Online Social Network Information Forensics

A survey on use of various tools and determining how cautious facebook users are?

Amber Umair  Priyadarsi Nanda  Xiangjian He
School of Computing and Communications
Faculty of Engineering and IT
University of Technology, Sydney
Australia
amber.umair@student.uts.edu.au

Abstract—Online Social Networks (OSN) such as Facebook, Twitter, LinkedIn, and Instagram are heavily used to socialize, entertain or gain insights on people behavior and their activities. Everyday terabytes of data is generated over these networks, which is then used by the businesses to generate revenue or misused by the wrongdoers to exploit vulnerabilities of these social network platforms. Specifically social network information helps in extracting various important features such as: user association, access pattern, location information etc. Recent research shows, many such features could be used to develop novel attack models and investigate further into defending the users from exposing their information to outsiders. This paper analyzes some of the available tools to extract OSN information and discusses research work on similar type of unstructured data. Recent research works, which focus on gathering bits and pieces of information to extract meaningful results for digital forensics, has been discussed. An online survey is conducted to gauge the cautiousness of users in social media usage in terms of personal information dissemination.

Keywords—Online Social networks; Facebook; Digital Forensics; social snapshot

I. INTRODUCTION

OSNs have become utilities like electricity, water and gas. People extensively use / misuse them. Terabytes of unstructured data is generated daily by the users on social networks. This propagation of private information often leads to disastrous results. But, law-enforcing agencies as an aid to reach, cease and prosecute criminals can also use this information. Social network information analysis / forensics is proving its role in digital forensics but still a lot of milestones need to be achieved. This paper discusses and compares the tools available for acquisition and analysis of social network data (Facebook) and combining the pieces of information to get an understanding of the big picture.

Research on the tools to obtain and analyse OSN information is still at infancy stage. The challenges include ever-changing nature of social networks format (APIs), huge data sizes and manual intervention requirement to obtain relevant data. Our contributions in this paper includes:

1. Analysis of various tools to collect, analyse and visualize data obtained from Facebook. Tools are selected on the basis of retrieved data quality, data access feasibility and analysis.

2. Study and discuss the research done in the area of OSN information forensics and its correlation with other types of data sources.

3. Conduct an online survey to understand and analyse the Facebook usage by people and their attentiveness towards the security challenges imposed on their privacy.

This paper is divided in six sections. Section II discusses the related research work and provides the background of using the OSN information for forensics purpose. Section III overviews the functionality of Facebook data tools. Section IV analyses the tools to obtain and compare the OSN information. Section V details on the online survey conducted and discusses its findings. We conclude the paper in Section VI with future work.

II. RELATED WORK

In recent years, use of technology for socializing has increased a lot. Relatives living across the globe, friends from the school, past and present office colleague are just a click away in sharing information. Moreover, these clicks are generated from disparate sources / devices. A typical computer is no more the origin of this type of communication. Now the origin varies from a laptop, IPad, Tab, Mobile phone (IOS, Android etc.), google watch and health monitoring device (e.g. FITBIT). Although use of such varying devices to socialize has increased but usually users are unaware of the device configurations to avoid dissemination of unnecessary information about them. They don’t realize the serious implications of such ignorant information propagation.

Socializing over web via OSNs, blogging, and emails can reveal a lot of information about an individual. Such information could be used by criminals to harm someone or to plan any harmful activity. Three information aspects about an individual could be identified by social information are:

- Relationship Strength (Weak or Strong tie)
- Type of relationship
- An individual’s predicted behavior / characteristics

Nowadays, Facebook is an important part of majority of individuals. People post, like and share a great deal about
themselves on Facebook. By observing the type of information shared by an individual, analyzing his/her joined groups, lot of information could be inferred about them. For example, the strength of a relationship between two individuals, user A and user B and their corresponding contacts can be estimated. The number of tags/picture sharing, similar subscribed groups and number of likes to one another’s post can help in guessing the relationship strength [1, 2]. This relationship strength can be demonstrated by the following equation:

\[
\text{Relationship strength} = \frac{\text{Number of likes from user A to B}}{\text{Total number of likes by A}}
\]

User profiling on the basis of tags and joined groups support the theory of homophiles in social networks [3], which suggests that users with similar interest are more likely to form social connections [4]. However, in some cases users socially influence each other’s behavior and they become similar over time [5].

