Extract Knowledge from Social Networks

THESIS

submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Hariton A. Efstathiades
born in Nicosia, Cyprus

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Center for Advanced Studies
IBM Netherlands
Johan Huizingalaan 765, 1066 VH
Amsterdam, the Netherlands
http://www.ibm.nl
Extract Knowledge from Social Networks

Author: Hariton A. Efstathiades
Student id: 4182480
Email: h.efstathiades@student.tudelft.nl

Abstract

The easy access to Online Social media platforms through a variety of powerful devices gives to different categories of people the ability of interaction with them. With the increasing popularity of Online Social Networks new opportunities arise for the researchers. Given the enormous volume and variety of information that is being transferred among the users, the collection and analysis of these data could provide useful results for different research areas; from social science and transportation to education and economics. Online Social Networks became a rich resource of information that can be used for knowledge extraction in benefit for the side who is using it.

The knowledge extraction using data retrieved from Online Social Networks is mainly done through a manual-programming procedure with the use of Application Programming Interfaces. However, the majority of social science researchers has usually low programming experience. Thus, the procedure from data retrieval to knowledge extraction is not a trivial task for them. With this research we aim in the development of a framework that provides the researcher who has low programming experience with the ability of extracting knowledge from information that is mainly available in Online Social Networking platforms. Additionally, we present a case study where we evaluate the existing framework by extracting knowledge for a private organization from a professional social network, contributing new insights to the study of social networks.

Thesis Committee:

Chair: Prof. Dr. Ir. Geert-Jan Houben, Faculty EEMCS, TUDelft
University supervisor: Dr. Ir. Alessandro Bozzon, Faculty EEMCS, TUDelft
Company supervisor: Robert-Jan Sips, Center for Advanced Studies, IBM Netherlands
Committee Member: Dr. Ir. Alberto Bacchelli, Faculty EEMCS, TUDelft
Preface

I would like to thank my supervisors from Web Information Systems Group, assistant professor Alessandro Bozzon and professor Geert-Jan Houben, for the continuous support they have provided me during the work on this research project. I thank them for the understanding that they have showed for the various problems that I have encountered at certain times and for their help in solving them. The opportunity that they provided me to work in a research lab was the trigger for my exposure to high-quality research, an important ingredient for the completion of this thesis project. Moreover, I would like to thank assistant professor Alberto Bacchelli for his participation in the thesis committee.

Additionally, I would like to thank Robert-Jan Sips, my supervisor from IBM Netherlands, for giving me the opportunity to work in a professional environment such as the Center for Advanced Studies in IBM Amsterdam. Dr. Alessandro Bozzon and Robert-Jan Sips were by my side from the first until the last step of this thesis providing me with daily guidance in this research project. I was very lucky for cooperating with them gaining value from their background and their experience. Special thanks to Pierre De Wit and Manfred Overmeen (Center for Advanced Studies, IBM Netherlands) for the useful discussions and their feedback for my work. I would also like to thank all the employees of IBM Netherlands who participate in this research, without any personal gain.

Moreover, I would like to thank the students who participate in Omicron meetings for their feedback and insights. Especially, I would like to thank Orestis Gartaganis, for the valuable and selfless help he offered to me. Finally, I would like to thank my family and my friends for their advices and guidance.

Hariton A. Efstathiades
Delft, the Netherlands
December 30, 2013
Contents

Preface iii

Contents v

1 Introduction 1
  1.1 Background .............................................. 1
    1.1.1 Social Networks ...................................... 1
    1.1.2 Social Network Types .................................. 1
  1.2 Motivation .............................................. 2
  1.3 Research Objectives ...................................... 3
  1.4 Document Structure ...................................... 3

2 Related Work 5
  2.1 Users’ motivation in their interaction with social networks and their psychological characteristics .............................................. 5
  2.2 Prediction Knowledge ...................................... 6
  2.3 Business value and knowledge ................................ 7
  2.4 Discussion .............................................. 9

3 A Framework for Social Network Analysis 11
  3.1 Dataset Collection ...................................... 12
    3.1.1 Permissions Granter .................................... 14
    3.1.2 Data Retriever ......................................... 14
  3.2 Preliminary data process .................................. 18
  3.3 OSN Data Analysis ...................................... 19
    3.3.1 OSN Analysis operations ................................ 19
    3.3.2 Supported tools ........................................ 20
  3.4 Discussion .............................................. 22

4 Case Study: IBM 23
  4.1 Experimental Methodology .................................. 23
    4.1.1 Research Questions .................................... 23
    4.1.2 Approach .............................................. 26
  4.2 Experimental Setting ...................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1</td>
<td>Data sources</td>
<td>27</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Dataset</td>
<td>28</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Tools and methods used</td>
<td>32</td>
</tr>
<tr>
<td>4.3</td>
<td>Findings - Results</td>
<td>33</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Research Question 1: Do employees use different social networks for different purposes?</td>
<td>33</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Research Question 2: How can we extract business knowledge from professional social networks?</td>
<td>41</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Research Question 3: How can we identify distinguishing properties of different categories of employees from professional social network profiles?</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions - Future Work</td>
<td>59</td>
</tr>
<tr>
<td>5.1</td>
<td>Conclusions</td>
<td>59</td>
</tr>
<tr>
<td>5.2</td>
<td>Future Work</td>
<td>60</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Future work for presented framework</td>
<td>60</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Future work for case study</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td><strong>Bibliography</strong></td>
<td>63</td>
</tr>
<tr>
<td>A</td>
<td>Survey</td>
<td>67</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

1.1.1 Social Networks

From the early 90’s until today, World Wide Web had a significant impact on almost all the areas of our everyday life; from the way information is created and transferred up to the way in which people communicate with each other. As technology evolves and becomes more affordable, the growing use of the World Web created increasingly more possibilities and needs. This fact led to the transition from a static way of passive consumption of information, the so-called Web 1.0, to the current annual interactive Internet, where every user can easily publish his/her own content and get in contact with other users.

One of the most important features of Web 2.0 is the spread of so-called Social Media and Social Networking Sites [8]. A subgroup of these online social services were designed exclusively for the adoption or preservation of existing social relations between people, while others aim to facilitate the sharing of files that have been created by users-members (User-Generated Content Sharing Sites). The popularity of those services, has offered a variety of possibilities and opportunities for the users, not only to interpersonal and creative level, but also in a professional level. Therefore, the study of these social networks and the interactions of the users in them are in high interest for the research community.

1.1.2 Social Network Types

Online social networks were mainly design for interaction. However, as in real life, different types of interactions and communities exist. For example, one may have the community of his/her friends with whom interacts in a different way than his/her colleagues. This fact lead to the design of different categories of online social networks, which are presented in this subsection [12].

Social - Personal oriented

The first category is the "Friendship Oriented" social networks, which are mostly being used to keep in touch with friends, relatives, family members etc. In these platforms
1.2 Motivation

Introduction

- **Users** tend to share information about their private personal life, such as demographic (age, ethnicity, gender, relationship and marital status), psychographic (hobbies, interests, feelings and life events), location (the place that they live, places that they visit or they are going to visit). Online social network platforms in this category are Facebook and StudiVZ.

**Professional - Career Oriented**

Social networks of this category are designed to provide easy access to career opportunities. These social networks are usually platforms where professionals can find their future employees and partners. Additionally, this platform can be used from job or employees seekers. Professionals can also keep in touch with people they know from their job industry (like classmates, colleagues, bosses, partners, etc) who could be a connection to a new career opportunity. They are considered as social networks because the user maintains his/her social graph, which contains connections with his/her professional environment. Such platforms support similar functionalities as Social - Friendship oriented social networks. However, recent studies show that those two categories have significant differences in the way that people use them [23]. Specifically, while personal oriented social networks are being used for socialize and entertainment purposes, career oriented are being used for professional identity maintenance and opportunities investigation. The most popular worldwide social network in this category is LinkedIn, while others like XING are popular country-wise (Germany).

**Communities of Interest**

Social networks of this category aim to bring together people who have the same interests. These networks were designed based on the hypothesis that people who have the same interests are also interested in interacting together. The people who belong in the same community of interest share their ideas about their passion but usually they do not know much information about each other life. Examples of platforms of this category are MySpace and Dogster.

1.2 Motivation

During the recent years a large amount of information created by users’ interaction is out there and is increasing rapidly every day. The data collection is mainly done through a manual-programming procedure with the use of online social networks’ Application Programming Interfaces. However, a social science researcher has usually low programming experience. Additionally, he/she wants to enrich the collected dataset with external resources or results and, in the most efficient way, preliminary prepare the data for use in the powerful data analysis tools that exist. With this research we aim in filling this gap by providing the researcher with a framework that will help him to extract knowledge from social networks.

Despite the fact that there are tools that can be used for social network data analysis, there is no framework that provides the researcher with all the data analysis steps and automates the procedure of knowledge extraction. We design and develop a framework that will help a researcher to build his/her dataset and analyze it using a powerful
data analysis tool that exist. Moreover, to test the features of the framework we conducted an experiment in order to extract knowledge from a professional social network. The motivation of this case study is presented in chapter 4.

1.3 Research Objectives

The high-level question that we aim to answer with this research project is the following:

RQ: How can we design a framework for the extraction of knowledge from multiple social networks, enriching them with information retrieved from other resources and analysis results?

To answer this research question we present a framework that provides the methods and the steps that should be taken in order to investigate and retrieve the knowledge that available in a social network. With the proposed framework a researcher with low programming experience will be able to collect a dataset from a social network, enrich it with information retrieved from external services, preliminarily analyze it and extract knowledge from the data analysis using existing powerful tools.

We performed a case study in IBM, where we define a set of research questions that will gain business knowledge to the organization by analyzing data retrieved from a professional social network and enriched with external sources and data analysis results. With the use of our proposed framework we have targeted a group of IBM employees to collect a dataset and analyze it in order to investigate and answer the research questions of our case study. The research questions, procedure followed and results of the case study are presented in chapter 4.

1.4 Document Structure

The rest of this document is structured as follows: Section 2 gives an overview about related work and existing research projects in this field. Section 3 presents in detail the proposed framework. Section 4 demonstrates the framework through a case study in IBM, where we present in detailed the procedure that we followed and the results of our analysis. Section 5 gives an insight to the work that could be follow and concludes the document.
Chapter 2

Related Work

Given the enormous volume and variety of the transferred information among communities of users, an interesting opportunity arises: to collect and analyze data that could conclude in useful information for a variety of study fields such as economy, market, social science, business, education etc. Social networks, referring either to physical or online communities, could be used as a rich source of knowledge extraction. Researchers’ communities have shown special interest in the last years in studying a variety of social network communities in order to extract any kind of valuable knowledge according to their studying interest.

2.1 Users’ motivation in their interaction with social networks and their psychological characteristics

A large part of these types of analysis, focus their interest in the reasons that lead users to participate in social networks and especially in the production and sharing of files. One of them, is the work of Nov et al. (2009) [21], which, by having their main focus in the photo-sharing communities and specifically Flickr, attempt to find the factors that have correlation with the behavior of the users, i.e. the number of photos that each of them uploads. The studied factors are related with the personal motives (such as pleasure or acquisition of reputation etc.) and structural properties of the community in Flickr (like the age of user’s profile and the number of contacts they may have). Based on the results of the analysis, there is a decrease in users’ contribution over the time, starting from the time of their enrollment in the site. On the other hand, users who have larger number of contacts as well as those that show a greater dedication degree in the community tend to share greater number of photos.

Similar is also the goal of Huberman’s et al (2009) [15] study, from which it follows that the interest accruing from the sharing files, where in Youtube case is the video, operates as a reward for users who produce it, by increasing their productivity. Vavliakis et al. [32] attempted to expand the studies of Nov et al. [21] and Huberman et al. [15] by developing a methodology for identifying the factors which are affecting the activity of the social networks users. For their research they have used their methodology in data retrieved from Flickr social network. These factors are related with user’s personality data, as they result from the incentives of sharing more files and specific events. Similarly, Ellison’s et al. (2010) [11] studied the Facebook
2.2 Prediction Knowledge

Studies that focus on the factors, which affect the popularity of the shared files, have also special interest. Szabo and Huberman (2008) [29] showed a method for long term popularity prediction of online content, by measuring the initial activity of the users. Specifically, they showed that by measuring the access to the online shared files within relatively small period of time after their publication, it is possible to predict their future popularity. By using Digg and Youtube as examples, they showed that these predictions are more accurate in cases of networks where the content is temporarily interesting.

