Design and Implementation of iMacros-based Data Crawler for Behavioral Analysis of Facebook Users

Mudasir Ahmad Wani[*], Nancy Agarwal[†], Suraiya Jabin[‡]and Syed Zeeshan Hussain[§]
Department of Computer Science,Jamia Millia Islamia, New Delhi-110025

Abstract

Obtaining the desired dataset is still a prime challenge faced by researchers while analyzing Online Social Network (OSN) sites. Application Programming Interfaces (APIs) provided by OSN service providers for retrieving data impose several unavoidable restrictions which make it difficult to get a desirable dataset. In this paper, we present an iMacros technology-based data crawler called IMcrawler,capable of collecting every piece of information which is accessible through a browser from the Facebook website within the legal framework authorized by Facebook.The proposed crawler addresses most of the challenges allied with web data extraction approaches and most of the APIs provided by OSN service providers. Two broad sections have been extracted from Facebook user profiles, namely, 'Personal Information'and 'Wall Activities'. The collected data is preprocessed into two datasets and each data set is statistically analyzed to draw semantic knowledge and understand the several behavioral aspects of Facebook users such as kind of information mostly disclosed by users, gender differences in the pattern of revealed information, highly posted content on the network, highly performed activities on the network, the relationships among personal and post attributes, etc. To the best of our knowledge, the present work is the first attempt towards providing the detailed description of crawler design and gender-based information revealing behaviour of Facebook users.

Keywords: *Online Social Network, Information Retrieval, Data Extraction, Behavioral Analysis, Privacy and Security.*

1 Introduction

Online Social Network (OSN) is a web application which focuses on the social life of netizens and provides them an excellent platform to build a network of relationships and share their social life. OSNs such as Facebook, LinkenIn and Twitter are increasingly becoming popular among internet users and turning to be an essential medium for people's daily activities. Since their structure significantly reflects the real life communities and they contain a huge amount of user's personal and social information, they are of scientific importance to the researchers and disciplines in different domains including marketing, sociology, politics etc. Marketers study OSNs to design a viral marketing strategy (?, ?), sociologists use them to study the human behavior (?, ?) and politicians use them to empower their political campaigns (?, ?).

For conducting any kind of analysis, sufficient amount of data is the preliminary requirement. OSNs like Facebook contain billions of user profiles and the service providers ensure that their data is protected which makes the process of data collection very challenging for researchers. OSN datasets are not publicly available because of privacy reasons and since the data is enormous, collecting it manually makes the task complex and time consuming. However, most popular social networking sites like Facebook, Twitter, Flickr, etc. provide methods for retrieving information from the network through their own well defined Application Programming Interfaces (APIs) like Graph-API[1] , REST-API[2] etc., but these APIs are associated with several unavoidable constraints such as data request rate restrictions, selective data access, etc. Web scrapping provides an alternative solution by automatically extracting the information from the web pages in a systematic manner. Although it can solve the problem of data collection to a great extent but writing a scrapper is itself a challenging task. Generally, social media platforms including Facebook have inbuilt bot detection mechanism (?, ?) which can recognize an automated activity on their platforms and, therefore data collection by means of software programs may lead to the suspension of the user account which is being used for data extraction. The feature of dynamic loading of content via various web technologies (e.g. JavaScript and Ajax) further complicates

[*] mudasirwanijmi@gmail.com
[†] nancy.agarwal02@yahoo.in
[‡] sjabin@jmi.ac.in
[§] szhussain@jmi.ac.in

[1] https://developers.facebook.com/docs/graph-api
[2] https://dev.twitter.com/rest/public

the task since this information is not available in the source code of a web page. Moreover, the call to dynamic content on the web page is mostly triggered by user interactions with the page which implies that there should be a mechanism to automate these interactions in order to load this dynamic content into the parent HTML document. Therefore, there is a need of a tool which is capable of circumventing the API constraints and can overcome the data scrapping barriers.

The paper focuses on the designing of data crawler, IMcrawler for Facebook network that addresses the above discussed challenges and assists the end user in extracting the data in an efficient and convenient manner. Facebook is one of the topmost networking sites and has the most complex privacy policy structure. It's API can be used to extract data of only those users who are already registered with the application. Unlike Twitter API, the Facebook API has to explicitly ask for permissions from its members to access their data. Users privacy settings and privileges granted to the application will decide the data that can be scrapped from their profiles. Facebook has also imposed constraints on the maximum request rate to limit the amount of data being scrapped. The crawler proposed in this paper does not face any of these hurdles and is able to extract every information of any amount which appears on the user profile. It is also capable of interacting with Timeline to load the dynamic content.The complete framework of the data collection is also described with step wise processes followed from crawling the network to obtain the featured dataset in a useful format.

The crawler has been designed to extract two broad sections viz profile information (profile features) and wall activities (post features), particularly from the friends of a profile. Profile information consists of details provided by the users about themselves and wall activities consist of actions performed by the users on the Timeline. The collected information is structured into two datasets to represent the attributes of a profile and it's posts and each dataset is used to perform social and behavioral analysis separately. By analyzing profile features data set, we observed what kind of information people tend to reveal about themselves on the social media and if there exists any gender bias in disclosing their personal details. One of our statistical analyses shows that females tend to more secretive and usually reveal attributes with the intention to establish the connections and stay in touch with their family and friends. We also observed that there are some personal attributes which are highly correlated to each other in terms of revealing. It means if users provide some details on one attribute, it is more likely that they will also mention about other correlated attributes. In the second dataset based on wall activities, we analyzed what type of content people mostly post on their timelines, which activities are highly performed on the network and most importantly, whether there exist any kind of monotonic relationship between different post attributes. The remaining article is structured as follows: Initially, the work relevant to data extraction from social networks and analysis on the information revealed by users have been discussed followed by novel contribution of the work. Then, the design of the IMcrawler and it's working have been explained. In the next section, we analyzed the collected data from various statistical perspectives to produce a number of interesting findings. Finally, the last section concludes the overall work carried out towards the design and implementation of the IMcrawler and the behavioral analysis performed on the extracted data.