From forensics perspective, text mining on social networks data could provide useful results, because topical similarity (words used by users in tags and statuses) can help in social link/relationship prediction [5]. Mining or searching for certain keywords used in the body of emails can help in extracting the type of relationship between the sender and recipient (e.g. manager and subordinate relationship). Two types of relationship ranking approaches could be used:

1. Traffic Based Relationship Ranking: The sender and recipient details of captured emails can reveal the organizational structure [6]. From a forensics perspective, emails propagated from a certain email id (employee) to any non-relevant id (like not manager or peer) or external id could be marked as suspicious. Moreover, an employee’s search patterns and social media usage could assist to detect a traitor for an inside attack detection [7].

2. Content Based Relationship Ranking: By filtering the emails with certain keywords in the email content can also help in understanding the relationship between the sender and recipients.

An individual’s similarity with another or with a group of people can be predicted on the type of information browsed, pages visited and groups joined. Email communication can also help in identifying the similar characteristics and relationship strength between two or more users. E.g., If a user A sends more emails to User B than User C, it is more likely that user A and user B have stronger relationship than user A and user C [8] at a specific time. This relationship strength could also be because of certain characteristics, shared interests, or similar age/demographics.

Fig. 1. Relationship Strength between users on the basis of emails exchanged.

III. ONLINE SOCIAL NETWORK TOOLS ANALYSIS

This section discusses various tools to gather social networks data and analyse gathered data to extract meaningful feature. There are limitations attached with every tool in terms of varying format of results and inevitable changes in Facebook Graph API. These changes are made to secure the privacy of users but for researchers they either limit the functionality of the Apps or make them completely unusable.

1) Social Snapshot Framework

Online Social Snapshot [9] comprised of 3 parts. 1. Social Snap Shot Client. 2. Automated Browser to add the snapshot client to the target user’s profile. 3. Third Party snapshot application.

Design: Due to the diversity of information available via OSNs, researchers in [9] propose a twofold approach: an automated web-browser in combination with a custom third-party application. The social snapshot application is initialized with a user’s credentials or authentication cookie. In the following, a custom third-party application is temporarily added to the target account. This application fetches the user’s data, pictures, friend list, communication, and more. Information that is unavailable through the third-party application is finally gathered using traditional web-crawling techniques [10]. By automating a standard web-browser and avoiding aggressive web-crawling, social snapshot simulates the behavior of a human OSN user, thus minimizing the risk of being blocked by the social networking site.

Depth of information collection: Fig. 2 shows an example of a social snapshot with depth = 2. For a given user all of her friends are first fetched, followed by the friend’s photos. The single path for photos of the friend’s user illustrates the magnitude of available paths and thus data. Defining a specific social snapshot depth enables us to limit the amount of fetched data. The amount of data grows exponentially with social snapshot depth.

Fig. 2. Example for elements fetched with Social snapshot of depth=2[9].

This research is supported by Australian Government Research Training Program Scholarship.
Social Snapshot introduced novel techniques to efficiently gather data from OSNs that may be used as criminal evidence. It gathered detailed data and made it feasible to link “online evidence” to traditional forensic artefacts. A prototype application for Facebook data collection was also developed and was shared as an open-source license for other researchers and helped significantly in Facebook data collection until the changes in Facebook API on April 2015.

B. Automated Identity Theft on OSN

Automated Identity theft on OSN is also possible with almost no manual / human intervention. This phenomenon has been explained by [11] in their prototype system which automates the identity theft process for four social network sites including Facebook and LinkedIn.

There are two main concepts of this approach:

- **Profile crawling:** a target is selected and his/her profile is crawled on OSN1 (e.g. Facebook) to get maximum information about him/her and respective contacts.

- **Profile Cloning:** The target ID’s existence is checked on other social networks OSN2 (e.g. LinkedIn). If the ID is not present on that OSN2 a clone ID is made containing same information, which, was obtained from the OSN. Request is sent to connect to the contacts of target ID. Which is usually accepted by the contacts assuming that the same person is now on this new network and now wants to be connected.

There are four (4) components of this system:

- **Profile Crawler:** collects information of the individuals with public profiles.

- **Identity Matcher:** looks for the target individual’s user id existence on other social networks. E.g. If user A is present on Facebook, does A also has a profile on LinkedIn?

- **Profile Creator:** creates a clone ID of the target Individual on the different network where he/ she was not present previously. It uses the same information (Profile picture, age, education etc.) for his/her profile obtained via the crawler module.

- **Message Sender:** sends the connection request to the target individual’s contacts (which were obtained from crawler module), usually people accept the request without any suspicion as the similar picture and details of the requestor are used.