Cha et al. (2007) [6] focus their interest on Youtube and other similar networks for sharing content produced by users. They highlight the need for approaching this study field with different methods than those that were traditionally used to predict the popularity of content that are not produced by users. Indeed, the popularity of these files is more ephemeral and, as they concluded from the analysis, shows exponential form with truncated edges. After that, they also examined more efficient systems for sharing content produced by users, such as P2P (peer-to-peer) and caching methods. Moreover, they investigate the possible advantages of using these technologies in the networks that they were examining. Finally, they noted the spread of sharing illegal or duplicated content and the mostly negative impacts of that.

In the case of Twitter, the processing of the strict structure of 140 characters messages and its effective functionality as a communication platform in real time, has aroused the interest of the scientific community from multiple different research areas. The examples have been presented on a variety of topics such as predicting earthquakes, hurricanes and other natural disasters (Sakaki et al. 2010) [24], to support actions such as e-government and e-participation in combination with open government data (Kalampokis et al. 2011) [16] and prediction of revenues for new films.
(Asur et al. 2010) [2].

Another related research is the one of Tumasjan et al. [30], in which the authors analyzed Twitter’s messages mentioning parties or politicians prior to the German federal election 2009. They found out that Twitter is indeed used as a platform for political deliberation and could be used as a valuable election predicting tool, as the majority of the tweets, during elections period, reflects voter preferences and comes close to traditional election polls.

2.3 Business value and knowledge

Professional social networks, like LinkedIn, belong to another interesting research area of social network community. Sharma (2010) [25] analyzed the LinkedIn data which is a valuable information repository, and presented the 10 most overused words in LinkedIn profiles.

Case’s et al. (2012) [5] focused on the useful information we could get from analyzing the data from a professional social network and specifically LinkedIn in combination with real data. The investigation concentrated on studying the entry-level jobs held by alumni, their subsequent career progress, and in extend the long-term outcomes of the alumni study program. For that experiment, authors have used the IS program alumni database of a university in the south-eastern US and a LinkedIn group that they created for the needs of that research. Specifically, they aimed to find and invite as much as possible alumni to join their LinkedIn group and study their profile data. Finally a total of 175 IS alumni profiles, were harvested from LinkedIn and have been studied. However, numerous factors have limited the signification of their conclusions from the analysis of retrieved data. The most significant limitation was the wide variation in the completeness and richness of LinkedIn profiles that IS program alumni have created. Although, there were a few fully filled profiles which provided significant insight into the professional development; the majority of the other profiles were incomplete. Information about the entry level job was often missing and only information about current job position was available. Another potential limitation was the fact that LinkedIn profiles contain self-reported data which cannot be easily verify. Moreover, limited female alumni’s profiles were found. A lot of female alumni could not be identified, possibly because they are currently using their married names instead of maiden names. Additionally, the investigation focused only on IS program alums, so its results may not be generalized to graduates from other computing degree programs. Despite the limitations, this investigation suggests that LinkedIn profiles of program graduates may be a valuable source of information for examining the career progress of IS professionals and testing career progress models. According to this research results, the first jobs accepted by IS alumni after graduation are likely to be technical in nature. However, within five years of graduation, numerous program graduates migrate from more technically focused jobs to more business and management oriented positions. It was also shown, that ten years after graduation graduates migrated toward increasingly more responsible management positions. Those insights can potentially help academic advisors guiding students in a more constructive way. Besides that, these information supports the view that within the IS curriculum there should be a balance between management/business and practical/technical skills. In
addition, alumni LinkedIn profiles could be used as an assessment of the program success regarding to its participants career success.

Social network analysis within an organization could provide useful information such as business value, how to improve knowledge creation and sharing within the organization, how to make a more effective employees’ structure etc. The study conducted by IBM Institute for Knowledge-Based Organizations [9] on a group of people in an organization conclude to valuable knowledge on how to improve knowledge creation and sharing within the company. According to this research social network analysis of an important group of people in an organization allow us to understand how the information flows within the organization, to whom people turn for advice and if there is an effective sharing between subgroups and departments. Authors performed a social network analysis of a petroleum organization, by analyzing it into two diagrams; the first one according to the formal organization structure and the second one based on the communication and information flow perspective. One of the most interesting results of the social network diagrams analysis was the identification that there was only one person who was the single point of contact between members of the production division and the rest of the network. That person, due to a facilitated session with this executive team, revealed that over time his reputation for expertise and responsiveness had resulted in becoming a critical source for all sorts of information. As a result, the number of informational requests he received and the number of projects in which he was involved grew excessively, which not only caused him stress but also frequently slowed the group down as a whole because he had become a bottleneck. Regarding to the social network analysis authors conclude that knowing what someone knows, gaining timely access to that person and trusting him are the factors which affect the knowledge sharing relationship within an organization network. Questionnaires and interviews were the main method for collecting the necessary data that had been used for the needs of that research.

The need for social networking platforms for companies and organizations lead to the development of protected-internal enterprise social networks, which can be used as a source of knowledge for the working environment. Wu et al [34] focus on analyzing the behavior of the users inside such a social network having as purpose to determine the type of interaction patterns that reveal real-life closeness between the colleagues. They investigate whether the relationship closeness can be predicted from these interactions by comparing them with users’ perceptions about their relationships with colleagues. Their results suggest that from the analysis of these interactions such prediction is possible.

Cao et al [4] take advantage from the existence of such network in a large international organization in order to study how the different professional relations between nodes have on impact on their interaction in an enterprise social network in large scale. Additionally, they investigate whether user’s geographic location and rank in company’s hierarchy have a relation on the way that the user uses the social network. It is concluded that users’ that hold a position higher in the hierarchy are more likely to receive replies from other users, as they receive more attention because of their influence. Moreover, a pair of users who share the same nationality or have small hierarchical distance are more likely to interact than a pair with different nationality and large hierarchical distance.
2.4 Discussion

These works show that researchers are interested in extracting knowledge from online social networks. Based on these facts we were motivated in designing and developing a framework that will help them in collecting their data, enriching and analyzing them using powerful analysis tools in order to be able to draw into conclusions. Our framework is based on the methodologies that have been used in the presented research projects. We strongly believe that the use of the proposed framework will help future researchers in their studies for extracting knowledge from online social networks.

For the case study our investigation is related to business knowledge extraction by studying the social network of IBM employees through LinkedIn online social network. IBM Institute for Knowledge-Based Organizations [9] study gave us the inspiration to work on business knowledge extraction by analyzing the social network of an organization enriched with data retrieved from external resources. Like Case’s et al. (2012) [21] we use LinkedIn and study profiles data of a group of people within IBM organization but at the same time we use questionnaires and interviews as well.
Chapter 3

A Framework for Social Network Analysis

Social networks popularity rapidly increases on a daily base. Due to their user friendly approach social network platforms attract users who belong to different categories according to their age, education, job, geographic location etc.. For this reason, they are an important source for data retrieval and analysis in order to drive into conclusions or validate theories in various fields such as social science, market analysis, transportation. However, the use of this powerful resource and the enrichment from other resources, is not a trivial task for a researcher with no programming experience.

The purpose of this framework is to provide the researchers with the ability of extracting knowledge using data retrieved from social networks. With this framework we aim in providing functionalities for data retrieval, data storage and data analysis that will help in the social networking analysis. The proposed framework also supports the data enrichment through other online or offline resources and the analysis with the use of existing powerful statistical analysis tools.

In this chapter we describe in detail our framework, the functionalities that it provides and the procedure that should be followed to extract knowledge from social networks. We present the characteristics of this framework and technical details for its implementation. Additionally, the purpose and features of each element are also described.

Framework Definition

In figure 3.1 we can see a diagrammatic representation of our proposed solution. It consists of the following elements:

- Users: The people who interact with online social networks and share their content
- Social Networks: Active social networking services
- Permissions Granter: In order to access the huge amount of data that is stored in those services, we need the permission of the user. Permission granter is an application that give us the ability to get permissions from the users
3.1 Dataset Collection

The increasing number of online social network users produces huge amounts of data with rich user interactions. This information is of high interest for researchers who want to retrieve and analyze it. There are two approaches to extract this information: 1) crawling using service provided Application Programming Interfaces (APIs), or 2) crawling needed information from rendered HTML pages [1]. For our proposed solution we chose to follow the first approach as the majority of online social networking services provide APIs (Twitter, Facebook, LinkedIn, Flickr, etc.).

The main challenge that we have to address in this step is how can we target the users in order to get permissions to access their social network profile. Our framework should be able to address the challenge and handle problems that can occur. For that reason we have designed and implemented the Permission Granter element. This

- Data Retriever: Element that provides access to users’ social networking data
- Data repository: Pre-processed data that have been retrieved from online resources
- Preliminary analysis: The element that provides us data in a format that is able to be analyzed
- Online Social Networks (OSN) Analysis: Analyzes the data in order to extract knowledge

Figure 3.1: Framework’s schematic representation.
element provides the researcher with the functionality of retrieving permissions from social network users in order to access their data. Additionally, it is designed in such way that, after some configurations, can be integrated with different online social networks’ APIs.

Procedure - Challenge

Before proceeding in describing the procedure that we follow with the Permissions Granter and OSN Data Retriever elements we will give a brief description about how the APIs work in the majority of social networks. In order to be able to access users’ data stored in social networks repositories there is a standard procedure:

1. The first step is to register the application in the social network and retrieve unique application key, consumer key and consumer secret key.

2. Generate an access token whenever a user logs in using the application key. The access token is unique for each user and gives the ability to the application to make API requests on behalf of the user. Before generating the access token, the developer should state which fields his/her application is going to access. When the user accepts to share the information, the access token is being generated.

A major problem in this procedure is the need of the user to be online during the retrieval process for some APIs, such as LinkedIn. Thus, we should implement a mechanism that is able to grant permission to access users’ profiles and provide our framework with the functionality of accessing those fields while the user is not online.

Figure 3.2: Permission Granter flowchart.
3.1 Dataset Collection

3.1.1 Permissions Granter

The Permissions Granter application provides the ability for the researcher to get permissions from the users in order to access his/her social network data. This application supports the generation and storage of user’s access tokens for a later use, in order to simulate user’s online behavior with respect to his/her privacy.

Design and implementation

In figure 3.2 we present the data flow diagram for this application. As we can see, the Permissions Granter first takes as input the consumer key and the consumer secret key. Those keys where given to the developer after the application registration procedure. The next input is the URL of the token request that will be sent to the API of the social network. In this token the developer states the exact permissions that his/her application wants to be granted. An example of this request is the following: ‘https://api.linkedin.com/uas/oauth/requestToken?=r_fullprofile+r_network’.

With this URL the developer requests from LinkedIn the full profile and the community (connections) of the user who access his/her application.

After sending the request the application has to retrieve the response from the social network’s API. This response contains two other tokens, the authorization (auth) token and the authorization secret token. These tokens are responsible for the authentication of the user in the online social network. Upon the receipt of the response, Permission Granter generates a URL, using auth token, that redirects the user to his profile page on his/her social network and asks for their permission to access the profile fields that were stated on the request token. When a user accepts to grant the application with the permissions, the online social network responds with a verification pin code. Permission Granter ask the user to give this verification pin code as an input in order to be able to enable auth and auth secret tokens. Then, it uses access token, access secret token and the verification pin code and executes a verification request to the social network. Now, the application can store the pair of auth and auth secret tokens and use it at a later stage in order to retrieve user’s social network data using Data Retriever application. The pair of auth and auth secret tokens are valid for about 60 days. If the developer wants them to be renewed, he/she should ask the user to use Permission Granter application.

However, we have to state that there are social networks which, according to their API’s terms of use, do not allow the simulation of users online behavior, despite the fact that they do not have mechanism to block it. For this group of social networks researchers are encouraged to follow their terms of use and not to take advantage of the functionality that our framework provides.

3.1.2 Data Retriever

Another important element for the dataset retrieval process is the Data Retriever application. This application executes data retrieval calls to online social networks APIs and manage their responses. The purpose of this application is to access the information of the users who have already granted permissions to the application. Additionally, it supports the enrichment of the collected data either from other resources or from
collected data analysis results.

**Design and implementation**

In figure 3.3 we present the architecture of Data Retriever, which is an online social network crawler. The basic elements in this architecture are the Resources (Online Social Networks, Organization internal platform etc.), the Scheduler, the Data Retrieval Management, the crawler, Tokens and framework data repositories.

- **Resources:** The main resource is the online social network from which we want to retrieve information and we already have permission to do it. For the connection with the OSN we use the corresponding API in order to execute requests to its data repository. The OSN API analyzes the requests and returns the requested data (or an error message) to the system.

- **Scheduler:** The scheduler is an important element in this architecture. Its purpose is to monitor the limitations of social network API and inform the Data Retriever Manager about the errors. For example, some APIs have limitations in the number of requests per hour or day. The scheduler is able to identify the error and inform the data retriever manager to pause the crawling procedure. When the application satisfies the requirements again, the scheduler informs the data retrieval manager and the crawling procedure can be resumed.