Related Work

Given the massive amount of online user data, its extraction is the key challenge for the researchers. A study (?, ?) has presented a detailed discussion on web data extraction techniques along with the application domains in which they are applied. In general, APIs and HTML scrapping are the two popular ways to retrieve data from OSNs. Although APIs provide well organized data but with several restrictions. The HTML scrapping techniques offer alternative solution that can circumvent the limitations imposed by APIs but at the price of technical complexities. The paper (?, ?) presented a semantic-based framework for collecting the social media data using APIs and anlysing it by the open source tools provided by GATE family [3]. The authors in (?, ?) extracted and analyzed the structure of four popular OSNs namely, YouTube, Flicker, LiveJournal, and Orkut. Flicker and LiveJournal networks are crawled using their APIs, Orkut data is collected using HTML scrapping technique, whereas both the techniques have been used to collect YouTube content. Data extraction schemes significantly depend on the policy of online social networks as in (?, ?), the authors have presented the process of extracting personal attributes and the list of top friends from MySpace social network without being its member. MySpace provides a rich source of data to the non members as well, whereas networks like Facebook, Friendster etc. expose either no content or minimal content to the external users. Several studies have been carried out on the Facebook network exclusively, since it is one of the most popular online social networking websites and has the most complex privacy mechanism (?, ?, ?). Netvizz (?, ?)

[3] https://gate.ac.uk/family/

is a Facebook application designed to help researchers in collecting features of a profile including personal networks, groups and pages. However, just like any other API, the working of Netvizz application is also limited by the permission and privacy model of the Facebook service. First, it requires logged in Facebook account. Second, it explicitly asks the user for the permission to access their different data and third, the user can further restrict the data availability to the application by their privacy settings. In (?, ?), the authors have implemented an HTML-based crawler by using PhantomJS, a headless browser [4], to extract the friend's network of Facebook users of a specific region in Macau. In addition, authors have also discussed the technical challenges and their viable solutions while designing the data extractor for OSNs.

OSN sites offer a great medium to its users to share varied amount of their personal details. A number of studies have been conducted to explore the pattern of information disclosed by the users of these websites. The paper (?, ?) examines the kind of information that has been disclosed most on the Facebook profiles and what type of people are likely to disclose it more. The authors categorized the revealed information into three groups viz identity information (schools, jobs, etc), sensitive information (email, profile picture, etc.) and stigmatizing information (religious views, political views, etc). The authors observed that people who mentioned details about their gender, relationship status and age have revealed more information about themselves in comparison to the users who did not provide the said details. Moreover, age and relationship status were found to be the salient features in predicting the disclosure of information. Another study (?, ?) has shown that the users who feel lonely, specifically females, reveal information about themselves that might encourage other users to contact them such as relationship status, address, etc. However, exposing too much on OSNs can cause serious problem to the users due to the severe risks associated with the revealed information such as identity theft, cyber stalking, etc. (?, ?). Given the enormous use and potential danger associated with the information disclosed, it is important for the users to understand the privacy settings provided by social sites so as to protect their information from getting in wrong hands. In (?, ?), authors have conducted a study on the Facebook network of a university campus. They investigated the patterns of personal information disclosed by its students and usage of privacy features offered by the site for limiting the visibility of the content. It has been found that while the personal details are generally revealed, privacy features are hardly concerned.

Unlike the web, which is primarily centered on content, online networking sites are concerned with the user life. These sites allow users to publish information about them, connect to each other, share content, disseminate information and so on. Hence all these activities can be analyzed to observe a variety of trends on the network. In paper (?, ?), the authors have extracted the data from *New Orleans* regional network of Facebook for extracting friendship links and wall activities and studied the evolution of relationship between users on the basis of their interaction on the wall. A study (?, ?) has used python programming language to obtain data from the Facebook network and proposed interaction graph to represent real social relationships. They showed that the active social links of the users are much lower as compared to their total friends in the network.

Contribution of the Proposed Work

Although a number of researchers (?, ?, ?) have crawled the OSN sites, but the least attention has been paid towards providing the technical details of the data collection process and its complexities. The presented IMcrawler positively overcomes most of the demerits coupled with available APIs and the challenges with HTML scrapping. The paper discusses the design and implementation of data collection approach in a well defined, comprehensive and systematic manner. It also introduces a new technique called iMacros for the collection of user-related information from OSN sites. The present work will prove to be fruitful for the upcoming researchers by providing them deep insight into the data collection mechanism and enable them to overcome the challenges faced while data collection. Furthermore, the paper also discusses the several statistical observations and findings from the collected datasets pertaining to user's behavioral analysis. We sampled around 10,000 profiles based on the four metropolitan cities (Bangalore, Delhi, Mumbai, Pune) of India, filtered using current city information from the whole data collected through seed profiles on the Facebook network. Our analysis is divided into two areas: *personal information disclosure* and *user wall activities*. From personal features extracted from profiles, our purpose is to study the pattern of information revealed by users and in particular, we studied how gender affects information revelation. From wall activities, we focused on the several aspects like what content is generally posted on Facebook, what activities are highly performed and the relation between different post features. The contribution of the work is briefly discussed as follows:

- An iMacro technology-based data crawler for the

[4] http://phantomjs.org/

Facebook network has been designed and implemented.

- The whole data collection procedure, its technical aspects and complexities have been explained in detail.
- Behavioral analysis of users based on personal information revealed by them along with their wall activities has been carried out.

Design of IMcrawler

Online Social networking sites contain rich amount of information about its users which is mainly stored in a semi structured format as HTML documents. Before performing any analysis, this huge amount of data present on any OSN needs to be extracted in a systematic and automated manner. Web extraction tools are one of the ways to efficiently collect data from a website with least human effort. There are a number of data extraction approaches such as Natural Language Processing (NLP)- based approaches, Ontology-based approaches, HTML-aware approaches, etc. (?, ?).Here, in this paper, we are discussing the implementation of an iMacros technology based data crawler for the Facebook social network. iMacros is a browser extension specifically designed to automate web interactions and extract data from a website [5]. It allows scrapping every piece of data from a web page which is accessible thorough a browser. iMacros add-on is available for almost all the prevalent Web Browsers including Mozilla Firefox, Google Chrome and Internet Explorer. However, the working may slightly vary from browser to browser. iMacros uses Document Object Model (DOM) tree (?, ?) to identify the data to be scrapped. DOM provides the structured representation of a web page where nodes denote the HTML tags and tree hierarchy represents the organization of the nested elements. Generally, dynamic applications bind the information from the databases with predefined web templates as per the user request. It implies that a large section of HTML structure of a web page remains almost static for different profiles, and only the content of the pages differs. These static elements in a page assist in extracting varied content from multiple profiles.

