Universidade do Minho

Escola de Engenharia

Departamento de Informática

Jorge Caldas

Analysis and Visualization of Dynamic Social Networks

December 2016

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Jorge Caldas

Analysis and Visualization of Dynamic Social Networks

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Pedro Rangel Henriques
Alda Lopes Gançarski

December 2016

ABSTRACT

This document represents the work developed under the master's thesis Analysis of Visualization of Social Networks until the present date. First we present and identify the problem and challenge of building a system for network analysis, we also mention the motivation, research hypothesis and goals. From here we start to search the science foundations for social networks, starting from the very basic theoretical concepts in chapter two, then we travel to the present and present Online Social Networks as the most known application of this science, in chapter 3 we do a more detailed study on Online Social Networks starting from exposing theme in a more generalist way and then narrowing and exploring some of then with more detail. In chapter 4 we cover the analysis theoretical with the tool we want to build in mind, being Social Network Analysis a very broad field we take advantage of our goals to narrow the research on this field, presenting only the concepts and tools that in a certain way built the path for Chapter 5, where we propose an yet non much detailed architecture schema and description of some features that we think that our system should implement.

RESUMO

O presente relatrio de pr-dissertao, representa o tabalho desenvolvido sobre o tema de Anlise e Visualizao de Redes Sociais Dinmicas, at presente data. Comeamos por apresentar e identificar o problema e desafio de construr um sistema para anlise de redes sociais, mencionando os nossos objetivos e motivao para o projeto. No segundo captulo comeamos a explorar as fundaes cientficas sobre redes sociais, apontando os conceitos mais bsicos. No terceiro captulo fazemos uma passagem pelas Redes Sociais Online, comeando por fazer um levantamento mais generalista acerca do tema e das redes sociais online, aps esta introduo, fazemos um estudo mais detalhado acerca de algumas redes sociais que selecionamos. No seguinte captulo apresentamos a base terica da Anlise de Redes Sociais, uma rea bastante basta e complexa. Por ser uma rea de grande dimenso e diversas aplicaes, limitamos o seu estudo ao propsito desta dissertao, sendo que o critrio para explorao de alguns conceitos tero como base o facto de nos ser til a sua compreenso para a implementao do sistema, cuja arquitetura ainda numa fase embrionria, apresentada no captulo final.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In this chapter we do an introductory overview on the work being developed along this master's dissertation. This chapter presents the essential introductory topics. First we present the problem and context where this project is framed, then we expose the motivation, followed by the research hypothesis, which concisely describes the possible outcome of this project. Finally we list the goals of this project in a generic and simple way.

1.1 CONTEXT AND PROBLEM

In the mid 1950s sociologists introduced the term social networks, that despite being a familiar term for today general public because of the OSNs platforms such as Facebook, Instagram or Twitter, it is a deeper and more mature concept. It was in the 2000s that much of the OSNs we know today start emerging, so it took at least ten years to people to adopt the concept and the new way of living, so today billions of people use these online platforms as channels for socializing, connect with each other and share their daily lives.

From the user's point of view we may consider that all the platforms offer a microscopic perspective from within the network, people have a public profile, and they can visualize their friend's profile (this is a typical scenario that we observe today in the majority of the OSNs), and normally have access to a timeline that displays friends activity. The point is that, to the users of these online platforms, it is not provided a mean to visualize and analyze their network structure in a more abstract and generalized sense, where users are given the opportunity to observe their social network from a macroscopic perspective, and, with that, all the metrics for measuring nodes and relationships within the network.

The problem that is being built in this section resides on general social structure observation and analyzes. This dissertation aims to fill the gap or struggle that online social networks users have in understand their network, how their relationships evolve along the time, what role they play within the network and how can they analyze and visualize their networks based on social properties such as mutual relationships, geographical position, personal tastes and preferences or hobbies.

Within the big data challenges, social network data analysis might be one of the biggest

demands that we face today, because besides of dealing with tremendous amounts of data, we are dealing with unstructured data. The unstructured data derives from the diversity of this platforms known as OSNs, and unstructured data adds complexity to the challenge of analyzing social networks data. The major challenges related with big data and unstructured data comes after the data extraction.

The steps for data analyzes and visualization

Next we present the steps trough data extraction to data visualization, that generally represent the structure and flow of data analysis and visualization systems.

