oikonomou, G., & Tryfonas, T. (2012). Forensic analysis of wireless networking evidence of Android smartphones. In *2012 IEEE International Workshop on Information Forensics and Security (WIFS),* (pp. 109 - 114).

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Proceedings 3rd International AAAI Conference on Weblogs and Social Media*, (pp. 361 - 362), San Jose, CA, USA.

Bezerianos, A., Chevalier, F., Dragicevic, P., Elmqvist, N., & Fekete, J. D. (2010). GraphDice: A System for Exploring Multivariate Social Networks, *Computer Graphics Forum*, 29 (3), 863 - 872.

Developer Rules of the Road (2013). Retrieved August 7, 2013, from http://dev.twitter.com.terms/api-terms

Documentation, Twitter Developers (2013). Retrieved April 26, 2013, from https://dev.twitter.com/docs

Downloading Your Info | Facebook Help Center | Facebook (2013). Retrieved April 26, 2013, from http://www.facebook.com/help/131112897028467/

Facebook account lockout means better scam prevention | Digital Trends (2013). Retrieved May 9, 2013, from http://www.digitaltrends.com/social-media/locking-you-out-of-your-facebook-account-is-the-new-way-to-prevent-scams/

Facebook has 83 million fake profiles. Retrieved May 9, 2013, from http://www.globalpost.com/dispatch/news/business/120802/facebook-has-83-million-fake-profiles

Facebook Platform Policies (2013). Retrieved August 7, 2013, from http://developers.facebook.com/policy

Facebook report – Wolfram | Alpha (2013). Retrieved April 26, 2013, from http://www.wolframalpha.com/input/?i=facebook%20report

Facebook's Name Policy | Facebook Help Center | Facebook. Retrieved May 9, 2013, from https://www.facebook.com/help/292517374180078

Facebook/fbconsole – GitHub (2013). Retrieved April 26, 2013, from https://github.com/facebook/fbconsole

ForToo (2013). Retrieved August 7, 2013, from http://www.fortoo.eu

Garfinkel, S. (2010). Digital forensics research: The next 10 years. *Digital Investigation*, 7, S64 – S73.

Gessiou, E., Athanasopoulos, E., & Ioannidis, S. (2011). *Digging up social* *structures from documents on the web.* (Tech. Rep. No. 412), ICS - FORTH, Hellas.

GEXF File Format (2009). Retrieved August 7, 2013, from http://gexf.net/format/schema.html

Global Smartphones Market 2012 – 2016 (2013). Retrieved August 7, 2013, from http://www.technavio.com/report/global-smartphones-market-2012-2016

Graph API – Facebook Developers (2013). Retrieved April 26, 2013, from https://developers.facebook.com/docs/reference/api/

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis, *Communications of the ACM*, 55 (4), 45 - 54.

Herman, I. (2000). Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 6 (1), 24 – 43.

Huber, M., Mulazzani, M., Leithner, M., Schrittwieser, S., Wondracek, G., & E. Weippl, E. (2011). Social snapshots: digital forensics for online social networks. In ACM, *Proceedings of the 27th Annual Computer Security Applications Conference*, (pp. 113–122). ACSAC '11, New York, NY, USA.

Is chat history stored on Skype servers? - Skype Community (2012). Retrieved April 26, 2013, from http://community.skype.com/t5/Security-Privacy-Trust-and/Is-chat-history-stored-on-Skype-servers/td-p/472379

Key Facts – Facebook Newsroom (2013). Retrieved April 26, 2013, from http://newsroom.fb.com/Key-Facts

Lampos, V., & Cristianini, N. (2011). Nowcasting Events from the Social Web with Statistical Learning, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3 (4), 72:1 - 72:22.

Lansdall-Welfare, T., Lampos, V., & Cristianini, N. (2012). Nowcasting the mood of the nation, *Significance*, 9 (4), 26 - 28.