In order to extract the data, we need to crawl through the targeted OSN. Basically, an OSN is a graph where each profile is considered as a node and friend-links represent edges between the nodes. However, crawling complete graph is generally not possible for a network like Facebook which has billions of profiles (nodes). Hence, a relatively small but representative sub-graph as per the problem to be solved can be considered to serve the purpose. Graph traversal techniques and randomly selecting profiles based on a certain condition, are the two well known techniques used to gather the data sample that describes the network. In this paper, we use Breadth First Search (BFS) as the graph traversal technique because it is easy to implement and has been extensively employed by different researchers for crawling profiles (?, ?). The BFS algorithm starts from a root profile, explores and discovers its friends and then moves to the next level.

In order to speed up the extraction process, we run iMacros scripts on different machines (we call them crawler agents) and remotely access and control these independent machines with a centralized host machine as shown in the Figure1.

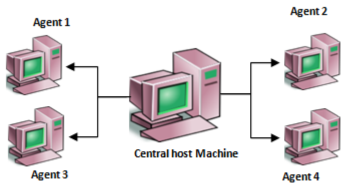

Figure 1: *Parallel Data Collection Approach*

A study in (?, ?) has presented a parallel framework to collect data from online auction websites. In order to further shorten the time for data collection, we created multiple user sessions on the agent machines to run multiple scripts simultaneously.

Crawlers are designed based on the problem under consideration. More extraction of attributes needs more time and technologies. Therefore, one can design an optimized crawler by knowing in advance about the data of interest. As far as our data of interest is concerned, we have extracted most of the sections of a Facebook profile including *basic information, family and relationships, places lived, pages liked, groups joined, post details, post emotions, post shares,* etc. and out of which we are interested in two main categories of attributes namely personal attributes and post attributes. Personal attributes include the number of *friends, birthday, e-mail address, phone number, family members, relationship status, gender, home town, current city,* etc. While the post attributes include the type of the post (*text, photo, video,* etc.), tags in a post, etc.

The complete pseudo code of the IMcrawler is shown in the Algorithm 1. Before starting the extraction process we need to select the seed profiles (registered user accounts) which act as an entry point for the crawler. The proposed crawling approach uses a set of seed accounts which have huge friend network and whose log-in credentials have been collected from their owners.

[5] http://wiki.imacros.net/Tutorials

This enables us to extract the friend-specific information as well. These seed profiles can be one or many, mainly depends upon the amount of data to be required. We manually selected the seed profiles from our friend lists based on the number of friends. We distributed the seed profiles among distinct crawling agents. Every agent extracts the friend links of a specified seed profile and visits each friend link one by one to extract the information.

As shown in the pseudocode, the CRAWL procedure requires two arguments: *seedFile* (contains seed profiles and their credentials) and *configFile* (contains the values of the several parameters to be used during crawling the data). Description of parameters in the configuration file is given in Table 1. The two attributes namely, *reextractLinksFile* and *seedProfile*, are basically used for re-collecting the data of the profiles whose data has not been extracted as expected. The CRAWL procedure reads the *parameters* from configuration file into the *conf* object. If specific *seedProfile* is not given, the data is extracted from all the friend links for each *seedProfile* mentioned in the *seedFile*, otherwise data is collected from the friend links

mentioned in the *reextractLinksFile* by using the user account indicated by parameter *seedProfile*. The extracted data is saved to file specified by the *outputFile* parameter. Basically, extraction process requires three things. First, a source file that contains the URL of target web pages (*friendLinksFile* and *reextractLinksFile*). Second, it requires the HTML tag identifier to locate the data to be extracted from a web page. And third, a file to store the extracted data (*outputFile*).

The presented crawler is not limited to above mentioned attributes only. Researchers can use this data crawler to extract the features of their interest or as per the problem under consideration. The complete framework of the IMcrawler is shown in the Figure 2 that describes the step wise processes from crawling the Facebook network to obtain a sampled dataset in the required format. The whole process is divided into 5 modules. The first module deals with the setting the parameters in the configuration file for the execution of crawler program such as reference to the input files, output file etc. In the second module, the execution of crawler program takes place to collect the data from the network. The output of the crawler program is a raw file which needs to be converted into a structured format before verifying it and this conversion process is done in step 3 (in our case the .csv file has been converted into excel sheet). Step 4 applies several verification tests to ensure whether the data is collected as expected. The extracted data in the files have been verified manually as well. There may be the case when junk data is collected for some profiles due to technical problems like a slow internet connection or machine overload. The data is re-collected by specifying those profiles and the seed account in the *reextractLinksFile* and *seedProfile* respectively. The *reextractLinksFile* is actually a parameter which specifies the file where the links for which the data need to be re-extracted are stored and *seedProfile* parameter indicates the seed profile in the *seedFile* that will be used for re-extraction. Finally, the step 5 filters the information from the mark up tags in the verified file and stores into the database. It may also be possible that seed profiles have mutual friends among them which lead to the extraction of some profile data multiple times. The primary key concept of database helps to resolve it by preventing the storing of same data multiple times.

Behavioral Analysis of Users

Several experiments were performed on the collected data. The data has been partitioned into two data matrices, namely personal information data matrix ($\mathrm{I_M}$) and post content data matrix ($\mathrm{C_M}$). It is to be noted that although our crawler extracts all the attributes mentioned in the pseudocode but there are few features like groups joined by the users, pages like by them, views and shares of the posts, etc. which we are not using in the present research. The data of these attributes will be used in our upcoming research projects. As far as the analysis of rest of the attributes is concerned, we have divided our observations into two sections. In the first section, we were interested to find out the personal information revealed by users on the Facebook network. In the second section, we analyze the activities performed by users on the Facebook.