- **Data Retrieval Manager:** This manager controls the data retrieval process. It is responsible for retrieving the tokens from the tokens repository and use them in order to achieve the connection with the Social Network. Moreover, it is responsible for managing the permissions and the fields that the crawler should access. Additionally, it is responsible for retrieving the crawled data from the crawler and execute the proper queries to store them in the data repository. Furthermore, it controls the operation of the crawler and monitors the system’s progress. The Data Retrieval Manager is also responsible for synchronizing the retrieval process based on the input that has from the scheduler.

- **Crawler:** The crawler is the element that uses social network API in order to retrieve the requested information. It is controlled by the Data Retrieval Manager and provides a mechanism to retrieve the permitted data, parse them and convert them in a storable format. It then provides the data to the Data Retrieval Manager who controls the storage procedure.

- **Tokens Repository:** The tokens repository holds the tokens that were generated from the Permission Granter. It can be a repository in temporal or permanent memory. It provides the access token and access secret token to the Data Retrieval Manager in order to be able to login to the Social Network API. Data Retrieval manager has read-only permissions to this repository.

- **Framework Repository:** This repository is responsible for holding the data that was collected from the Data retrieval process. After the retrieval of the data from the crawler, the Data Retrieval Manager executes storing requests to the repository. At the end of this procedure the Framework Repository should hold
3.1 Dataset Collection

A Framework for Social Network Analysis

Figure 3.3: The architecture of data retriever element.

the social network data that have been crawled. This repository can be held in temporal or permanent memory (Database, file etc.).

Data retrieval process

The first input in Data Retriever are the consumer and consumer secret keys that were retrieved after the registration of the application in the OSN. This step ensures that the access tokens will be used through this application. Then the application asks for the fields that the Data Retriever should access. These fields can be the same or a subset of the ones that application has access to.

The next step is to retrieve the pair of access and access secret tokens, from the tokens repository, for the user who has granted permissions to the application. Data Retriever manager can receive the tokens either from a text file or from a database, depending on the storage method that was used for the Permission Granter. This pair of tokens gives the ability to our application to create an API client for each user and simulate his online behavior. Thus, Data Retriever can now make requests on behalf of the user and access the permitted social network fields through the API. Specifically, Data Retriever Manager gives the command to the OSN crawler to start the crawling procedure.

The approach for the crawling procedure is first to access the permitted data of current user profile and then iterate in user’s community and access also the permitted data. If there are more than one users, Data Retriever iterates and stores the tokens for each. This procedure can be done while the user is either online or offline. At specific moments, the Data Retrieval Manager pauses the crawling procedure and stores the data in the framework data repository. At the end of the data retrieval process the framework’s data repository will hold all the permitted data retrieved from the online social network.

The choice of storing this data locally sometimes is restricted from the terms of use of some APIs. However it is supported from our framework. The data can be
A Framework for Social Network Analysis

3.1 Dataset Collection

Figure 3.4: A representation of the social graph of a dataset collected using the proposed framework.

replicated to the data repository and used for the next steps of our framework.

OSN Data Enrichment

The information retrieved from social networks can be very valuable for different reasons. If this information is combined with additional information retrieved from user’s environment will provide the researcher with a much more valuable dataset. For example, the social network profile information of a user sometimes is more valuable for his/her manager than a completely stranger. User’s manager can connect the data with user’s professional behavior and gain more value from the combination. Thus, we conclude that our application should support data enrichment functionalities.

Data enrichment can be divided in two categories. The first category is enrichment from data that comes from external sources. These data can be retrieved from surveys, web sources, organizations data, blogs or other social networking services in order to
be linked and enrich user’s profile in the dataset. An example that lies in this category is the use of organization’s data in order to store in the dataset the department that the user works. The second category describes information that can be the result of an analysis of the existing data (online natural language processing API, gender identification etc.). By analyzing that dataset new information can occur about the user. This information can be stored and used as additional information of user’s profile. An example is to analyze if the user is connected to his/her manager and link the answer with user’s profile.

The proposed framework supports data enrichment of both categories, providing the research with the ability of using all the available sources of information. When this information is combined with the already collected dataset, it provides the researcher with better and more accurate analysis capabilities that will help him/her to extract knowledge for his/her target group.

### 3.2 Preliminary data process

The preliminary data process can be done during or after the completion of the dataset collection procedure. Before proceeding into the data analysis phase and extract knowledge from the data, this procedure is needed. In this step our framework takes care about the cleaning of the data, duplicated entries and data alignment. At a later stage network and statistical analysis tools will be used, thus the conversion to the proper format and the production of an import to these tools are necessary.

During the design of our framework we have considered that the majority of statistical analysis tools (SPSS, R, Microsoft Excel, etc.) are able to take as import Comma Separated Value files (.csv). Additionally, except from statistical analysis tools we aim in the support of classifiers that will provide us useful results. Again, the majority of data mining softwares that support classifiers’ training are able to take as import .csv type of files. For that reason we chose to use this format as the primary export for our framework.

The approach that we will follow for the preliminary data process is strongly related with the type of our data storage repository. The procedure could be different if data repository is stored in the temporal memory, text files or database. The preliminary data process in the following cases would be:

- **Temporal Memory**: The data are stored in data structures in the temporal memory. Our framework provides the ability of exporting those data in a .csv file and then use it in the statistical analysis tools.

- **System files**: The data can be stored in different data format files (.xml, .csv, .txt). Our framework can provide an export of the data in a supported format from the statistical analysis tools. Additionally, the preliminary data analysis process can be done in parallel with the data retrieval process. Thus, the framework is able to store the data directly in the proper format.

- **Database**: The data are stored in a Database. To access the data and export it in the desired format we can execute queries on the database based on the database schema. Most of the database engines support the exportation to .csv
files. These files can be used directly as an import to the statistical analysis tools. Additionally, the majority of the database engines support a number of statistical analysis methods, which can provide us descriptive statistics about our dataset in a simple way.

After the preliminary analysis of our raw data we can proceed to Online Social Network data analysis.

### 3.3 OSN Data Analysis

OSN Data Analysis element is important for extracting knowledge from social networks. From this analysis we are able to map and measure the relations and flows that are developed between people, groups and other network nodes. The network nodes can be the users and their profiles, while the connections can be the relations between them. With the OSN data analysis we can proceed to visualized or statistical results. The framework that we proposed provides the ability to researchers with low programming experience to export data from social networks and proceed to their analysis in order to support their assumptions or draw into useful conclusions.

#### 3.3.1 OSN Analysis operations

Existing literature reports that there are different operations that a researcher can follow in order to analyze and draw into conclusions regarding social networks’ data analysis [7]. The proposed framework facilitates the following operations based on how the data is stored and exported. A part of these operations has been used in our case study that is presented in chapter 4.

- **Statistical Analysis:** Milgram [19] with his work supports that extensive analysis of large scale networks structural properties can reveal interesting information. With the statistical analysis in social networks we can examine the connectivity behavior of nodes in the network and identify patterns. For example we can investigate the average size of a community and identify the reasons behind those numbers. Moreover, we can investigate the clustering trends based on descriptive characteristics. As the social graphs in the network evolve in the time, new nodes may be added in the network. Additionally, a node may change its fields or a part of them. The investigation of this behavior can reveal interesting information about user’s real-life, as we can identify his/her new interaction patterns. This analysis is facilitated from our framework as it provides exports to tools that are able to provide such results like SPSS and Weka.

- **Community Detection in Social Networks:** Community detection is an important challenge in the area of social network analysis [7]. This challenge is highly related with clustering problem and tries to identify parts of the network that have similarities on their linkage behavior. However, the dynamic nature of social networks raises more challenges in the area of community detection. The proposed framework supports the researcher of this field by pre-processing and providing the data in a format that can be used in tools which have implemented
3.3 OSN Data Analysis

A Framework for Social Network Analysis

powerful clustering algorithms and classifiers and can provide interesting results for this field.

- **Evolution in Dynamic Social Networks**: Because of the dynamic nature of online social networks where new members join or leave the social graph, many changes take place regarding the structure of the network and the communities. In this area we can analyze which communities change most and identify their patterns. Additionally, we can provide information about the nodes that are involved in these dynamic behaviors. With the presented framework the dataset collection during different time periods is supported. Thus, with the support of powerful analysis tools the framework facilitates the analysis of the evolution in dynamic social networks.

- **Social influence analysis**: Social networks primary action is the interaction between different users, nodes. Such interactions can lead to the influence of the nodes who participate in the interaction in terms of behavior. An example is the influence of having your manager in your community in your posts. Again, this type of analysis is also supported, as the framework provides the ability of retrieving such data, enrich them with external sources and export in a format that can be used from powerful tools.

- **Visualizing Social Networks**: Visualizations provide a natural way of summarizing the information, and present them in a more user-friendly way. Additionally, a visualization of a social network can be the trigger for a more extensive research for a current phenomenon. The proposed framework again provides this capability, by supporting the extraction of the collected data in a format that can be imported in a variety of graph visualization tools.

3.3.2 Supported tools

In this section you can find information about the analysis tools that can be used, among others, for our framework.

**Network analysis tools**

The following network analysis tools were tested and can be used for the proposed framework.

**Gephi**

Gephi [3] is a tool for graph and large scale network analysis. It is an open source software that can be also used for 3D visualizations. This tool provides services for importing, filtering, analyzing and exporting for all types of networks, taking into advantage of multi-core processors which are able to handle large scale computations.

Gephi’s user interface is user-friendly and provides a user with limited programming experience with the ability of using the services that it provides, like filtering the dataset, execute complex network calculations and generate visualized reports. The results of the analysis can be exported as SVG or PDF files.
This tool can be used in our framework by connecting it with our data repository. We only need to define which are the nodes and the edges of the collected social graph with the use of a simple query to the corresponding tables. Additionally, we can import the produced CSV files of the preliminary analysis step.

**VivaGraphJS**

VivaGraphJS is a graph visualization library for Javascript. This library can be used for the representation of the social network graph, where the researcher can visually identify the nodes that need a further research. This library is free of use, and supports rendering with different engines like WebGL and SVG. Additionally, it supports different layout algorithms for the representation of the graph. The functionalities of this library can be used through its Application Programming Interface.

The graph on figure 3.4 was generated using VivaGraphJS, and represents a social network’s graph. Each node is the user and the edges represent the connections between them.

VivaGraphJS gives us the ability to execute queries on existing data repositories. Thus, we can use its functionalities by executing the queries on our repositories in order to import the collected and processed datasets.

**Statistical analysis tools**

The following statistical analysis tools were tested and can be used for the proposed framework.

**IBM SPSS Statistics**

IBM SPSS is a statistical data analysis tool, which provides the user with options for analyzing and modeling data, generating reports with graphical representations. It has several statistical functions for data analysis through a user-friendly GUI. With the help of SPSS all stages of the analytical process are completed under a unified interface covering data analysis dimensions.

This tool can be used also for normality and significance tests. With its user friendly GUI, the user can easily retrieve descriptive statistics about the data and extract knowledge from the results. IBM SPSS provides the ability to the user to export the results in different formats like Excel, CSV and PDF.

After the preliminary analysis step we are able to export the produced data in a format that can be used as an import in IBM SPSS Statistics tool. After importing the files in the tool we are able to use its statistical functionalities on our datasets.

**Weka**

WEKA [13] stands for Waikato Environment for Knowledge Learning and it is a machine learning tool that was developed by the University of Waikato in New Zealand. It is a free software written in Java, with graphical user interfaces. It contains a collection of algorithms and visualization tools for data analysis.

WEKA supports machine learning algorithms for data classification, clustering, regression and visualization. Additionally, it provides simple descriptive statistics for
3.4 Discussion

In this section we have presented a framework that provides the researcher with the ability of efficiently following the procedure for extracting knowledge from a social network. We have presented the different elements that this framework has and the functionality of each element. Moreover, the steps that each element follows in order to achieve its design goal are also presented.

One factor that we have considered during the design of the proposed framework was the efficiency in the procedure that should be followed from the first step until the extraction of knowledge. We have in mind that data analysis tools are being developed for many years so they can provide the researcher with the ability of high level analysis. For this reason, we did not develop a data analysis tool as there are many powerful tools that can accomplish the goals and fulfill researchers requirements. Thus, we have developed a framework that is able to facilitate the researcher in the procedure of data collection, data enrichment, preliminary analysis and provide export in a format that can be easily used by existing analysis tools. With the use of this framework the steps that precede the data analysis part are being automated.