Mao, H., Shuai, X., & Kapadia, A. (2011). Loose tweets: an analysis of privacy leaks on twitter. In ACM, *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, (pp. 1 – 12). WPES '11, New York, NY, USA.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. In *Proceedings of the SPIE 7868, Visualization and Data Analysis 2011*, 786809, CA, USA.

Mattar, N., & Pfeiffer, T. (2011). Interactive 3D Graphs for Web-based Social Networking Platforms. *International Journal of Computer Information Systems and Industrial Management Applications*, 3, 427 – 434.

Mutawa A. N., Baggili, I., & Marrington, A. (2012). Forensic analysis of social networking applications on mobile devices. *Digital Investigation*, 9, S24 – S33.

Munzner, T. (1997). H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. In *Proceedings of IEEE Symposium of Information Visualization*, (pp 2 - 10), Phoenix, AZ, USA.

Nearly 75 Percent of U.S. Android Smartphone Users Accessed Facebook's Website or App From Their Phones in March, According to NPD's Connected Intelligence (2012). Retrieved August 7, 2013, from http://www.prweb.com/releases/2012/5/prweb9519156.htm

Parker, G., Franck, G., & Ware, C. (1998). Visualization of Large Nested Graphs in 3D: Navigation and Interaction. *Journal of Visual Languages and Computing*, 9, 299 – 317.

Pretorius, A. J., & Van Wijk J. J. (2008). Visual inspection of multivariate graphs. *Computer Graphics Forum*, 27 (3), 967 – 974.

Read, H., Blyth, A., & Sutherland, I. (2009). A Unified Approach to Network Traffic and Network Security Visualization, ICC'09, In *Proceedings of the 2009 IEEE international conference on Communications*, (pp. 614 – 619). IEEE Press.

Read, H., Xynos, K., & Bluth, A. (2009). Presenting DEViSE: Data Exchange for Visualizing Security Events. *Computer Graphics and Applications, IEEE*, 29 (3), 6 – 11.

Reda, K., Febretti, A., Knoll, A., Aurisano, J., Leigh, J., Johnson, A., Papka, M. E., & Hereld, M. (2013). Visualizing Large, Heterogeneous Data in Hybrid-Reality Environments. *Computer Graphics and Applications, IEEE* , 33 (4), 38 - 48.

Reports of Facebook and Twitter crimes up by 400% in Merseyside, police figures show - Liverpool Echo (2012). Retrieved April 26, 2013, from http://www.liverpoolecho.co.uk/liverpool-news/local-news/2012/12/28/reports-of-facebook-and-twitter-crimes-up-by-400-in-merseyside-police-figures-show-100252-32504694/

Skype Logs Reader/Viewer (.dbb and main.db files) (2012). Retrieved May 8, 2013, from http://www.nirsoft.net/utils/skype_log_view.html

Skype Privacy Policy (2013). Retrieved August 7, 2013, from http://www.skype.com/en/legal/privacy

SkypeHistoryViewer | Free software downloads at SourceForge.net (2013). Retrieved May 8, 2013, from http://sourceforge.net/projects/skypehistory/

Sparrow K. M. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects, *Social Networks*, 13 (3), 251 - 274.

sqlite3 - DB-API 2.0 interface for SQLite databases - Python v2.7.4 documentation (2013). Retrieved May 8, 2013, from http://docs.python.org/2/library/sqlite3.html

Supported Graph Formats (2012). Retrieved August 7, 2013, from http://gephi.org/users/supported-graph-formats/

Twitter Blog: Your Twitter archive (2012). Retrieved April 26, 2013, from http://blog.twitter.com/2012/12/your-twitter-archive.html

Twitter To Surpass 500 Million Registered Users On Wednesday – AllTwitter (2012). Retrieved April 26, 2013, from http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

Van der Land, S., Schouten, A. P., Feldberg, F., Van den Hooff, B., & Huysman, M. (2013). Lost in space? Cognitive fit and cognitive load in 3D virtual environments. *Computers in Human Behavior*, 29 (3), 1054-1064.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis.* Cambridge, United Kingdom: Cambridge University Press.