Algorithm 1 : IMcrawler

```
Require: $seedFile$ and $configFile$
procedure CRAWL($seedFile$,$ConfigFile$)
    $conf$ ←read parameters from $configFile$
    if !$conf.seedProfile$ then
        for all $seedProfile$ in $seedFile$ do
            $profileCredentials$ ←read $seedProfile$ credentials from $seedFile$
            $login$ to $profile$ using $profileCredentials$
            visit $friends\ option$ on $timeline$
            extract $friend\ links$
            save $friend\ links$ to the file $config.friendListFile$
            ExtractPersonalAttributes($conf.friendLinksFile$,$conf.outputFile$)
            ExtractPostAttributes($conf.friendLinksFile$,$conf.totalPost$,$conf.outputFile$)
            $logout$ from $profile$
        end for
    else
        $profileCredentials$ ←read $conf.seedProfile$ credentials from $seedFile$
        $login$ to $profile$ using $profileCredentials$
        ExtractPersonalAttributes($conf.reextractLinksFile$,$conf.outputFile$)
        ExtractPostAttributes($conf.reextractLinksFile$,$conf.totalPost$,$conf.outputFile$)
        $logout$ from $profile$
    end if
end procedure
procedure ExtractPersonalAttributes($friendListFile$,$outputFile$)
    for all $UserLink$ in $friendListFile$ do
        visit $about\ option$ on $timeline$
        extract $basic - information - section$
        extract $places - lived$
        extract $family - and - relationship$
        extract $number - of - friends$
        extract $pages - liked$
        extract $groups - joined$
        save extracted data to the file $outputFile$
    end for
end procedure
procedure ExtractPostAttributes($friendListFile$,$totalPosts$,$outputFile$)
    for all $UserLink$ in $friendListFile$ do
        for $i \leftarrow 1, totalPosts$ do
            extract $post - title$
            extract $post - content$
            extract $post - date$
            extract $post - time$
            extract $post - comments$
            extract $post - emotions$
            extract $post - shares$
            extract $post - views$
            extract $post - reactions$
            save extracted data to the file $outputFile$
        end for
    end for
end procedure
```

Personal Information Analysis

Here, in this section the statistical analysis on user profiles has been carried out to semantically map the personal information revealed by the users to their gender. Let $\mathrm{I_M}$ be the Personal Information data matrix of size $n \times m$, where n denotes the number of unique users in the dataset and m denotes the number of personal attributes under consideration. Table 2 shows the personal attributes present in the $\mathrm{I_M}$ which represent the information provided by users about themselves on their profiles.

The first column in $\mathrm{I_M}$ holds the gender value which can be Male, Female and NR (who have not revealed their gender). All other columns contain personal attributes in binary form. The value is 1 if the information of the attribute has been revealed by the user, 0 otherwise. Each row (X_i) in the matrix is the m dimensional vector representing the attribute values of the particular user which can be written as:

$$\begin{gathered} X_i = (x_{i0}, x_{i1}, ...x_{im-1})\ ,\ \forall\ i = 1..n \\ x_{ij} \in (0,1)\ ,\ \forall j = 1...\ m-1 \end{gathered} \tag{1}$$

Here, X_i denotes the i^{th} user in the data and x_{ij} represents the value of j^{th} attribute of the i^{th} user. x_{i0} holds the gender information of the user and for the rest of the features; the value is 1 if i^{th} user has revealed the respective attribute, 0 otherwise. The personal information has been analyzed in the following subsections.

Proportion of personal information revealed

In this section, our main intention is to find out which attributes are most likely to be disclosed by the users. We divided the disclosed information into three sections on the basis of gender. Let $K = \{$"Male", "Female", "NR"$\}$ be the gender vector. The number of attributes revealed by the users can be calculated as follows:

$$y_{kj} = \frac{\sum_{i=1\ ,\ j_0 \in k}^{n} (x_{ij})}{n} \times 100\ ,\ \forall\ k \in K\ ,\ j = 1...\ m-1 \tag{2}$$

Here, $y_{\mathrm{k,j}}$ represents the percentage of j^{th} attribute disclosed by males, females and NR (as denoted by K vector) out of total users n. ${j_0}^{\mathrm{th}}$ column vector holds the value of gender of the user in the dataset.

Figure 3 displays the percentage of information revealed by the users for each personal attribute in the collected data sample. The information revealed is divided into three groups i.e. the information revealed by the male group, female group and the group whose gender information was not revealed (NR). It is clearly visible in the plot that the female group has disclosed more information in all the attributes. Attributes like birthday (85%), hometown (75%) and number of friends (70%) have been revealed more among all the 14 attributes. While the attributes like email address (3% only), home address (10%) and website address (10%) have been revealed by very few people.

Birthday and number of friends (or simply friends) are considered as the two compulsory attributes in a social network like Facebook. Providing value to birthday attribute is mandatory at the time of registration and value of friends attribute is auto maintained by the website as per the connections. However, users have the privileges to hide these attributes but either they deliberately don't do it or pay less attention to it. Also, birthday is the only attribute which may make you talk of the day among your friends without any effort. And since we have extracted the data from the friends of seed accounts, it is less likely that users hide number-of-friends information from their friends. Therefore, these two attributes can be seen as highest revealed by the users.

Gender based information revealed

The Figure 3 above shows that the number of females providing information is significantly greater than the males. But this information would be incomplete to conclude which gender exactly tends to disclose more personal details on their profiles, since it has not taken into account the ratio of males to females in the dataset. In order to assess exactly which gender is more revealing towards a particular personal attribute, we measure the percentage of males and females who have included the value of the attribute on their profiles out of a total male and female population respectively. The objective is to study the gender differences between the kind of personal content revealed. Let $K = \{$"Male" ,"Female"$\}$ be the gender vector (here we did not include users who have not revealed their gender information i.e. NR). Mathematically, we measure

$$y_{kj} = \frac{\sum_{i=1\ ,\ j_0 \in k}^{n} (x_{ij})}{\sum_{i=1\ ,\ j_0 \in k}^{n} (1)} \times 100\ ,\ \forall\ k \in K\ , j = 1...\ m-1 \tag{3}$$

Here, y_{kj} represents the percentage of males and females who have revealed the j^{th} attribute out of total males and females respectively. The ${j_0}^{\mathrm{th}}$ column vector holds the value of gender of the user in the dataset. The graph in Figure 4 shows the percentage of each attribute disclosed by males and females respectively. The information which is most favored by females includes the number of friends, relationship status, family members and birthday. Semantically, it

Table 1: Description of parameters in the configuration file

Parameters	Value	Description
friendLinks	Path to the file	Contains the path to the file which contains the friend links extracted from the seed profiles.
outputFile	Path to the file	Contains the path to the file where the extracted data is to be stored.
totalPost	Numeric	Specifies the number of posts to be extracted from each user profile.
reextractLinksFile	Path to the file	Contains the path to the file which contains the friend links for which data is to be re-extracted.
seedProfile	Numeric	Value that represents the specific seed account to be used to extract the data.