The proposed framework was also designed in a way that can be easily extended. The addition of new elements and steps is highly supported and will not affect the current functionalities. The reason is the use of the extensible Data Retriever element, which handles everything that will be stored in the database repository. Additionally, the Preliminary Analysis element handles everything that is retrieved from the repository, in order to maintain a consistent export.
Chapter 4

Case Study: IBM

To showcase the utility and the potentials of our framework we engaged in a case study performed at IBM Netherlands, where we aim to extract business knowledge from a professional social network using the proposed framework. We perform this case study in cooperation with IBM Netherlands, with whom we agreed on a set of research questions, which we will try to answer using a dataset collected from organization’s employees.

In this case study we aim to analyze users’ behavior inside professional social networks and extract knowledge from this interaction. We hypothesize that the same person handles a profile in more than one social networks, of different types and that their interaction patterns differ between "Social" social networks and "Professional" social networks. Moreover, we know from existing researches that there are factors that have an important role in the determination of users’ social networking behavior in Facebook [31] such as gender and culture.

With this research we aim to investigate if these findings also hold in the case of professional social networks and identify also other factors that influence users’ behavior. Additionally, by analyzing the users’ communities and their interactions inside a professional social network we aim to extract business value, which will be useful for companies and organizations, such as IBM.

4.1 Experimental Methodology

In this section we describe the methodology that we follow for this case study. We present the research questions that we aim to answer and the approach that we followed.

4.1.1 Research Questions

For this case study we have agreed with IBM a set of research questions. We then use the proposed framework to collect and analyze a dataset retrieved from LinkedIn, enriched with information from internal IBM organization, in order to answer our research questions.
4.1 Experimental Methodology  

Research Question 1: Do employees use different social networks for different purposes?

Existing literature reports that there are different types of social networks [12]. Additionally, Skeels and Grudin show in their research that there are differences in the behavior of the users in Facebook and LinkedIn, the most popular social and professional social networks [27]. Based on this findings, with research question 1 we aim to support our hypothesis that the same person handles a profile in social networks of different types, for a different purpose. In specific, we aim to show that the same person is a user of a "Social" social network and a "Professional" social network in parallel. To answer this generic question we perform a survey inside IBM, aiming to investigate if this hypothesis holds between IBM employees. The results of this survey are used as an input in our framework, and will help us to define and investigate the rest of the research questions. Research question 1 has been divided into the following sub-questions:

**RQ1.1: Are there differences in his/her interaction with them?**

We hypothesize that the same person handles a different profile in different platforms for "social" social network and a different one for "professional" social network. This might be because, of the nature of the corresponding platforms and the purpose that they have been designed and developed. Thus it is easier for the user to use the corresponding platform for the purpose that he wants. With this sub-question we aim to investigate if there is a difference in the interactions between the same user and his "Social" and "Professional" social network.

**RQ1.2: Is there a difference in the composition of his/her communities?**

Additionally, we hypothesize that the same person, who handles a profile in a "social" social network platform and another one in a "professional" social network platform, tend to construct his communities with different criteria. We test the hypothesis that the type of the network is an important factor in the decision of adding a new member in the community and try to identify the reasons behind the addition of a new member in user’s community.

Research Question 2: How can we extract business knowledge from professional social networks?

By their definition professional-career oriented social networks aim in the representation of user’s professional background. These networks encourage the user to upload content related with his/her curriculum vitae such as skills, education, experience, current career status. With this question we aim to investigate if this information can be valuable for the organizations that employee-user work for and if we can extract business value from information published by the users.

To answer this generic question we divide it into the following research questions:

**RQ2.1: Do employees reveal internal organization information in their professional social network?**

A person in his/her CV reveals as many details as possible about his experience, past
and current positions that he/she is involved with. A user’s profile in a professional social network is directly related with his/her CV [27]. With this sub-question we aim to analyze if the way that user’s promote their self in a professional social network reveals internal organization information.

**RQ2.2: How can an organization’s key employees be identified from professional social network profiles?**

As stated by Skeels and Grudin [27], user’s LinkedIn page is relatively static, apart from his/her community. Additionally, based on previous work which suggest that social networks can reinforce meaningful relationships within the working environment [28], we aim to gain value from the analysis of the static but also dynamic part of the employee’s profile.

**RQ2.3: How can we infer information about changes in an organization’s structure from its employees’ professional social network?**

Using the information published from the user in his/her profile, enriched with information retrieved from the organization, we investigate if it is possible to identify changes in the internal structure of the organization. We aim to identify changes such as departments growing and company’s hiring periods.

The answer to these questions can be found thanks to the support of our framework. After collecting the dataset, social network information will be analyzed in order to identify factors that are related to the construction of user’s community inside his professional social network. From the analysis we aim to answer the following research questions:

**Research Question 3: How can we identify distinguishing properties of different categories of employees from professional social network profiles?**

With this research question we concentrate in analyzing the dynamically changed field of user’s community in relation to static properties such as user’s country, industry, position in the organization that he/she works, organization’s size and type, etc. Our goal is to identify properties of different categories of employees using data retrieved from professional social network.

In order to answer this research question we have divided it into the following two questions:

**RQ3.1: How does a department and its operations influence the community of the user?**

Given the size and extend of IBM organization, we test if the operational organization affects the community of the user. By raising this sub-question for a professional social network, we aim in identifying different factors that are related with user’s working environment and influence his/her community structure. Specifically, we investigate in what extend do employees tend to establish connections with users related with their working environment like their colleagues, clients or managers. After having our conclusions we try to reason about the facts that we identify.
4.2 Experimental Setting

Case Study: IBM

**RQ3.2: Which are the factors that influence the size of the community of the user?**

From literature we know that there are factors, such as gender and culture, that influence the behavior of the user inside a social network [31]. With this sub-question we want to investigate if these factors also hold in the case of professional social networks. Specifically, we test if user’s nationality, user’s industry and properties of the organization that user works for like organization’s size and industry influence the size of his/her community in a professional social network.

4.1.2 Approach

Our approach for answering these research questions starts with the investigation of RQ1 and the definition of "business" value. For that reason, at the beginning of the project we decide to perform a set of qualitative interviews. During this research we investigate the views, understandings and ideas of a selected group of IBM employees in order to answer our research question. After supporting our hypothesis and defining the term of "business" value, we continue our research project with the survey.

With the survey we intend to validate the results of our interviews with a larger group of people and build a part of our ground truth that will later help us in the general validation of our case study. Additionally, the results of this research have helped us to choose the proper professional social network that will later be used in our framework for the dataset collection and analysis. Before releasing our survey, we have developed a website that explains to the visitors the project’s purpose, and a simple web application that provides them with information about their professional social network. The purpose of this website is to attract them to participate in our research and later give us permission to access their professional social network profile through the proposed framework.

The step that follows is to define the experiments and the data analysis that we are going to complete with the use of the proposed framework, having as goal to answer research questions 2 and 3. We proceed to the dataset collection, data analysis and knowledge extraction using our framework and existing tools.

4.2 Experimental Setting

In this section we describe the data sources that have been used in our case study. Additionally, we describe our approach of collecting the dataset and the data model using the proposed framework. Moreover, we will give information about the approach that we have used for validating our results.

Figure 4.1 shows how our framework was used in this case study. More details about each element will be presented in this section. As we can see the users that have been used for the case are the IBMers and their profiles in LinkedIn online social network. The dataset has been enriched from IBM bluepages framework, survey answers and information retrieved from finance department. Permissions granter has been used for the collection of the permissions and data retriever to access the profiles and link them with external resources. Moreover, a preliminary analysis was done in order prepare the data for analysis (replace non-unicode characters, remove duplication, remove non-completed fields etc.).
4.2.1 Data sources

The main data sources that we have used is data retrieved from IBM employees (survey), LinkedIn professional social network and Bluepages internal network.

IBM employees

IBM employees were our target group and the basis for this case study. Firstly, we get insight from a selected group of IBMers via interviews. Then, in order to verify and generalize our findings we target a larger group of IBMers and ask them to fill a questionnaire (Appendix I). Additionally, IBM employees who have answered the survey they are asked to provide us permissions to access their LinkedIn profiles through the proposed framework. Moreover, we have used data from the internal organization related to them in order to enrich the collected dataset.

LinkedIn

LinkedIn is an online social network which focuses on professional information. It encourage users to build and update their digital CV, while they also maintain their professional community. LinkedIn is mainly used for professional networking and it belongs in the category of professional-oriented social networks. From June of 2012 LinkedIn reports more than 175 million registered users in more than 200 countries. This social network is available in many different languages like Dutch, English, French, German, Italian, Spanish etc..
The main goal of LinkedIn is to provide career opportunities for professionals from all over the world. This network does not only target persons who are searching for jobs but also companies who are searching for employees. Moreover, it provides the user with the ability of uploading information about his educational background, professional career and interests. The user can also create a community by sending/accepting connection requests to/from his friends, colleagues, clients etc. in order to keep in touch with them and follow their updates.

Douglas [10], referring to his personal use of LinkedIn, identifies that a LinkedIn user could be a recipient of an inquiry (e.g., from a job hunter), an intermediary (e.g., passing messages along my network on IS behalf of friends), or a job hunter (e.g., seeking employment).

Hempel [14], in a try of emphasizing the useful and important role of LinkedIn in someone’s recruitment and professional career, he mentioned that "If you don’t have a profile on LinkedIn, you’re nowhere." Moreover, a lot of tools have been implemented in order to help graduate students and professionals to find new career opportunities easier. Characteristic example of this type of tools are the LinkedIn search platform which has been improved in 2008 and 2009 [17] [18] and the Career Explorer (Watters, 2010) [33]. On the other hand, Schuen [26] considers LinkedIn similarly valuable to salespeople who are seeking contacts as potential clients.

Bluepages

Bluepages is an IBM’s internal framework which provides IBMers with the ability of finding the right person to contact in order to solve a problem that they are struggling with. Every IBMer has a profile in bluepages, which states in which building, department office he/she works, phone number, manager’s name and specialities.

Using this framework an IBMer can get in contact with professionals that are able to help him with what he works on. For example if an IBMer wants to work with RFID, he/she can search for an IBMer who is professional in that technology and get his valuable help. Moreover, from Bluepages, an IBMer can get information about different operations and identify persons that work in the same area as he/she and start a new internal cooperation.

Information from Finance Department

Another interesting source that has been used for the data enrichment it was information retrieved from the Finance Department. From this department we retrieved information about the department’s code, name, building and operations and we linked them with the already collected data.

4.2.2 Dataset

Web site

In order to attract people to participate we have developed a website for our project. In this website we host interesting information about the purpose of this research and describe the gains that some could have by participating. Additionally, we developed
a web application, which provides interesting insights using data from user's "Professional" social network profile.

**Recruitment**

Our basic method for recruitment was through personal contact with IBMers. With this method we try to spread our quantitative survey inside the company. First, we tried to contact employees who were working inside IBM Netherlands, via email. We also contacted managers and ask them to spread it in their departments. Then, we share the link of our website inside IBM groups in LinkedIn "Professional" social network. Note that our survey was hosted in IBM’s internal network, thus only users working in IBM could visit it. These two methods yield 134 responses.

**Procedure**

This study ran from January until March of 2013. Using the methods described above, participants were invited to fill out an online survey and, using the proposed framework, give read-only permissions to our application for their "professional" social network profile. We then gain access to their personal information that they share in their LinkedIn profile and the 1st degree connections that they maintain. The questionnaire we used was developed in English.

**Dataset Information**

In our dataset we have the following categories of users: (Figure 4.2)

- **Core IBMers:** Survey participants. We have access to their full LinkedIn profile and bluepages information
4.2 Experimental Setting

Case Study: IBM

Figure 4.3: Case study: Data model. The fields that LinkedIn provides but are not filled by our participants are marked with red colour.

- **IBMers**: IBM employees who belong in Core IBMers communities. We have access to basic LinkedIn profile and bluepages information.

- **Basic Users**: Non IBM employees who belong in Core IBMers communities. We have access to their basic LinkedIn profile.

**Data model**

In this section we describe the data model of our dataset (figure 4.3). The first type of user that we can find in our dataset is the **Basic User**. This user is a non IBM employee who belongs in at least one Core IBMer’s community. For this user we can access his/her basic profile information fields:

- **Headline**: User’s personal description of his/her current position.

- **Country**: User’s country code.

- **Industry**: The industry that the user has indicated that he/she belongs to. A LinkedIn user has to choose from a list of industries that LinkedIn provides.

- **Community Size**: The number of connections a user has. If this number is smaller than 500 then we have the actual number. If it is not, the value of this field is 500.