- Data extraction trough social media APIs or trough web crawlers (also known as web scrappers);

- Saving data, and more importantly know what data to store; in order to have an efficient system that provides a good structure for data analysis, one needs to selected data carefully;

- What to do with the data, what applications the stored data may have, how can the system digest and transform data in order to make it useful or interesting for the end users;

- How to present/show the transformed data, despite the science of visualization represents only a small part of the data scientist work, it has a huge impact on the end user, mainly when targeting a general audience.

1.2 MOTIVATION

As we see in the previous section, social media data analysis represents a major challenge for data scientists in every aspect, since the extraction all the way to the visualization. Despite representing a major technological challenge, social media data analyzes has an additional motivation, that is the massive daily usage in every country across the planet making OSNs an universal tool for communication, such as radio or television but with the technological flavor of the 21st century.

   OSNs as we will see along this dissertation, are today a *"digital mineral"* in terms of exploration potential, we do not only pretend to have a generalist perspective of the analyzes of data that flows within this platforms, we will try when appropriate to demonstrate the most narrower applications as possible of analyzing social networks, this applications may go from health analyzes within social structures, to strategic marketing planning supported by the analysis of the already mentioned unstructured data.

1.3   RESEARCH HYPOTHESIS

With this master's dissertation, we aim to prove that a software tool may be designed and implemented in order to actually improve the analysis of social phenomena, allowing not only sociologists but also the public in general to explore with greater detail the connections of individuals within a network, being OSNs the base of analysis for such a tool.

1.4   GOALS

The main goal of this project is to build a useful software tool in the context of social network analysis. Along the process of building and investigating, the following are some of the goals that are also very important to achieve:

- Understand the theory of *Social Network (SN)* in sociology;

- Understand how OSNs came to such a massive use nowadays;

- Perceive the roles of Online Social Networks in society and their potential applications in various fields;

- Study and analyze the most used Online Social Networks, learn how to interact with those systems and how to learn and profit from them;

- Design a system of analysis and visualization that matches the desired goals and requirements;

- Explore new technologies and choose the appropriate tools to build the specified system;

- Implement the system.

SOCIAL NETWORKS IN SOCIOLOGY

Nowadays, it is hard to find something that is not organized as a network, if one tries to understand something about the world around us, then definitely one needs to know something about networks.

Curiously, if we look up the term SN in the Dictionary (2002), we may face the following:

> *"a website or computer program that allows people to communicate and share information on the Internet using a computer or mobile phone"*

But, even if today we automatically think in SNs as websites (or web applications), deep down we know when talking about SNs, we refer to a much more broader term, that said, we may consider a SN as the following:

> *"A social structure made of nodes that are generally individuals or organizations. A social network represents relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. The term itself was coined in 1954 by J. A. Barnes."* Beal (2016)

One may say that networks work like pipes, and through them things flow, from individual to individual inside the network. Trough networks, big institutions can organize themselves, and actually add value to society despite the large number of individuals.

2.1 ORIGINS OF SOCIAL NETWORKS

> *"The network concept is one of the defining paradigms of the modern era."* Kilduff and Tsai (2003)

The network concept is broadly used across multiple fields of study, including, physics, biology, linguistic, anthropology, mathematics, computer science and more recently computer networks.

But why is the network approach so adopted in such diversification fields? According to Kilduff and Tsai (2003), the answer is because networks allows us to capture the interactions of any individual unit within the larger field of activity to which the unit belongs.

Before reviewing the concept of network (Section 2.2), it is important to talk about it in a sociological perspective.

2.2 SOCIOLOGY PERSPECTIVE

"(...) many people attribute the first use of the term "social network" to Barnes (1954). The notion of a network of relations linking social entities, or of webs or ties among social units emanating through society, has found wide expression throughout the social sciences. (...)" Wasserman and Faust (1994)

The SN concept has been around for many years now, maybe not in the exact format than nowadays, we are familiarized with the *"web way"*, in a manner of speaking, but in a more abstract sense, applied in real life within real connections. Wasserman and Faust (1994), refer that this term has first came into discussion in 1954, introduced by Barnes, J.A.

"Social relations in Bremnes, Norway, fall into three categories: relatively stable formal organizations serving many different purposes, unstable associations engaged in fishing, and interpersonal links that combine to form a social network and on which perceptions of class are based. In fishing situations, orders are given and obeyed; in the other social settings, consensus decisions are reached obliquely and tentatively." Barnes (1954)

In the above citation, John Arundel Barnes, does a very well succeed reflection about the relationships of the people from Bremnes (Norway).