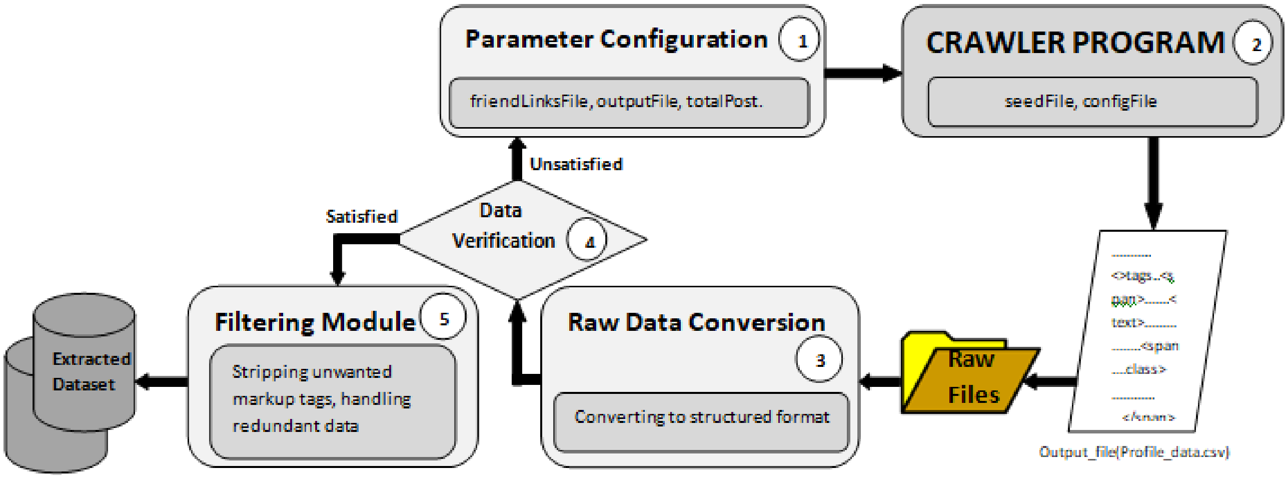

Figure 2: *Framework for Data Collection*

reflects that females are generally using social media to establish connections and stay in touch with friends and family. On the other hand, the information of religious views and political views have been largely seen on male profiles. The observation supports the report(?, ?),which states that men tend to influence the society and enhance their social status.Moreover, it can also be observed that men are more professionally inclined because of scoring higher revealing percentage for attributes like website links, social links and email address. This fact is more supported by the LinkedIn (professional networking site) network where male accounts are more in number than female accounts [6]. Furthermore, the graph also shows that the attributes which are strictly personal, e.g. address and phone number, are usually avoided by females to disclose them openly.

Overall information revealed by Gender

The above two sections have focused on the distribution of individual personal attributes. In this section, we assess the overall information revealed by each gender. Let p be the new column vector that stores the information revealed by each user. Then, the number of attributes a user has disclosed (x_{ip}) can be calculated mathematically as:

$$x_{ip} = \sum_{j=1}^{m-1} (x_{ij}) \ , \ \forall \ i = 1...n \qquad (4)$$

where m is the number of columns in the dataset, n denotes the total number of users and x_{ij} is 1 if j^{th} attribute has been revealed by i^{th} user, 0 otherwise. Our purpose is to measure what proportion of information has been revealed by the males and females out of total personal information space available to male and female population respectively. Let $K = \{$"Male" ,"Female"$\}$ be the gender vector. Mathematically, we measure

[6] https://blog.linkedin.com/2014/04/18/the-next-three-billion

Table 2: Description of personal attributes considered in the dataset

Index	Personal attribute	Value
0	Gender	Male, Female, NR (Not Revealed)
1	Friend Count	0,1
2	Relationship Status	0,1
3	Family Members	0,1
4	Interested In	0,1
5	Languages	0,1
6	Hometown	0,1
7	Birthday	0,1
8	Phone No.	0,1
9	Address	0,1
10	Email Id	0,1
11	Political Views	0,1
12	Religious Views	0,1
13	Social Links	0,1
14	Website Address	0,1

$$y_{kp} = \frac{\sum_{i=1\ ,\ j_0 \in k}^{n}(x_{ip})}{m \times \sum_{i=1\ ,\ j_0 \in k}^{n}(1)} \times 100\ ,\ \forall\ k \in K \qquad (5)$$

where y_{kp} represents the percentage of attributes disclosed by each gender with respect to their total information space. We found that males have disclosed more information to their friends (42.5%) as compared to females (39.7%). From Figure 4, we have observed that the number of personal attributes disclosed more by males is greater than the attributes disclosed by females.

Pattern of revealed information by Gender

In the previous section , we concluded that males are more revealing in terms of quantity of overall information. Here, we are interested to know whether the pattern remains the same or gets fluctuate at different level on information revealed scale. We divide the information scale into the sequence of intervals as 5%, 10%, ... 95% and measured the percentage of males and females who have disclosed their information out of total males and total females in the dataset. Figure 5 shows the pattern of revealed information for both the gender groups. It is clearly visible in the graph that irrespective of gender, as we increased the level of the scale, the percentage of population who are revealing information gets lower. One interesting thing to be noticed here is that initially on the information revealed scale, the pattern of males and females is almost same and between 20% and 30%, the pattern gets fluctuated and proportion of women becomes slightly higher than men. It indicates that tendency of revealing information is somewhat more in women if less information is considered (below 30% in this case) rather than complete.

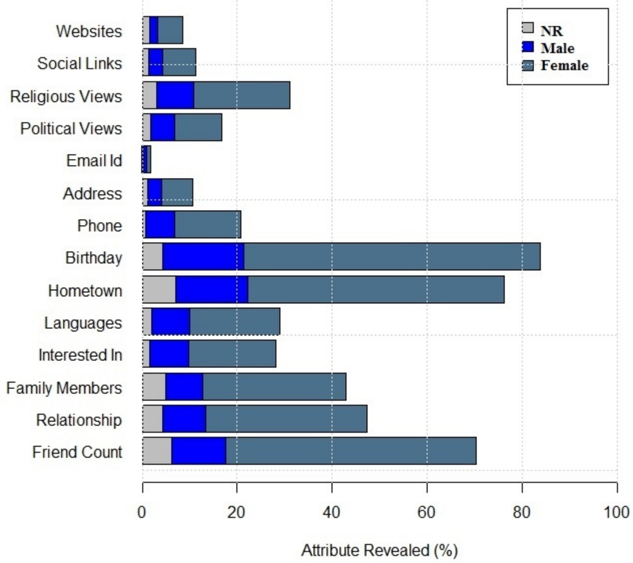

Figure 3: *Proportion of personal attributes revealed by the users and participation of gender vector (male, female and NR) in each disclosed attribute*

Correlation among Personal Attributes

Correlation helps to determine if there exists an association between two variables. We are interested in uncovering whether there exist any relationship between a particular pair of personal attributes of a user profile. In other words, if one attribute is revealed by a user, how likely the user will also disclose another attribute. Since in our case personal features hold binary values, phi measure has been used to compute the association (?, ?). Unlike chi-square, phi does not depend on the sample size and hence can also measure the strength of the relationship between the two variables. Moreover, Pearson correlation coefficient for two binary variables yields the same value as phi. The value of phi coefficient indicates the degree by which two attributes are bound to each other. The value has a range from -1 to 1. Values greater than 0 signify positive relationship while values less than 0 signify negative relationship.