- **Positions**: Information about current positions that a user reports in LinkedIn.
Case Study: IBM

4.2 Experimental Setting

- **Job Role:** User’s personal description about the current position
- **Start date:** The date that this position has started
- **Company:** Information about the company that user works for this position
  - **Industry:** The industry in which the company operates
  - **Size:** The number of the employees in the company. LinkedIn provides a list of choice for this field
  - **Name:** The name of the company

The second type of user is the **IBMer.** IBMer is an IBM employee who belongs in at least one Core IBMer’s community. For this user we can access the information that we can access for Basic User, plus information retrieved from IBM internal organization.

- **Bluepages profile:** The profile that every IBM maintains in this framework

  - **Employee number:** His/her employee number inside IBM. A unique number for each employee.
  - **Is_manager:** Flag that indicates if the current IBMer manages at least one other employee in IBM.
  - **Department:** The number of the department that he/she belongs.
  - **Job role:** A short description about his/her job role inside IBM.
  - **Manager:** The employee code of his/her manager.

- **Department:** Information about the department that the IBMer works for.

  - **Code:** The code of the department.
  - **Operation:** The operation of the department.

The third type of user that can be found in our dataset is the **Core IBMer.** A Core IBMer is the most completed user that we have, as he/she is also an IBMer and a Basic User. According to his/her full profile, we can retrieve information about his/her email, date of birth and interests. Moreover, we can access Core IBMer’s languages, skills and information about his/her education. Educational information provide us information about the field of study, the degree, start and end dates and school name. Additionally, is the only type of user in our dataset that we can access information about his/her connections.
4.2 Experimental Setting

Demographic Information

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>73.37</td>
<td>101</td>
</tr>
<tr>
<td>Work in country of origin</td>
<td>85.82</td>
<td>115</td>
</tr>
<tr>
<td><strong>Countries of origin</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>51.49</td>
<td>69</td>
</tr>
<tr>
<td>Belgium</td>
<td>12.6</td>
<td>17</td>
</tr>
<tr>
<td>USA</td>
<td>11.9</td>
<td>16</td>
</tr>
<tr>
<td>Denmark</td>
<td>7.46</td>
<td>10</td>
</tr>
<tr>
<td><strong>Educational</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School Diploma</td>
<td>8.96</td>
<td>12</td>
</tr>
<tr>
<td>BSc Degree</td>
<td>29.10</td>
<td>39</td>
</tr>
<tr>
<td>MSc Degree</td>
<td>42.54</td>
<td>57</td>
</tr>
<tr>
<td>Ph.D Degree</td>
<td>6.72</td>
<td>9</td>
</tr>
<tr>
<td>Other Diplomas</td>
<td>12.69</td>
<td>17</td>
</tr>
</tbody>
</table>

A part of our survey is organized to collect demographic information as presented in Table 1. From demographic data we can see that the 73.37% of our sample is male and 24.63% female. Additionally, from educational data we can see that 42.54% of the participants hold a MSc Degree, 29.10% a BSc Degree, 6.72% a Ph.D degree, 8.96% a High School diploma and 12.69% other diplomas. Moreover, 85.82% works in the country of origin, having the rest 14.18% working abroad. In Table 1 we can also see the countries where the participants work, having Netherlands ahead with 69 participants, which is the 51.49% of the total number.

4.2.3 Tools and methods used

Limesurvey has been used for designing, developing and releasing our survey. Additionally, through this tool we have collected and store the survey answers.

Limesurvey

LimeSurvey is a free and open source on-line survey application written in PHP based on a MySQL, PostgreSQL or MSSQL database, distributed under the GNU General Public License. As a web server-based software it enables users to develop and publish on-line surveys, and collect the responses. Limesurvey provides also the ability of descriptive data analysis and visualization. It has different options of export like Excel, CSV, SPSS, R, Text.

Other tools

For the statistical analysis we have used IBM SPSS Statistics tool. For the classifier training, social graph analysis and social graph visualization, WEKA, Gephi and Vi-vaGraph have been used. The above mentioned tools and library are being presented in section 3.3.2.
4.3 Findings - Results

4.3.1 Research Question 1: Do employees use different social networks for different purposes?

RQ1.1: Are there differences in his/her interaction with them?

Interviews

From the results of our qualitative research we were able to investigate the first part of our research question regarding the behavior of a user inside online social networks. We found out that our small sample of participants holds a profile in a "Social" social network and a different one in a "Professional" social network at the same time. Additionally, from our discussions it was clear that they interact differently in these two profiles by sharing different kind of information and having differences in the frequency of visits and the amount of time they spent in each network.

Moreover, all of them believe that their community in their "Social" social network is different than their community in their "Professional" social network as they have different criteria when they are accepting or sending a "connect" request. All of them believe that there are other factors that may define a different behavior inside a professional social network, like the department that an employee works for, culture, the industry type and size.

Specifically, employees of bigger companies or more related to the computer and network industry, may have differences in the structure of their network than users with other characteristics. Additionally, employees who work at the same company but in different departments may have differences in their community because of different partners and colleagues. The culture and the personality characteristics of every person could also affect his social network behavior.

Furthermore, all of them believe that social networks can be used in order to extract business value and they shared their ideas with us. They strongly believe that the communities in a "Professional" social network hide much more valuable information from what the business world thinks. We also believe that the community analysis of "Professional" social networks is a strong tool that can provide the companies with many advantages.

After the completion of the qualitative research we conclude that the following points can be investigated with surveys and data analysis:

1. Same person handles a profile for his "Social" social network and a different one for his "Professional" social network

2. Interaction patterns are different between "Social" social networks and "Professional" social networks

3. There are factors that have important role in the determination of social networking behavior

4. User communities interaction analysis can lead us in extraction of business value
4.3 Findings - Results

Case Study: IBM

Figure 4.4: Left: Which social network do you use for your social connections? Right: Which social network do you use for your professional connections?

Business Value

According to IBM Netherlands managers that have been interviewed, this term has direct relation to the organization and its business clients. In particular, they all have in mind a term that, among others, describes the need of information about the internal organization. An example of such information is the department and/or the job role of an employee, exactly as it is defined by the internal procedures of the organization. Additionally, business value also describes information about an organization’s structure. Such information could be the hierarchy of the organization, or a part of it, employee or department-wised. Moreover, organization’s milestones or important chronological periods are also enclosed in the term of business value. The identification of such events can help the organization to identify its weak and strong points and try to improve them.

Every organization is connected with its business clients in some degree. Sometimes, more than one department can be connected with the same client for different purposes; some for business purposes and some for other purposes like family relations, friendship, ex-colleagues, etc. The identification of the strength of the connection between different parts of an organization and the business client is also enclosed in the term of business value. In this way, information about the connection between an organization and the client, that the organization wasn’t aware of, can be revealed.

Survey: General statistics

At the end of the interviewing procedure, we were able to get an insight for our research question. The next step was to support our results based on a larger sample with the quantitative research. From the results of our survey we are able to drive into conclusions regarding the behavior of the users in online social networks.

The research through this survey give us a general understanding of how do IBM employees interact in social networks. Based on our findings from the qualitative research we develop a quantitative survey and release it inside IBM.
The sample of our quantitative research is 134 participants. They are all employees of IBM, working in different departments, buildings and countries. The results are presented using graph charts and tables.

According to social networking questions section, we can see that 70.9% of the participants mostly use Facebook for their social connections (as "Social" Social Network), having LinkedIn with 14.18% and Twitter with 8.21% come second and third respectively (figure 4.4). From figure 4.4 we also note that the mostly used social network for professional connections is LinkedIn with a percentage of 90.3%.

For the question of "How old is your "Social" social network profile?" the majority of the participants, 32.84%, answered between three and five years, followed by more than five years with 31.34%. The results are similar with the corresponding question...
for professional social network, with the answer between three and five years having 28.36% and the answer more than five years followed with a percentage of 27.61% (figure 4.5).

Moreover, according to the participants’ answers, we can see that the majority of them (36.57%) visit their "Social" social network more than once daily, and at the same time they visit their "Professional" social network more than once a week with a percentage of 32.84% (figure 4.6). From what we can see, people tend to visit their "Social" social networks more frequent from their "Professional" social networks. Most people visit their "Social" social network daily, while most of them have more than one visits during the day. A part of our sample visits their "Social" social network on weekly basis while less visit it in monthly basis. On the other hand, we have the results of "Professional" social networks. From these results we can see that the majority of the users visit their "Professional" social network more than once per week but less than once per day. A large part visit their networks daily, while the most of them visit it once daily.

Finally, for the question about the time that they spend during the interaction with their social network they choose the answers of between 1 and 2 hours and between 2 and 5 hours per week, with percentages of 32.84% and 26.12% in the social social network, and 47.01% and 33.58% for their professional social network (figure 4.7).

In addition, we can get interesting insights by not just analyzing the frequency of social networks visits but also by analyzing the time that users spend during their visits. In figure 4.7 we can see that the majority of users in both social networks spent less than one hour per week in their interaction with social networks. However, we can see that the number of the majority is different, having visits that last less than 1 hour be 31% less in "Social" social networks than the ones in "Professional" social networks. Moreover, we can see that users who spent between 2 and 5 hours in their "Social" social network are more than the users who spent the same amount of time in their "Professional" social networks.
Discussion

From the results we can derive that a person handles a different profile in different types of social networks. From figure 4.4 we can see that people in our sample prefer Facebook for their "Social" social network interactions and LinkedIn for their "Professional" social network. This fact also guides us to choose LinkedIn "Professional" social network for the network that will be used in our research as it seems to be the most popular in our sample.

Furthermore, from the results of figure 4.5 we can derive that in the majority of people in our sample the age of their profiles in both types of social networks is more than three years old. However, we can also derive that during last three years more people tend to create a profile in a "Professional" social network, while less tend to create a profile in a "Social" social network. From this fact we can conclude that a part of our sample created a "Social" social network profile before the creation of their "Professional" social network profile.

Additionally, by comparing figure’s 4.6 results of "Social" social networks with "Professional" social networks we can see that the same users visit more frequently their "Social" than their "Professional" social network. This could be because of the difference in communities structure. From what we described above, the communities in "Social" social networks are build based mostly on friendship. On the other hand, "Professional" social networks are build based on professional reasons. By combining all information together, we can derive that people of our sample tend to interact with their friends more frequent than with their professional partners and clients.

The information on figure 4.7 can lead us to the general conclusion that the average user of our sample spends less time during his visits in "Professional" social networks, having the majority spending about 1 hour per week. In addition, the minority which is spending more than 10 hours per week is bigger in "Social" social networks than the one in "Professional" social networks. Again, this could be reasoned from the difference in the communities structure of these types of networks.

Research Question 1.2: Is there a difference in the composition of his/her communities?

Moreover, we can see that for a question according to the reasons to connect with someone in their "Social" social network 89.55% answered friendship, 36.57% colleague, 20.90% Business (e.g Client, Potential Client), and 20.15% choose other. For the same question, according to their professional social network, 91.04% choose Business (e.x Client, Potential Client), 77.61% Colleague, 41.79% Friendship and 11.94% other reason (figure 4.8). From this figure we can see the reasons why people of our sample tend to add a new member in their community. In the case of "Social" social networks we see that most of the users connect with another user because they have a friendship relation. A smaller number creates a connection because they are colleagues and a smaller one because of business relation or other reasons like family relations. In the case of the "Professional" social networks things are different. The majority(122 users) is connected with other users because the are doing, or going to
4.3 Findings - Results

Discussion

From these results we can drive to the conclusion that "Social" social network communities are different than "Professional" social network as users tend to construct them based on different reasoning. This fact can be reasoned in the description of these categories of social networks. Specifically, "social" social networks are described as personal-oriented and advice the users to connect with others who share a personal relationship with them. On the other hand, "professional" social networks are described as career-oriented and advice the users to maintain their curriculum vitae and build connections with a more professional career-related social graph.

Survey: Data analysis

Having the results of the previous section we decided that we could proceed to deeper research. In this part of our research we aim to analyze the different combinations of statistics in order to proceed into conclusions about the behavior of the user in a professional social network and the factors that influence it. Specifically, we analyze the relation between the different factors that we examine through our survey.

Firstly, we investigate the relation of the different reasons that users have to add a new member to their professional social network’s community with the following factors: age of the profile, frequency of visits, duration of visits and educational level. The main reason that we investigate is the friendship. By definition, a professional social network encourage users to build their online professional-oriented community. Thus, we want to investigate the group of users who also add friends in this community. We found out that a percentage of 41.18% adds another user in the community because...
of that reason. Here we proceed into a deeper analysis to identify the different types of users who contribute to this percentage.