The author points out that relations can form organizations for serving a specific purpose, and today we clearly see that the chosen path of SNs and also OSNs, was narrow down SNs to very specific purposes, such as professional networks. So one may say that John Arundel Barnes not only coined the term SN, but also was one of the first who described **interest-based social networks**.

2.3 FUNDAMENTAL CONCEPTS

The concepts listed below are of key importance and are the basis of comprehension of SNs (Wasserman and Faust (1994)).

- *Actor* - It is important to understand the linkages among social entities and the implications of these linkages, these social entities are described as actors. Actors are are discrete individual, corporate, or collective social units.

- *Relational Tie* - Actors are linked to one another trough *social ties*. The type of ties may be extensive, and it describes the nature of the connection. Some example of ties:

  - **Evaluation** of one person by another;

  - **Transference** of resources (business transactions);

  - **Association** (to social event or cause);

  - **Behavioural** interactions (communicating);

  - **Moving** between places or statuses (migration, social or physical mobility);

  - Others may be: physical connection (roads, rivers), formal relations (authority), biological relationship.

- *Dyad* - The most basic relationship that can be established is a dyad, a connection between two actors.

- *Triad* - A relation established between three actors. Many studies included breaking SNs down to small groups (triads), this allowed a more clear conclusion about the transitivity of the connections.

- *Subgroup* - It defines any subset of actors in a SN (conceptually, subgroups come after dyads and triads).

- *Group* - A finite set of actors who for conceptual, theoretical or empirical reasons are treated as a finite set of individuals in which network measurements are made.

- *Relation* - A collection of ties of a specific kind among members of a group is called a **relation** (e.g. a connection in *LinkedIn* is a relation while evaluating our connections of sending them messages are ties).

- *SN* - At last, with the definitions of actor, group and relation, a SN consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relation information is critical and defining feature of a SN.

Next, we present a two more advanced and abstract concepts but still fundamental concerning SNs in the context of this project.

Homophily

In a New York Times Magazine article ([Retica](2006)) it is mentioned that the term *"ho-mophily"*, was coined in the 1950s by sociologists and in a more literal sense it means *"love the same"*. This term emerges from the natural tendency we have to link to other individuals that are similar to us.

Quoting the sociologists [McPherson et al.](2001), **Similarity breeds connection**, basically similarity is considered a generator of connections among individuals, being the result of this phenomena homogeneous SNs.

The term *homophily* has been cited in light of many different themes, from teenagers choosing friends who drink and smoke similar amounts to theirs, or in explaining how homophily influences the matches of partners in online social dating, this proving that one likes, most of the time, someone like oneself, on or offline ([Fiore and Donath](2005)).

From another point of view, this trend could be seen as a threat to diversity and globalization. It is said that diversity can be a synonym of power, when bringing different cultures and different ways of thinking together we could achieve great things, but homophily is already a cemented concept/pattern that sociologists observe among SNs, and maybe we could find ways to battle in favor of diversity, or maybe homophily is a fundamental property in order to structure society.

Heterophily

In order to complete the previous presented concept (*homophily*), we now present the opposite that is *heterophily*, that translates in literally the opposite idea, being *heterophily* the trend of individuals belonging to diverse groups thus connecting with different people.

2.4   ABSTRACTION AND GENERALIZATION

In a more abstract sense networks are merely abstractions that are originated by the generalization both of individuals, and relationships.

> *"When we study social organization of a simple society, we aim at comprehending all the various ways in which the members os the society systematically interact with one another. For purposes of analysis we treat the political system, the pattern of village life, the system of kinship and affinity, and other similar areas of interaction as parts of the same universe of discourse, as tough they were of equal analytical status, and we strive to show how the same external factors, principals of organization and common values influence these different divisions of social life. "* [Barnes](1954)

In the above citation, the author describes a generalist approach on analyzing social networks. The two main characteristics of this approach are **generalization** and **abstraction**. First generalization because we are trying to simplify reality by minifying different kinds of connections (political, affinity etc.), this will allow us to treat networks as part of a world where they can fit in the domain of the exact sciences, being mathematical the way networks express themselves in order to measure metrics and behavior analysis.

Abstraction comes naturally in the way as the process of generalization takes laces, we could see abstraction and generalization as synonym in this specific case, but it also may be seen as a tool to see through the generalization process. Also fitting (at least try) networks and their analysis within the domain of exact sciences, requires the abstraction of the generalization that toke place before. In Chapter 4 we will cover with much more detail the field known as *Social Network Analysiss (SNAs)*, that is responsible of deriving conclusions from analyzing social structures.