All the possible relationships between the profile attributes have been shown in the Figure 6. The areas of circles show the absolute value of corresponding correlation coefficients. The circles are filled clockwise for positive values and anti-clockwise for negative values. It can be seen in the graph, the strong positive relationships involve 4 attributes, namely, language, interested-in, political views and religious views

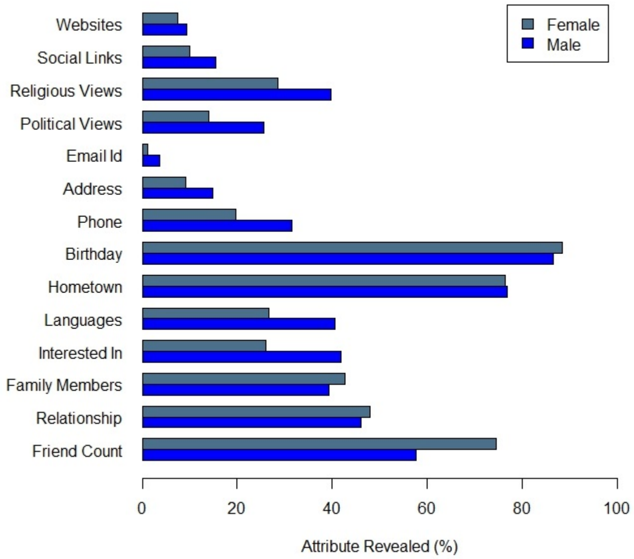

Figure 4: *Proportion of each personal attribute revealed as a function of gender (male and female)*

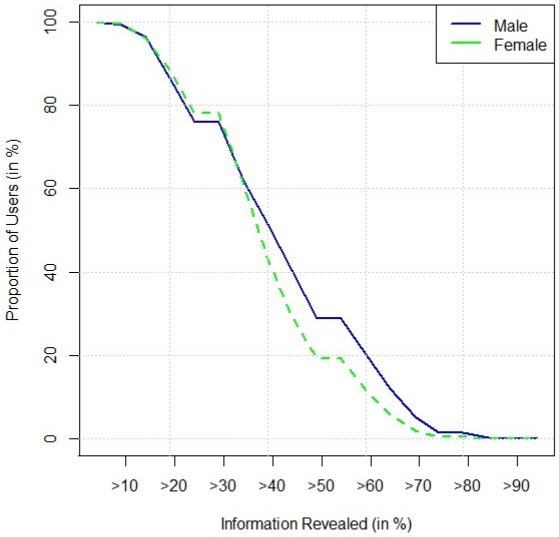

Figure 5: *Pattern of proportion of males and females providing personal details at different levels of information revealed scale*

(p <0.001). One thing to be noted is that all these attributes are the part of basic information section of a user profile. There may be the possibility that attributes in the same section are likely to be revealed together. Secondly, the basic information section on the Facebook is not considered as personal as other sections, thereby increasing the possibility of these attributes to get revealed simultaneously on user profiles. From Figure 6, it can also be seen that religious views and political views are most positively correlated with each other (r=0.56, p <0.001). It implies, if a user gives any opinion about political views, there are greater chances that he will mention about religious views as well and vice versa. Also according to the political philosophy [7], religion and politics go hand in hand in most of the sections around the world. This gives a stronger foundation to our observation that religious views and political views attributes have a strong relation to be revealed together.

User Timeline Analysis

Timeline (previously called wall) serves as one's personal space on Facebook that offer users and others (if permitted) to post content on it for other people to see. The main focus in the section is to map the user-activities to their social life based on the statistical analysis performed on user post attributes. Let the $\mathrm{C_M}$ be the Post content data matrix of size $p \times q$, where p denotes the number of unique posts in the sampled dataset posted on different user Timelines and q represents the number of post attributes. The post attributes include post id, user id, actor, object, action and tags. Table 3 presents a detailed description of post attributes. We have analyzed the post features from different perspectives as discussed in the following sections.

Distribution of Types of Post Content

Based on the content, a post can be a photo, video, text, memory of any past post, event, link to any URL, etc. We identified five most prevailing types of post content, namely, photos, plain-text, videos, links and memories as shown in Figure 7. We observed that photos contribute highest to the kind of content posted on Facebook. This makes sense as photos are supposed to be more engaging, easy to digest and an efficient way to gain the peoples attention. The miscellaneous group includes the kind of content which appear very less on the Timeline such as events, friendversaries (friendship anniversaries), etc.

Distribution of Types of Post Activity

Post activities can also be of several types. Activity can be sharing, uploading, displaying emotions, check-in, traveling and so on. We recognized top four activities which are performed on the Timeline, namely, upload, share, writing text and showing emotions. Figure 8 shows the distribution of activities performed on user Timeline. Upload has been observed as the top most action (46%), which implies, a large proportion of users are interested in uploading photos and

[7] http://www.iep.utm.edu/rel-poli/

Table 3: Description of post attributes considered in the dataset

Index	Post attribute	Description	Possible Values
0	Post Id	Represents unique post on network	Post URL
1	UserName	Represents owner of the timeline	Cover Name
2	Actor	Represents the user who posted on Timeline	User Name
3	Object	Represents the content posted on Timeline	Photo, Video, etc
4	Action	Represents the activity performed on Timeline	Upload, Share, etc
5	Tags	Represents the number of users tagged in a post	Number