We investigate the relation of friendship reason with profile age. As we derive, a percentage of 75% of user’s who decide to add a new member to their community because of friendship are user’s who have a profile for more than two years. Additionally, from this 75% the 62,50% are users who have a profile for more than three years, while the majority of them has a profile for more than five years.

Moreover, according to the users who add colleagues in their community we can identify an interesting result. In specific, a percentage of about 86% of users with profiles with age smaller than six months tend to add their colleagues in their communities, while the percentage of user’s that belong to this group for profiles older than six months drops to 75%.

Another interesting category that we investigate is the frequency that users who add friends in their professional social network community visit their profiles. From
the statistical analysis we can see that 71.42% of this group of users visit their profiles more than once a week, having the majority (48.21%) of them visit it once or more than once daily. Furthermore, 25% from this 48.21% visit their profiles in a frequency higher than once daily.

For the case of the relation with the duration of visits we can see that the results are similar with the general case, thus we cannot derive into conclusions with the sample that we use. The results are similar also with the education status, having the majority of this group of users holding a BSc (26.79%) or an MSc (48.21%) degree.

Additionally, we investigate the relation between the age of the profile and the frequency and duration of visits. According to the duration, it is a general finding that users do not spend more than two hours per week in their professional social network. As we can see the majority of users having a profile for less than one month is visiting its social network for once or more than once a week, while the duration of visits follows the general finding. Furthermore, if we investigate users who have a profile for more than a month but less than a year, we can drive into the conclusion that they are the most frequent visitors of professional social network. Specifically, 61.54% of users who have profile for between one and six months visit their network once or more than once daily, while the same applies for the 62.50% of users who have a profile with age between six months and one year.

Furthermore, the investigation of the relation between educational level and the age of the profile is also interesting. By analyzing the answers we can see that the majority of the users have a profile older than two years. However, in the case of users who hold a University degree (BSc, MSc, Phd) who can see that the majority has a profile for more than three years. If we examine these categories, we can derive that from users who now hold a MSc degree, a percentage of more than 60% is in these categories. Namely, a percentage of 31.03% has a profile with age of three to five years old while a percentage of 29.31% has a profile older than five years. Moreover, the case of PhD degree holders is similar, having 55.56% and 33.33% belonging in these categories, respectively.

Discussion

From the deeper analysis in survey data we are able to provide more insights about users’ interaction with social networks. Firstly, from these results we can derive that user’s with older profiles, tend to add their friends in their professional social network communities. Specifically, we can conclude that the older the profile is, the less colleagues get added to the community. This conclusion might be due to the fact that employees who work in a company for long time, they have already added their colleagues during the early age of their profile. Thus, while their profile grows regarding its age, their community grows with other users than their colleagues (as they have been already added).

Additionally, as we can see, there is a relation of frequency of visits with the fact of having friends in the professional social network community. Specifically, we can derive that users, of our sample, who add friends in their professional social network’s community tend to have frequent visits to their profiles.

Moreover, regarding the relation between the age of the profile and the frequency of visits we can conclude that users in our sample that have a profile with age less than
1 year tend to visit their professional social network profile more frequent than others. For the relation between the education level and the age of the profile we conclude that highly educated users in our sample build a profile in a professional social network earlier than the other groups.

4.3.2 Research Question 2: How can we extract business knowledge from professional social networks?

In order to answer the high level research question, we first divide the problem into smaller pieces by defining 3 lower level research questions. In the following sections we explain the approach that we follow for each question and present our results.

RQ 2.1: Do employees reveal internal organization information in their professional social network profile?

In our dataset we have matched LinkedIn profiles that belong to IBMers with their profiles from internal IBM network, bluepages. To answer this question we rely on the most public information available in LinkedIn, which is the headline. The reason behind this choice is the fact that LinkedIn headline is available for 9100 IBMers, and not just 130 who were the direct participants in our project. Thus, the sample that we want to investigate grows exponentially.

We compare LinkedIn headline’s text with job responsibilities text field of bluepages framework. With this comparison we want to investigate if IBM employees reveal their internal job responsibilities into a public field in a professional social network.

Data: The data that we use to answer this question are the "Headline" field from LinkedIn and "Job description field" from Bluepages, for 1000 randomly selected IBMers.

Approach: To achieve our goal we retrieve from our dataset the "headline" field from LinkedIn and the "job responsibility" field from bluepages, for the same user. Because of the need of maximum accuracy in the semantic analysis and the meaning of the text used, the analysis was done manually. The procedure that we use is the following:

1. From IBMers in the dataset we randomly choose 1000
2. Export LinkedIn "Headline" and Bluepages "Job responsibility" for each and every user
3. Compare LinkedIn "Headline" and Bluepages "Job responsibility" and categorize them in: 0. Same, 1. Almost Same, 2. Different. With the term "Same" we mark the users who have the same LinkedIn headline and bluepages job responsibilities. With "Almost same" we mark the user’s who have a similar description in the corresponding fields, but with few differences. With "Different" we mark the user’s who have completely different descriptions.
4. Extract findings and investigate the reasons behind the results
4.3 Findings - Results

Case Study: IBM

Figure 4.9: Results of the comparison between LinkedIn headline and Bluepages job responsibilities.

The results are presented in figure 4.9. As we can see from the pie chart more than 75% of the total sample have the same information in LinkedIn and Bluepages fields. Furthermore, we can see that a percentage of about 13% has a LinkedIn headline that is almost same as their Bluepages job responsibilities field. The users who belong in this category can be divided into two subcategories: 1. The ones who use the same role but they describe it with different terms and 2. The ones who publish their general position in LinkedIn and their specific position and/or the client they work with in Bluepages.

In the first category we identify users who rephrase their description of one network and use it to the other. That means that they provide the same information in both networks. An example of this category is a user who has "CICS Product Line Manager at IBM" in LinkedIn, while his internal job role is "CICS TS Product Manager".

In the second category we can find users who provide similar information about what they do in both networks. However, the users in this category try to generalize in LinkedIn the job responsibilities that they have in Bluepages. A reason might be because they want the public to understand what they are doing, and despite the fact that they work for a client, they actually work for IBM. An example of a user in this category is the one with "Service Sales at IBM" in LinkedIn and "Service Sales Public/ING" in Bluepages.

A percentage of 23.5% belongs to the first category, while the majority, 76.5%, belongs to the second.

In the third category we identify cases where the fields expose a completely different information, having 12% of the users belonging in this category. Here, we also performed a deeper analysis in order to understand the reasons of the difference. We found out that the main subcategories are: 1. users who publish their general job role in LinkedIn, while they have their exact specialties in Bluepages. An example of this category is a user who has "Associate Project Manager IBM" in LinkedIn and "Mobility phones, data cards, blackberries Services" in Bluepages. 2. users who publish a personal description in LinkedIn and their actual position and responsibilities in Bluepages. An example is a user who describes his self in LinkedIn as "Passionate
provider of Competitive Advantage for IBM” and as "Committed Talent Mentor? European Sales Execution Leader” in bluepages.

Discussion

From the results of this step we can see that the majority of IBMers reveal their internal organization job role in a public field in LinkedIn, while a minority does not and they intend to give a more general or personal description of their current positions. The latter category seems to modify their job description in order to make it more appealing and understandable for the visitor of their profiles. Here we have to note that there was no policy from IBM according to the information that IBMers can publish about their job role, at the time of dataset collection.

RQ 2.2: How can an organization’s key employees be identified from professional social network profiles?

As we can understand from the previous results, IBM employees do reveal information about their responsibilities in LinkedIn. With this subquestion we aim in investigating if we can infer key organizations’ employees from IBM employees profiles in LinkedIn. The first assumption that we make before proceeding to the analysis is that managers are key employees. In specific, a manager in IBM is a person who manages at least one IBM employee. Thus, for the organization, an employee that is a manager has many responsibilities and may reflect the way that the organization works. Moreover, the distinguish between employees who are managers and those who are not will also be valuable in the extracting details about the structure of the organization.

To answer this subquestion we try to identify patterns in the terminology that managers and non-managers use in their job description in their professional social network profile.

Data: The data that we use to answer this question is the headline retrieved from IBMers LinkedIn profile and information about who is manager and not from Bluepages.

Approach: The general approach that we follow was to analyze the terms used by people identified as manager and non-managers in Bluepages.

By analyzing our dataset we found out that a number of 2233 IBMers’ in our dataset are reported as managers from bluepages, while a number of 6873 are not. The functionality of "Manager” in bluepages states that the IBMer who holds it manage at least one other employee.

Single terms analysis

The next step is to analyze their headlines and try to identify the terms that each category is using.

Table 4.3 presents the top 10 terms that are being used by managers and non-managers. As we can see, despite the fact that these two categories are clearly distinguished based on their organizational level, they have terms in common. The interesting part is that the overlapped terms that we identify are terms which clearly state
4.3 Findings - Results  

Case Study: IBM

Figure 4.10: Coverage of managers and non-managers for the most popular terms used by managers.

Table 4.3: This table presents the top 10 terms that are being used by managers and non-managers in their LinkedIn headline. The common terms between the two categories are highlighted. The categorization between managers and non-managers was done using information retrieve from Bluepages.
that the person who uses them is a manager. Thus we need to investigate the reasons behind why non-managers are using terms that do not reflect their organizational level. For that reason we proceed into analyzing the terms that are used by managers and are being overlapped from terms that are used by non-managers.

In figure 4.10 we can see the overlap of the most popular terms used by managers. As we can derive, occurrences of terms like "leader", "executive" and "president" are mostly covered by managers. However, other terms that clearly state a manager’s functional level, like "manager", "management" and "senior" are mostly covered by non-managers. Thus we decided to proceed into a further investigation.

We did an investigation inside IBM to identify the reason of having the terms of "manager" and "senior" higher in the list of most popular terms between non-managers. As we have stated before, bluepages internal framework reports as manager someone who manages at least one other employee. However, the terms of "manager" and "senior" are more general terms that can also be used for different meaning. From our discussions with employees who are not managers and use a term which states that they are, we found out that terms similar to "manager" ("director", "partner", "leader", "executive") also state the fact that they are responsible for something, like a program or a partnership. Thus, they are using a text which includes these terms but for a different role than the one defined by bluepages. Additionally, the term of senior is popular between non-managers because it also indicates the level of knowledge and experience that they have in their specific field.

**Bi-grams analysis**

The next step in our investigation is to analyze the headlines of these two categories, using bi-grams, and identify which pair of terms each category is using.

Figure 4.11 presents the 50 most popular pair of terms that can be found in managers’ headlines. As we can see the pairs of "vice president" and "sales manager" are
4.3 Findings - Results

Case Study: IBM

The pair of "program director" follows having "associate partner" 4th. By observing the general results of figure 4.11 we can see that the distribution of bi-grams in the headlines is different than the distribution of terms, where term of "manager" was by far the most popular.

From figure 4.12 we can see that the most popular pair of terms between non-managers are "it specialist" and "it architect". Additionally, by observing figure 4.12 we can see that the majority of bi-grams do not overlap with the bi-grams that are being used by managers. Namely, managers and non-managers have very different distribution as the 84% of top50 terms are different.

Discussion

These results show that managers and non-managers have different job description style and vocabulary. However, we can see that the two categories have some very relevant commonalities. For example the terms of "project manager" and "managing consultant". In order to investigate the reasons behind these results we have some discussions with IBMers. The results of our discussions were similar with the investigation of single terms: a manager is not just the person who manages at least one other but also the person who is responsible for something, like a program. Thus, there are many IBMers who are not managers but are using bigramms in common. For example a "project manager" (or "program manager", "marketing manager", "sales manager") is possible to be a manager according to bluepages definition but it can also be an employee who is responsible for a project but does not manage any other employee.

From our analysis we can conclude that there is a distinction in the vocabulary that managers and non-managers use. This fact provides the ability of the identification of key employees by analyzing the terms that they use to describe their job role. Moreover, we have identified a list of terms that is being used by both categories and the reasons behind this fact.
Do employees in different operations show different compositions of companies’ industry types in their community?

Company’s industry is a field that a company completes when it creates a LinkedIn account. In order to complete this field, LinkedIn provides a list of pre-defined choices. Thus, each company has an industry, and each employee’s position is linked with a company. Furthermore, each department in IBM is linked with an operation and each IBM employee is linked with a department.

With this sub-question we aim in the investigation of the community of an employee, and try to identify differences based on the operation of the department that the employee works for. The results can lead us to identify patterns that IBM employees have in their communities in relation to their department’s operations.

Data: The data that we use to answer this question was retrieved from our Core IBMers who work in the Netherlands. The reason is that for them we also have the department that they work for and the operation of this department.