- **Captcha Analyser:** is used to break the Captcha (Completely Automated Public Turing test to tell Computers and Humans Apart) of corresponding OSN site. Every site has a different captcha mechanism and requires different methods to break them. The crack of LinkedIn CAPTCHA [11] has not been achieved.

C. Netvizz V1.41

Netvizz is a research purpose tool that gets information from Facebook for in depth analysis. But its functionality is limited to Facebook pages and groups only. The limitations of this tools are that it is only a Facebook page and group specific and doesn’t provide information about a user’s profile[12].

Another limitation is that the target group should be an open group, closed group data could not be retrieved even if the logged in user is a member of the closed group. We have extracted the data for analysis from different open groups on Facebook. Most of these groups are used for buying and selling goods online. The following modules are currently available and are able to extract information:

- **Group data Module:** creates networks and tabular files for user activity around posts on groups

- **Page data Module:** creates networks and tabular files for user activity around posts on pages

- **Page like network Module:** creates a network of pages connected through the “likes” between them

- **Page timeline images Module:** creates a list of all images from the “Timeline Photos” album on pages

- **Search Module:** interface to Facebook’s search function

- **Link stats Module:** provides statistics for links shared on Facebook

The data is extracted in gdf format (a simple text format that specifies a graph) as well as statistical files using a tab-separated format. These files can then be analyzed and visualized using graph visualization software such as gephi (https://gephi.org/). We have shown the “page like network” of an Open Facebook Page which sells baby products and connects with pages that are in similar business. Data is gathered using Netvizz Page Module and visualization is generated with the help of gephi, shown in the Fig. 3:

![Fig. 3. Page like Network by Gephi.](image)

D. NVivo 11 Pro with N capture addon for chrome

NVivo 11 Pro analyses and gives in-depth insight in qualitative and mixed method research[13]. This tool is an additional point for digital forensic because the digital evidences are scattered, unstructured and of various formats. NVivo can gather and analyse data from popular social networks (Facebook, LinkedIn & Twitter), it also helps in discovering the hidden connections which are impossible to find manually[14].
NCapture is a web browser extension, to capture content like web pages, online PDFs and social media for analysis in NVivo 10 for Windows, NVivo 11 Pro for Windows and NVivo 11 Plus for Windows[15].

NCapture extension gathers web pages and online PDFs. It can gather the signed in user’s Facebook wall posts and comments, it also gathers the wall posts and comments from people, organizations or groups. Then import them into NVivo as a dataset source. Our experimentation with NVivo NCapture revealed that it is most appropriate and gives much deeper insight for twitter as compared to Facebook. Searches and data analysis and feature extraction in twitter could be based on a particular word, phrase or hashtag.

Twitter content: Use NCapture to gather Tweets from Twitter for example, Tweets that include a particular word, phrase or hashtag, or Tweets by a particular user. Then import them into NVivo as a dataset source.

LinkedIn group discussions: Use NCapture to gather LinkedIn group discussions that are relevant to your research. Then import them into NVivo as a dataset source. Note: In May 2015 LinkedIn limited access to their web service (API). As a result, you can no longer capture a LinkedIn group discussion as a dataset using NCapture. You can still capture a group discussion, and any other page in LinkedIn, as a PDF.

IV. ANALYSIS OF THE USES OF TOOLS

This section analyzes the tools on the basis of their usability, skills required to operate them, types of data they capture, analyse and visualize.

TABLE I. FEATURES OF TOOLS FOR ONLINE SOCIAL NETWORK INFORMATION ANALYSIS

<table>
<thead>
<tr>
<th>Tools name</th>
<th>Social Snapshot</th>
<th>NVivo</th>
<th>Netvizz</th>
<th>Gephi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of use</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Programming needed</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Captures FB Group/Page Data</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Captures User data</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Well suited for research</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Working after April 2015 FB Graph API changes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* only the logged in user’s wall as a pdf document

TABLE II. CAPABILITIES OF TOOLS IN TERMS OF DATA CAPTURE

<table>
<thead>
<tr>
<th>Features</th>
<th>Netvizz</th>
<th>NVivo</th>
<th>Social Snapshot</th>
<th>Gephi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group / Page Data Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gathers Facebook group data(Open / close)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Analysis of page data</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Capture Images</td>
<td>Links only</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Visual Analysis</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

User Data Analysis

| User profile data collection        | Not now | No    | Yes   | No    |
| Friends details capture            | Not now | No    | Yes   | No    |
| User posts collection              | Not now | No    | Yes   | No    |

V. FACEBOOK USAGE SURVEY

An online survey was conducted to understand the Facebook usage behaviour of individuals and their perception about the security challenges with this social networking site, the applications, quizzes and games associated with it. Moreover the survey also inquired the way people accept requests from other Facebook users and what information they seek prior to accepting any connection request[16]. The majority of respondents aged between 20-40 years, educated and employed individuals. The survey was completed by 60 individuals.