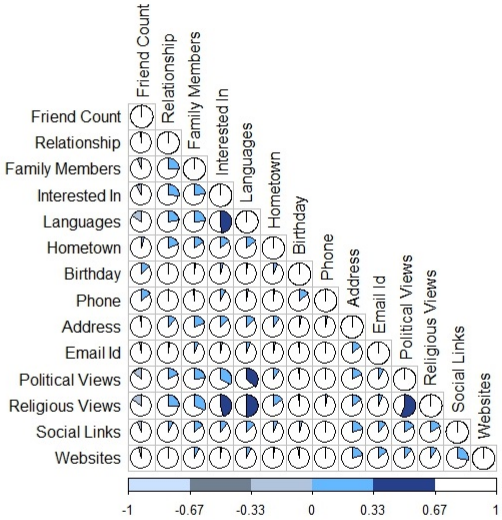

Figure 6: *Pattern of proportion of males and females providing personal details at different information revealed scale*

videos including profile picture and cover image on the Facebook network. Next is the sharing activity (29%) which is considered as one of the best ways to spread the content. People can share any type of content, though photos, videos and links are the most common elements shared on Facebook. Activities like posting a simple and plain-text message on the wall either by the owner or his connections are included in the text activity (18%). A significant number of users also use Facebook as a medium for showing their emotions like feeling happy, blessed, joyful etc. which we have grouped under emotional-activity (4%). All other activities like check-in, flying-to, traveling-to etc. are considered in the miscellaneous group since their proportion is comparatively very low. Table 4 shows the distribution of post activities on the basis of gender.

Correlation among Post Attributes

Similar to profile attributes, we also measured the correlation between post attributes, with different interpretation. We are interested to know if there exists any monotonic relationship among the post attributes.

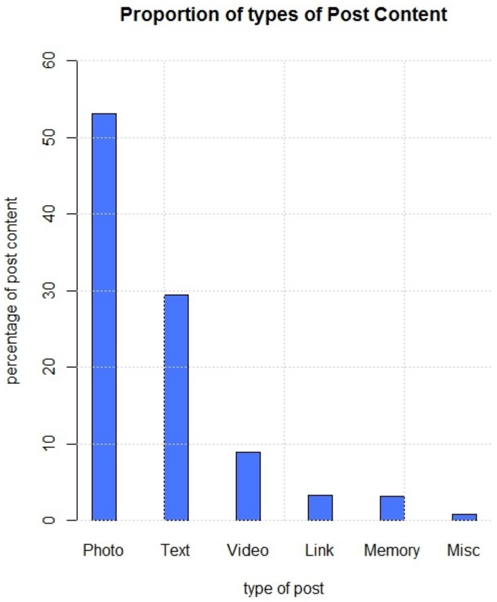

Figure 7: *Distribution of popular content posted on wall*

We determine the correlation between the top five types of content posted on the Timeline, top four activities performed on the Timeline, the proportion of activities performed by the owner(self posts) and the number of tags in the posts. However, we have categorized the text activity into user text (plain-text posted by a user on his timeline) and message (plain-text posted by others on the user timeline). Likewise, video and photo content is merged into one attribute named media. Spearman's rank correlation coefficient (?, ?) has been used to determine the monotonic relationships. Spearman correlation has the advantage of being a non parametric measure which is less sensitive to outliers and is capable to evaluate the degree of association between the variables with non-normal distribution.

Similar to personal features, the relationships among post features have been presented in the Figure 9. Clockwise filled areas represent positive relationships whereas anti-clockwise filled areas represent

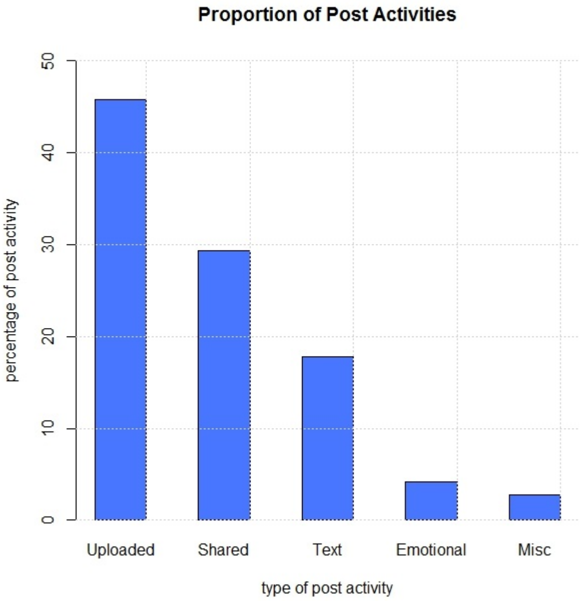

Figure 8: *Distribution of popular activities performed on wall*

Table 4: Distribution of post features on the basis of gender

Activity	Male	Female
Shared	37.31%	30.02%
Uploaded	26.72%	29.23%
Emotional Status	4.94%	4.20%
Text	7.56%	9.70%

negative relationships. It can be seen in the graph, self posts are negatively correlated with tags (r=-0.53, p <0.001) and messages (r=-0.38, p <0.001). The relationship between self-posts and tags implies that there is a decrease in the proportion of self activities with the increase in the number of tags. From the actor's point of view, the posts have been divided into self- posts (content posted by the user on his timeline) and other-posts (content posted by others on the user timeline). Posts which are without tags will appear on the user Timeline only, while the posts with tags will also appear on the Timelines of users who have been tagged in the post. It implies that the number of other-posts will increase on the network with the increase in the numbers of tags in the posts, which can cause the proportion of self posts to decrease. However, the negative relationship between self posts and messages is very clear since messages include plain-text which have been posted by others on the user's Timeline. It can also be noticed in the Figure 9 that plain-text is more positively correlated with the message (r=0.5, p <0.001)) than the user text activity (text label as shown in the graph, r=0.12, p <0.001). It indicates that plain-text content on the user wall is posted more by the user connections than the user by himself. Birthday wishes can primarily be the reason when the friends are implicitly forced to write on one's Timeline.

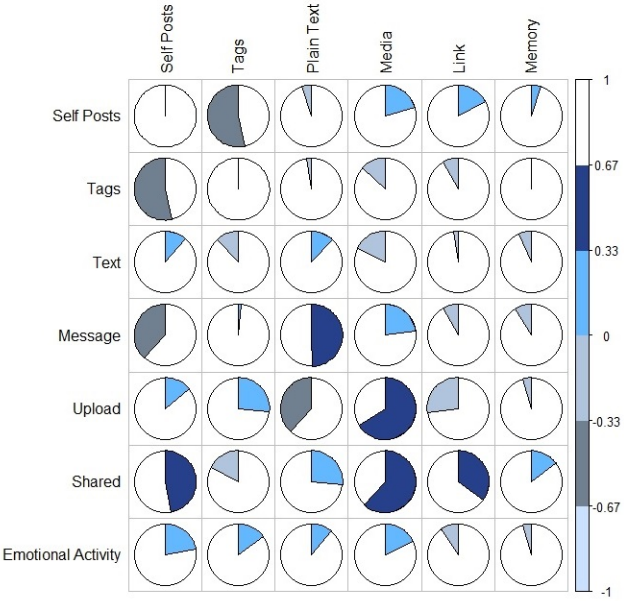

Figure 9: *Correlation among top activities (text, message, upload, share, emotional activity), top content (plain-text, media, link, memory), tags and self posts*