Approach: The approach that we follow in order to answer this question is the following:

1. For each operation we calculate the "connections to industry type" coverage (figure 4.13)
2. For each industry we calculate the "connections to operations" coverage (figure 4.14)
3. Extract findings and investigate the reasons behind the results

In figure 4.13 we can see the coverage of each operation according to the industries. As we can see the majority of the industries are mostly connected with ConS (Consulting) operation. The only exception is the Defense & Space industry where 45.7% of its connection are with public sector operation. Except from Defense & Space industry, Public Sector operation also covers 33% of Military industry. Experience suggests that Defense & Space and Military industries are related with the public sector also in real life. They are both industries that with their operations interact and help this sector. As we can derive, these facts can be reflected from our results.

Moreover, AS operation covers the 21% of Airlines/Aviation industry and 18% of Consumer Goods industry in relation to their LinkedIn connections. Sales operation covers the 19.2% of Consumer Electronics industry. After discussions with IBMers we found out that these results reflect the reality, as these operations actually cooperate with companies in those industries.

In figure 4.14 we can see the coverage of each industry according to the operations. As we can see here the connections are well distributed in the operations. The majority of the operations are mostly connected with Computer Software industry, with exception ConS (Consulting) operation which is mostly connected with Management and Consulting industry with a percentage of 13.5%. From experience we can understand
that IBMers, due to company’s global operations, have many connections with Computer Software industry, as they have a variety of reasons to be connected with (Clients, Ex-Colleagues, Ex-classmates etc.). Additionally, the operation of Consulting is very similar with LinkedIn’s pre-defined Management and Consulting industry. Thus, an IBMer who works for a department who operates in Consulting, is reasonable to be connected with other users of the same industry.

Moreover, Software Sales operation has 36% of its connections covered by Computer Software industry. Again, here if we use information that we already know about IBM’s general operation and follow what experience suggests we understand that this fact reflects reality. Public Sector operation has 10% of its connection covered by Defence & Space industry, which is the 2nd after Computer software. This should be normal as the corresponding industry is highly related with public sector. One more interesting insight from this figure is that Finance Sector is the only operation that
Banking and Financial Services industries cover about 10% of its connections each, following again what our experience suggests. Additionally, Sales is the only operation that Computer Hardware covers 8% of its connections.

To summarize, in this section we can see that after a statistical analysis we are able to confirm our research hypothesis and extract business knowledge about the internal organization and key organization’s employees. Additionally, we can retrieve information about the department and the company, by analyzing their connections, which in some cases give us the ability to identify the department and its operations.

RQ 2.3: How can we infer information about changes in an organization’s structure from its employees’ professional social network profiles?

With this question we aim to investigate if we are able to retrieve information about changes in an organization’s structure by monitoring the profiles of its employees in a professional social network. Nowadays professional social network users use their profiles as a digital form of their Curriculum Vitae and update their positions whenever there is a change. With this research question we aim in taking advantage of this action and investigate if it is possible to proceed into conclusions about the organization that the employee works for.

LinkedIn, gives the ability to the user to publish his positions in his profile. The information that a user can publish is a description about the job role, the company that he works for, start and end date (day,month,year). To answer the current question we use the following:

Data: The data that we use to answer this question was retrieved from IBMers who work in the Netherlands. For them we can retrieve information about their current positions.

Approach: For all the IBMers who work in the Netherlands of our dataset:

1. Identify their current positions that are inside IBM.
2. Find the latest current position (as it is the one who is reported as current position also in Bluepages).
3. Analyze the start date from LinkedIn and compare the results with actual information from Human Resources department.

We have started the analysis procedure by identifying the busiest years of hiring/moving according to LinkedIn data. We identify the years of 2012, 2011, 2013, 2010 and 2008 as the top five years with 168, 146, 138, 98 and 66 reported start times for a position respectively. We then identify the months when the changes took place.

From Figure 4.15 we can see the distribution of hirings/movings during the top 5 busiest years. A first insight is that January is the year with the biggest number of hirings/movings during all the years, having 40.41% and 32.7% of total year’s hirings/movings respectively. The difference is larger during 2013 and 2012 than in the other 3 years. During 2013 we can see that June, April and February follow with 13.7%, 11.64% and 10.27% respectively. Moreover, we can see that July, August,
4.3 Findings - Results

Case Study: IBM

Figure 4.15: This figure presents the distribution of hirings/movings during the top years according to number of changers.

October, November and December are the least busiest months with a percentage of 0.68% of the total hirings/movings each. During 2012 we can see a better distribution, having July as second and September as third with 13% and 10.1% respectively. All other months cover between 4-6% of total number, while May and December are the least busiest months with 2.38% and 3.57% respectively. January is the busiest month also during 2011 with 23.3% of total hirings/movings. During this year the 2nd busiest month is June, having 13.7%. The least busy months are February and December with 2.73%, while the rest months cover between 4-7% of 2011 hirings/movings. In 2010, we can see again that January leads with 23.46%. However it is interesting that October comes second with a percentage of 18.36%, which is only about 5% less than January. During this year we can see again that December is the least busiest month with about 2% of the total hirings/movings.

From figure 4.16 we can see the distribution of hirings/movings during an average year. As we can see, January is by far the busiest with about 30% of total year’s hirings/movings. At the second place we can find July with 9.09%, while the third place is being shared between April, June and September having 7.46% each. October, February and November follows with a percentage of 6.81%, 5.52% and 5.19% respectively. March and August follows with 4.87% and 4.54%, while the month of December is the least busy with a percentage of only 2.27% of total year’s hirings/movings.

Discussion

In order to investigate the reasons behind these results we had some discussions with employees of IBM’s human resource department. What we wanted to investigate here was if our research hypothesis holds and we are able to extract knowledge about changes in IBM’s structure by monitoring the profiles of its employees in LinkedIn. In specific, we mainly investigate the statement that January is the busiest month of the year regarding changes in reality. The results of our discussions were that it is true,
Case Study: IBM 4.3 Findings - Results

Figure 4.16: This figure presents the average distribution of hirings/movings per month during the busiest years.

as many hirings and movings take place. Thus, employees update their profiles based on these real world events. Because of this behavior, the analysis of data retrieved from LinkedIn can provide us with the needed information. These facts validate our research hypothesis.

Figure 4.17: This figure presents the departments’ growing during the years of 2011, 2012 and 2013.

Departments growing procedure

With the next investigation that we proceed we aim to identify information about the departments’ growing procedure. Specifically, we want to investigate the departments’
growing during last 3 years and extract knowledge about it. The approach that we follow is to analyze the start date of users’ current position and linked them with the department that they work.

In figure 4.17 we can see the departments’ growing during last 3 years (2011, 2012 and 2013). Here, we can get an insight about the years that were important for IBM. As we can see, it seems that last years many departments(10) have hired more than 60% of their total employees. Additionally, the department with number 240920 has hired 100% of its employees.

Discussion

For the investigation of the reasons behind these results we again have discussions with human resources department. Our goal in this procedure was to understand the results and investigate if they hold. From our discussions with HR, the reasons of our results were that some departments actually grow during the period of last 3 years. Moreover, a number of departments have changed their names during this period so they were marked as new positions from LinkedIn users. Thus, users’ reports in LinkedIn actually reflect the reality and can be used for extracting knowledge about changes in the organizational structure of IBM.

In summary, from the results and the investigations that we proceed to address this research question, we conclude that by using data retrieved from IBM employees we were able to support our research hypothesis. Namely, we were able to extract accurate information about changes in IBM’s organizational structure from its employees’ LinkedIn profiles.

4.3.3 Research Question 3: How can we identify distinguishing properties of different categories of employees from professional social network profiles?

With this research question we aim to investigate if the behavior of the user of a professional social network is influenced by external factors such as user’s country and industry, user’s position in the organization, organization’s size and industry, clients. Here, we try to identify these factors and the actual properties that they influence it.

RQ 3.1: How does a department and its operations influence the community of the user?

With this questions we aim to investigate in what extend do employees tend to connect with users who have a relation with their working environments. Such employees could be their colleagues, clients, managers etc. In order to answer this research question we raise 3 other sub-questions.

Do employees tend to connect with their managers? Do they tend to connect with their manager’s manager?

Data: The data that we use to answer this question was retrieved from our Core IBMers because is the only group that we have access to their communities. We use
Case Study: IBM

4.3 Findings - Results

<table>
<thead>
<tr>
<th>Manager</th>
<th>Connected with Manager</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>47</td>
<td>103</td>
<td>45,63</td>
</tr>
<tr>
<td>Y</td>
<td>15</td>
<td>26</td>
<td>57,69</td>
</tr>
</tbody>
</table>

Table 4.4: This table presents the number IBMers (Managers and Non-managers) and their connection with their managers

<table>
<thead>
<tr>
<th>Connected with Manager</th>
<th>Average Community Size</th>
<th>Sample Size N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>496,38</td>
<td>62</td>
<td>48,06</td>
</tr>
<tr>
<td>FALSE</td>
<td>325,44</td>
<td>67</td>
<td>51,94</td>
</tr>
</tbody>
</table>

Table 4.5: This table presents the relation between the connection with the manager and the community size

Core IBMers who’s manager is reported in Bluepages (129 IBMers).

**Approach:** The approach that we follow in order to answer this question was the following: For each core IBMer:

1. Identify employee’s manager from bluepages
2. Check if the employee is connected with his/her manager
3. Check if the employee is connected with his/her manager’s manager
4. Investigate the results and try to find a relation with the size of the community of each employee

From tables 4.4, 4.5 we can see the relation between the type of the employee and their connection with their manager. As we can derive from the first table, the majority (57.69%) of the employees who are marked as managers are connected with their managers. In the of employees who are not managers we can find a percentage of 45.63% being connected with their managers. If we proceed in a deeper investigation, we can see that employees who are connected with their managers seems to have a larger community. Specifically, employees who are connected with their managers have 496,38 connections in average, while the rest have 325,44.

**Do employees tend to connect with their manager’s manager?**

The next step is to investigate if employees tend to connect also with their manager’s manager. As we can see from the table 4.6 about the 20% of both categories (Managers and Non-Managers) tend to connect with their manager’s manager. Additionally, from table 4.5 we can see that employees who are connected with their manager’s manager seem to have larger community.

Table 4.8 summarizes the results of the comparisons, so we can see a clear comparison. As we can see employees who are connected with both manager and manager’s manager have the largest average community size with about 633 connections. Those who are connected with only their manager are second with about 426 connections.
### 4.3 Findings - Results

**Case Study: IBM**

<table>
<thead>
<tr>
<th>Manager</th>
<th>Connected with Manager’s Manager</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>18</td>
<td>103</td>
<td>17.47</td>
</tr>
<tr>
<td>Y</td>
<td>6</td>
<td>26</td>
<td>23.07</td>
</tr>
</tbody>
</table>

Table 4.6: This table presents the number IBMers (Managers and Non-managers) and their connection with their manager’s manager.

<table>
<thead>
<tr>
<th>Connected with Manager’s Manager</th>
<th>Average Community Size</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>589.83</td>
<td>24</td>
</tr>
<tr>
<td>FALSE</td>
<td>361.59</td>
<td>105</td>
</tr>
</tbody>
</table>

Table 4.7: This table presents the number of IBMers (Managers and Non-managers) and their connection with their manager’s manager.

<table>
<thead>
<tr>
<th>Connected With Manager</th>
<th>T</th>
<th>T</th>
<th>F</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected With Manager’s Manager</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>41</td>
<td>3</td>
<td>64</td>
</tr>
</tbody>
</table>

| Average Community Size | 633.71 | 426.05 | 282.67 | 327.45 |

Table 4.8: This table presents the number of IBMers their connection with their manager, manager’s manager and the relation of the results with their community size.

while those who are not connected with anyone or are connected only with their manager’s manager have about 327 and 283 connections respectively.

**Discussion**

From our analysis we can derive that despite the fact that managers tend to connect with employees positioned higher in organization’s hierarchy, the percentage of employees who have a connection with their managers is lower. Additionally, there is a relation between the community size and the fact that an employee has a connection with his manager. Specifically, employees who are connected with their managers have larger communities than the ones who are not connected. Moreover, employees who are connected with both manager and manager’s manager have the largest communities.

**In what extend do employees connect with their colleagues and clients?**

**Data:** The data that we use to answer this question was retrieved from our Core IBMers who work in the Netherlands. For this question we also use the answer of quantitative survey for the question "With which clients do you cooperate?".

**Approach:** Firstly we have matched the survey answers with the core IBMers of our dataset. Then we calculate the community coverage based on clients, ex-colleagues and colleagues. Finally, we investigate the results according to the community size of the employees.
The results of our analysis are presented in figure 4.18. As we can derive, a percentage of about 47% of the total user community is ex-colleagues, colleagues and clients. Specifically, Ex-colleagues and colleagues cover the 36.75% of the total community, while Clients cover 10.22% in average.