<table>
<thead>
<tr>
<th>Occupation Status</th>
<th>Percentage of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>58.3%</td>
</tr>
<tr>
<td>Self Employed</td>
<td>10%</td>
</tr>
<tr>
<td>Unemployed</td>
<td>11.7%</td>
</tr>
<tr>
<td>Student</td>
<td>18.3%</td>
</tr>
</tbody>
</table>

This survey aimed to analyse the following aspects of user’s Facebook usage:

1. Facebook usage device (Phone, IPAD, Tablet, PC)
2. Do users give location information via check-ins?
3. Users’ understanding of privacy issues
4. Users’ awareness about the type of information obtained by Facebook applications, quizzes and games

The responses collected from the survey suggest that majority of the respondents use Facebook from their cell phones Fig .4. With a variety of platforms available for cell phones, numerous security challenges arises. Cell phone related unaddressed risks combined with social media exposure magnifies the severity of data theft, identity theft and financial losses[17].

Most of the respondents showed that they care with whom they are being friends with but results show that most of the people, post or pre accepting a request do not check that when this profile was created Fig. 5. Anyone can make a fake ID by using the information available on another social network[11]. Afterwards social information can be an aide in a targeted social engineering attack[18].
As shown in Fig. 6, users use the check-in options but not very frequently. Check-in information can be used to understand the routine, and spending behaviour/level of a user.

The information obtained by Facebook apps, quizzes and games is also a big challenge. Although most of the respondents mentioned that they check what information is accessed by an app Fig. 6. But malicious code could be easily attached with links to access these apps, quizzes/games and more information could be captured in the background[19].

Table IV presents value for the survey responses categorized between types of users such as, Employed/self-employed, Unemployed and student. From the data collected through the survey process represents that the employed and self-employed respondents have certain behavior regarding Facebook usage and their personal information disclosure on online media. The individuals with this behavior demonstrate informed decision making traits. They are careful in choosing the people they add in their circle of friends. Moreover they don’t accept the second friend request of people already in their network until and unless they are sure about the authenticity of their friends new ID. This type of people are less vulnerable to identity theft attacks and if they become part of an investigation, it is difficult to get much information from their online social profiles. We also observed that, 42% of employed/self-employed Facebook users mentioned that they check what information an application/game/quiz is accessing from their Facebook profile. And 81.2% of employed/self-employed mentioned that they didn’t accept a friend request if it is coming from a friend already present in their Facebook network as they see such action as a suspicious activity. The survey revealed less privacy cautious behavior towards Facebook usage by the unemployed and student respondents. Only 8.3% of unemployed individuals showed that they check what information is accessed by various applications on Facebook. While 13.3% student respondents showed their cautiousness regarding personal information. On the other hand students showed a more casual approach regarding dual friend requests from their contacts. As only 6.2% of the student respondents mentioned that they would not accept a new friendship request from an existing friend on Facebook.

The results suggest that academic organizations, student portals, supportive groups/organizations for unemployed people need more stringent security controls and need to spread awareness about the pitfalls of OSNs.

### VI. Conclusion and Future Work

With the excessive use of OSNs by individuals of every age, race and background, criminals are now able to get a lot of
information about an individual or a group of individuals / organization. Information available via social networks serves as a good augment for social engineering attacks and targeted phishing attacks. Law enforcing agencies to get digital evidence against criminals/suspects could use this social network activity and resultant data. But a lot of challenges are faced in terms of data extraction from Facebook. The tools with ability to get in depth information usually require the authorization of targeted ID. For research purpose it is important to take consent from users whose accounts are used to extract information. Another challenge is Facebooks rapidly changing APIs and format. Many good research based applications have stopped working because of these rapid changes to protect privacy of Facebook users.

Collaboration between research community and OSNs to enhance social network data extraction, data reduction mechanisms and feature extraction algorithms is required to obtain meaningful court admissible results.

REFERENCES