Self posts are found to be most positive correlated with share activity (r=0.47, p <0.001) among all other actions performed on timeline indicating that user has a tendency of sharing the content more than any other activity on their walls. Moreover, the share activity shows the strong positive correlation with media (r=0.68, p <0.001) and links (r=0.35, p <0.001) which means that users are spreading more photos, videos and links on their wall. Furthermore, since upload action is related to photos and videos uploaded by a user, the correlation plot shows upload activity is positively associated with media (r=0.71, p <0.001) while negatively with plain-text (r=-0.38, p <0.001).

Summary and Conclusions

As OSNs are a huge source of information about people, analyzing social networks may result in interesting findings and useful knowledge. But unfortunately researchers are facing a number of challenges while analyzing the social networks because of the deficiency of ground truth data. Although OSN service providers have provided APIs to extract the data but untowardly with several unavoidable restrictions which makes it difficult to get a problem specific dataset. In this paper, we have designed and implemented an iMacros technology based data crawler called IMcrawler for Facebook network and conducted behavioral analysis on the extracted user data. IMcrawler is independent of any API and overcomes most of the challenges faced by the current researchers while extracting the data from the OSN sites.For the proposed work, we have filtered more than 10,000 profiles from the extracted data based on their current city information. We analysed the collected data to draw several behavioral aspects about the users based on the personal information disclosed and the prevalent wall activities performed by them.

As far as our analysis is concerned, we observed that Birthday and Friend count (total friends) attributes have been highly revealed among Facebook users whereas attributes like Email address and Home address have been least revealed. We also observed that Birthday and Friend count attribute have been disclosed more by female users whereas Political views and Religious views attribute by male users.However, if we consider the whole information (all attributes at once), males have proved to be more revealing. Interestingly on scaling down the level of the amount of information revealed, we found at certain levels (20% to 30%) female group becomes more revealing than males. Furthermore, we have also identified some associations among profile attributes, out of which political view and religious view have a strongest positive correlation with each other. From the user's Timeline point of view, photos, text and videos are the top most content posted on the Timeline whereas upload and share are the highly performed activities. We also identified the monotonic relationships between several post attributes such as the increase in the number of tags in a post decrease the proportion of self activity on the network. Also, the positive correlation between self posts and shares implies that users are more involved in spreading the content than other activities (uploading, showing emotions, etc). In addition, the sharing activity is found to be most positively correlated with media and links.

The scope of presented data crawler is not limited only to the attributes discussed in the paper. Researchers will be able to conveniently extend the IMcrawler as per their requirements. Although there are some extracted attributes that have not been used in the present study but they play a significant role in our upcoming research work. Furthermore, the collected data can be analyzed from different perspectives like sentimental analysis of user posts, identification of anomalous user behavior, etc.

References

Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In *International workshop on privacy enhancing technologies* (pp. 36–58).

Alim, S., Abdul-Rahman, R., Neagu, D., & Ridley, M. (2009). Data retrieval from online social network profiles for social engineering applications. In *Internet technology and secured transactions, 2009. icitst 2009. international conference for* (pp. 1–5).

Al-Saggaf, Y., & Nielsen, S. (2014). Self-disclosure on facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behavior*, *36*, 460–468.

Bilge, L., Strufe, T., Balzarotti, D., & Kirda, E. (2009). All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on world wide web* (pp. 551–560).

Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Crawling facebook for social network analysis purposes. In *Proceedings of the international conference on web intelligence, mining and semantics* (p. 52).

Chau, D. H., Pandit, S., Wang, S., & Faloutsos, C. (2007). Parallel crawling for online social networks. In *Proceedings of the 16th international conference on world wide web* (pp. 1283–1284).

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, *32*(4), 556–577.

Dwyer, C., Hiltz, S., & Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, 339.

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323.

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010). Walking in facebook: A case study of unbiased sampling of osns. In *Infocom, 2010 proceedings ieee* (pp. 1–9).

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 acm workshop on privacy in the electronic society* (pp. 71–80).

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003). Dom-based content extraction of html documents. In *Proceedings of the 12th international conference on world wide web* (pp. 207–214).

Karpinski, A. C., Kirschner, P. A., Ozer, I., Mellott, J. A., & Ochwo, P. (2013). An exploration of social networking site use, multitasking, and academic performance among united states and european university students. *Computers in Human Behavior*, *29*(3), 1182–1192.

Kondor, D., Pósfai, M., Csabai, I., & Vattay, G. (2014). Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PloS one*, *9*(2), e86197.

Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM Sigmod Record*, *31*(2), 84–93.

Maynard, D., Roberts, I., Greenwood, M. A., Rout, D., & Bontcheva, K. (2017). A framework for real-time semantic social media analysis. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th acm sigcomm conference on internet measurement* (pp. 29–42).

Nosko, A., Wood, E., & Molema, S. (2010). All about me: Disclosure in online social networking profiles: The case of facebook. *Computers in Human Behavior*, *26*(3), 406–418.

Rieder, B. (2013). Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th annual acm web science conference* (pp. 346–355).

Staab, S., Domingos, P., Mike, P., Golbeck, J., Ding, L., Finin, T., … Vallacher, R. R. (2005). Social networks applied. *IEEE Intelligent systems*, *20*(1), 80–93.

Stein, T., Chen, E., & Mangla, K. (2011). Facebook immune system. In *Proceedings of the 4th workshop on social network systems* (p. 8).

Vergeer, M., Hermans, L., & Sams, S. (2013). Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style. *Party Politics*, *19*(3), 477–501.

Vermeren, I. (2017). *Men vs. women: Who is more active on social media?* https://www.brandwatch.com/blog/men-vs-women-active-social-media/. (Accessed: 2017-04-2)

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in facebook. In *Proceedings of the 2nd acm workshop on online social networks* (pp. 37–42).

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., & Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th acm european conference on computer systems* (pp. 205–218).

Wong, C.-I., Wong, K.-Y., Ng, K., Fan, W., & Yeung, K. (2014). Design of a crawler for online social networks analysis. *WSEAS Transactions on Communications*, *3*, 264–274.