From table 4.9 we can see a comparison between the community size and the connections with colleagues and clients. As we can see, when user's community size is under the median, the connections with his/her colleagues cover a larger part of his/her community, while the connections with his/her client cover less than when the user’s community is larger than the median.

**Discussion**

We believe that a reason for these results might be that during the early age of his profile (where the user has less connections), the user mostly initiates connections with his/her colleagues. After passing this period and having more connections, the user is already connected with his colleagues.

If we compare our findings from this part with the findings from the survey we can conclude that we can validate our results. Namely, from the survey we have the conclusion that 90% of the users who hold a profile in a professional social network not older than six months, tend to add a new member to their community because he/she is colleague of them, while this percentage reduces in users with older profiles (72%-75% of them belongs in this category).
4.3 Findings - Results  

Case Study: IBM

Figure 4.19: Do country and industry affect the size of the community? T-Tests results.

RQ 3.2: Which are the factors that influence the size of the community of the user?

Data: The data that we use to answer this question was retrieved from Basic Users with community size less than 500 connections. The reason was that LinkedIn provides us the ability to retrieve the exact size of a user’s community if this number is smaller than 500.

Approach: The approach that we follow in order to answer this question was the following: Extract subset for each different factor

1. Checks which group of values satisfies the requirements for tests of significance (Size, Normality tests).

2. Proceed to test of significance for the different factors

The factors that we have access and we investigate were the following:

- Country: The country of the user as reported by LinkedIn
- Industry: The industry of the user as reported by LinkedIn
- Company’s Industry: The industry of the company where the user currently works as reported by LinkedIn
- Company’s Size: The size of the company as reported by LinkedIn

After extracting the data, we proceed to normality tests and t-tests in order to investigate the significant difference. The results for the values which have passed the normality tests are presented in the next tables.

Figure 4.19 presents the results of the tests of significance for country and industry. According to countries, we can see that there is a significant difference in the community size of profiles come from the Netherlands and Belgium, Netherlands and
Case Study: IBM  

4.3 Findings - Results

Figure 4.20: Do company’s characteristics affect the size of the community?

United States, Belgium and Great Britain. However, we can see that there is no such a difference between Belgium and United States, United States and Great Britain.

The next factor that we investigate is user’s industry. Here we can see that in the majority of the comparisons there is a significant difference between the different values. More specifically, users who state that they work in the Computer Software(CS) industry and those who work in Management and Consulting(MC) industry have a significant difference in their community size. The same applies to the comparisons of CS and Financial Services(FS) industries, CS and Banking(BA) industries, MC and FS industries. As we can derive, users of Computer Software industry seems to be different than users in other industries, in relation to their community size.

Figure 4.20 presents the results of tests of significance for factors that are related to the company that the users work. Namely, we examine the factors of company’s industry and company’s size. Here we can see again that user’s who work in a CS company seem to have different community size than the users who work in other industries. However, the equality of users who work in MC and Higher Education(HE) industry cannot be rejected. The same applies for user’s who work in HE and BA industries. According to the company size, we can see that in the majority of the cases equality cannot be rejected. The only exceptions are the pairs of companies with size larger than 10000 employees and size between 1001 and 5000 employees, companies with size of 11-50 employees with those with size between 1001-5000 employees.

Discussion

With this investigation we show that the community size of an employee can be influenced by external factors that are directly related with his/her real life environment. Specifically, from the comparisons we can derive that in the majority of the cases the results were that country, industry, company’s industry and company’s size could be factors that influences the size of the community of the user in his professional social network as in most of the pairwise comparisons t-tests identify significant differences.
Chapter 5

Conclusions - Future Work

This chapter concludes the work done in this thesis project. In section 5.1 we present our conclusions and in section 5.2 we describe the future work.

5.1 Conclusions

Extracting knowledge from social networks is, nowadays, vital for researchers and organizations who want to keep track of changes that take place in the market and are related to their operations and their fields. With such information they will be also able to better understand people trends in their daily interactions and get informed about what is happening in their lives. Moreover, for organizations, social networks analysis will provide them with useful information about internal organization’s operations, monitor their updates and changes and get insights of how they can improve them.

As part of this thesis, a prototype framework for extracting value from online social networks has been developed. The presented framework provides the ability to the researcher to collect a dataset from online social networks. Additionally, functions for enriching a social network dataset with additional information that will increase the quantity and the quality of the data have been implemented. Moreover, the presented framework provides functions for preliminary data analysis in order to convert the collected data in a proper format for powerful network and data analysis tools.

We performed a case study where we used the presented framework to extract business value in IBM Netherlands. For this case study, we had several meetings with IBMers in order to define the research questions and the business knowledge that is important for them to be extracted. We performed a set of interviews to investigate their views, understandings and ideas that they have. Additionally, we generalized the insights from the interviews by conducting a survey inside IBM. After defining the research questions that we aimed to answer we proceeded to the dataset collection and analysis using the presented framework. We present our results in chapter 4, where we conclude that we are able to accurately extract knowledge about an organization using employees’ social network profiles. A part of our results has been validated through discussions with IBMers who were related with the field that we were investigating.

This research has contributed in the field on online social networks’ analysis with the following ways:
5.2 Future Work

Conclusions - Future Work

- **Framework:** We present a framework that provides the researchers with the ability to extract knowledge and value from online social networks. Tools for receiving access permissions for the users, collecting professional social network dataset, enriching this dataset with other resources have been developed. Additionally, this framework supports the data analysis with powerful analytical tools. Overall, the proposed framework provides the researcher with the ability of investigating high level research objectives such as comparison of different social network interaction patterns, identification of internal organization information and key employees, investigation in organization’s evolution, identification of factors that influence the behavior of the users in an online social network.

- **Case Study:** To evaluate and present the potentials of our framework we performed a case study at IBM Netherlands, where we aimed to extract business knowledge from a professional social network using the proposed framework.
  - Survey: A survey has been designed and released for our case study. The purpose of this survey was to validate our hypothesis and answer a part of our research questions.
  - Dataset: A dataset of 40102 basic users’ profiles has been recorded. This dataset has been enriched with data retrieved from IBM internal resources (Bluepages, Financial Sector repository).
  - Results: An in-depth analysis has been done on the collected dataset and the results have been presented in this research. From the results we can understand that using the presented framework we are able to extract business value that will help the organizations who use it. Moreover, from the research done in the case study we answered the research questions and gave insights regarding the possibilities and the accuracy that social networks’ data analysis has. With our results we show that information about an organization’s structure, operations, evolution and influential factors can be discovered from professional social network data analysis.

5.2 Future Work

In this section we differentiate between future work related to our case study at IBM and future work related to the presented framework.

5.2.1 Future work for presented framework

For this thesis project we have designed, developed and presented a framework for extracting knowledge from online social networks. With this framework we integrate different services to achieve our goals.

The development of a Graphical User Interface, which will provide a researcher with no programming experience the ability to use its functionalities is highly ranked in the list with future work. Additionally, the GUI will also help in the representation of the dataset collection and preliminary analysis procedures providing a better user friendly approach.
The automation of the transition from preliminary dataset analysis to network and data analysis is also an interesting work to be done. The majority of network and statistical analysis tools provide the functionality of using their capabilities through their Application Programming Interfaces. Thus, instead of processing the dataset and providing an export in a format that is able to be imported from the tools, the framework could use these APIs and provide directly the results based on the researcher’s choice.

5.2.2 Future work for case study

We have used our framework for extracting business knowledge from IBM Netherlands and achieve to collect a dataset from a professional social network, enrich it using information retrieved from Bluepages internal framework and financial sector, analyze it and present our results.

Despite the fact that the analysis that has been done led to interesting and useful conclusions, it was done using a small part of core users (134). Furthermore, during the time that the dataset collection took place many of the profiles were not fully up-to-date. It would be interesting if IBM repeats the dataset collection in order to update the current profiles and retrieve information about a larger group of people and proceed to the same analysis. Another idea is to collect profile snapshots in different periods of the same group of users in order to analyze their behavior in respect with the time, age of the profile etc..

Moreover, for the analysis of the behavior of the user inside social networks, IBM should collect another dataset from a different type of social network for the same group of user. Thus, it would be possible to compare our survey results with the exact analysis results.
Bibliography


[25] Manu Sharma. Did you use one of these 10 most overused buzzwords in your linkedin profile this year?, 2010.


Appendix A

Survey

In this appendix you can find attached the survey that our participants have answered.
IBM: Knowledge from social networks

The results of this survey will be used in a research project with topic "Extract knowledge from social networks".

We ensure you that the collected data will not be used for any other reason than this research and will not be published inside or outside IBM.

The procedure will not take more than 7 minutes.

Thank you in advance.

There are 19 questions in this survey

Professional Questions

1 [Q00]

In which company do you currently work?

* 

Please choose only one of the following:

- IBM Anguilla
- IBM Antigua and Barbuda
- IBM Aruba
- IBM Austria
- IBM Bahamas
- IBM Barbados
- IBM Belgium
- IBM Denmark
- IBM Bermuda
- IBM British Virgin Islands
- IBM Finland
- IBM Canada
- IBM Cayman Islands
2 [Q01]
With which companies do you professionally interact? (Separate by comma if there are more than one companies)
*
Please write your answer here:
Social Networking Questions

3 [Q10]
Which social network do you mostly use for your social connections? ("Social" Social Network)

* 
Please choose only one of the following:

- LinkedIn
- XING
- Facebook
- Twitter
- Google +
- Other

4 [Q11]
Which social network do you mostly use for your professional connections? ("Professional" Social Network)

* 
Please choose only one of the following:

- LinkedIn
- XING
- Facebook
- Twitter
- Google +
- Other
5 [Q12]
How old is your "social" social network profile?

* 

Please choose only one of the following:

- Less than one month
- Between one and six months
- Between six months and one year
- Between one and two years
- Between two and three years
- Between three and five years
- More than five years

6 [Q13]
How old is your "professional" social network profile?

* 

Please choose only one of the following:

- Less than one month
- Between one and six months
- Between six months and one year
- Between one and two years
- Between two and three years
- Between three and five years
- More than five years
7 [Q14]
What are the reasons to connect with someone in your "social" social network?

* 
Please choose all that apply:

☐ Colleague
☐ Friend
☐ Business (eg. Client, Potential Client)
☐ Other: ________________________________________

8 [Q15]
What are the reasons to connect with someone in your "professional" social network?

* 
Please choose all that apply:

☐ Colleague
☐ Friend
☐ Business (ex. Client, Potential Client)
☐ Other: ________________________________________
9 [Q16]
How frequent do you visit your "social" social network?

* Please choose only one of the following:

- More than once daily
- Once daily
- More than once a week
- Once a week
- Less than once a week but more than once a month
- Once a month
- Never
- Other

10 [Q17]
How frequent do you visit your "professional" social network?

* Please choose only one of the following:

- More than once daily
- Once daily
- More than once a week
- Once a week
- Less than once a week but more than once a month
- Once a month
- Never
- Other
11 [Q18]
How many hours do you spend on your "social" social network? (per week)
*
Please choose only one of the following:

- More than 20 hours
- Between 15 and 20 hours
- Between 12 and 15 hours
- Between 10 and 12 hours
- Between 5 and 10 hours
- Between 2 and 5 hours
- Between 1 and 2 hours
- Less than 1 hour
- None

12 [Q19]
How many hours do you spend on your "professional" social network?
*
Please choose only one of the following:

- More than 20 hours
- Between 15 and 20 hours
- Between 12 and 15 hours
- Between 10 and 12 hours
- Between 5 and 10 hours
- Between 2 and 5 hours
- Between 1 and 2 hours
- Less than 1 hour
- None
Personal Questions

13 [Q20]
Gender

*  
Please choose only one of the following:

☐ Male
☐ Female

14 [Q21]
Education

*  
Please choose only one of the following:

☐ High School
☐ BSc Degree
☐ MSc Degree
☐ Ph.D Degree
☐ Other

15 [Q22]
Country of Origin:

*  
Please write your answer here:
16 [Q23]
Working in country of origin?

* 
Please choose only one of the following:

☐ Yes
☐ No

17 [Q24]
Please enter the IBM department that you currently work:

Please write your answer here:

18 [Q25]
Please enter your IBM email address: (Your email address will not be matched with your answers and will not be published anywhere inside or outside IBM)

* 
Please write your answer here:

19 [Q26]
Do you want to participate in future surveys for this project and receive insights about your social network's community?

* 
Please choose only one of the following:

☐ Yes
☐ No
Thank you for completing the survey!
You can now navigate to project's website!

Submit your survey.
Thank you for completing this survey.