ONLINE SOCIAL NETWORKS

People need to connect other people, and the urge for connection brings to us what today are known as OSNs. These web sites allow us to define a profile as an individual, and to share and visualize content with other individuals in the network, therefore connecting.

> *"We define Online Social Networks as web-based services that allow individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site.* Ellison et al. (2007)

OSNs have been around for more than a decade now, but these systems have gain world wide popularity since the global adoption of platforms such as Facebook, Youtube or Twitter, which are platforms that are today massively used across all cultures and age groups, and represents a paradigm shift on social interaction that we not yet fully understand.

The earlier referenced OSNs, belong to the top of the most visited web sites in the world, that's because these systems not only represents a new way to keep in touch with friends, but also represents for many, a new way of living, basically we live in network.

In this chapter we are going to explore OSNs, their history, how are these systems being adopted among Internet users, and for some OSNs, a more detailed and deep study will be conducted for they are important objects of study of this master's thesis.

But first, with intent of obtaining a macroscopic perspective of the different OSNs in the Internet, what they offer that makes them different from one to another causing many of the users using multiple OSNs at the same time, we present next a table featuring some of the most used OSNs.

Name	Year of launch	Registered Users	Active Users	Description/Purpose
Facebook	2004	>1 712 000 000	1 712 000 000	**General**. Photos, videos, blogs, apps.
Google+	2011	1 600 000 000	300 000 000	**General**. Google+ is an interest-based social network that is owned and operated by Google.
Youtube	2005	>1 000 000 000	1 000 000 000	Allows billions of people to discover, watch and share originally-created videos. Provides a forum for people to connect, inform, and inspire others.
Qzone	2005	>652 000 000	652 000 000	**General**. It allows users to write blogs, keep diaries, send photos, listen to music, and watch videos. It's only available in Chinese.
Twitter	2006	645 750 000	313 000 000	**General**. Micro-blogging, RSS, updates.
Tumblr	2007	>555 000 000	555 000 000	Microblogging platform and social networking website.
Instagram	2010	>500 000 000	500 000 000	A photo and video sharing site.
LinkedIn	2003	>450 000 000	106 000 000	Business and professional networking.
Sina Weibo	2009	300 000 000	282 000 000	Social microblogging site in mainland China.
VK	2006	249 409 900	100 000 000	**General**, including music upload, listening and search. Popular in Russia and former Soviet republics.
Reddit	2005	234 000 000	120 000 000	Social media, social news aggregation, web content rating, and discussion website.
Vine	**2013**	200 000 000	100 000 000	Short-form video sharing service where users can share six-second-long looping video clips.
Pinterest	2010	176 000 000	100 000 000	The worlds catalog of ideas. Find and save recipes, parenting hacks, style inspiration and other ideas to try.
Flickr	2007	112 000 000	92 000 000	Helping people make their photos available to the people who matter to them. Enable new ways of organizing photos and video.
Meetup	**2002**	27 590 000	-	World's largest network of local groups. Meetup makes it easy for anyone to organize a local group or find one of the thousands already meeting up face-to-face. meetup.com (2016)
Couchsurfing	2004	12 000 000	-	Couchsurfing connects travelers with a global network of people willing to share in profound and meaningful ways, making travel a truly social experience. Is commonly used by travelers to find free hosts across the globe. couchsurfing.com (2016)
ResearchGate	2008	>11 000 000	-	Built by scientists, for scientists. Connect the world of science and make research open to all. researchgate.net (2016)

Table 1: Table describing most used OSNs. (statista.com (2016), expandedramblings.com (2016))

Table 1 lists the most used and popular OSNs, **ordered by the estimated number of registered users**. Also notice that, for those OSN where the number of registered users is unknown, we will assume that it is a larger value than the monthly active users represented by the column *Active Users*.

The first obvious comment on the listed OSNs is that general purpose OSNs have more users (social networks with the word *General* in bold), being Youtube an exception, since it is not a general purpose OSNs, neither is focused on individuals, it is build around **social objects**, the videos.

The grey scale in the first column of Table 1 divides OSNs in three groups: the first and smallest, the 1 billion or more users OSNs; the second the OSNs with less than 1 billion users and more then 100 million; finally, the third group, OSNs with less then 100 million users. At this point, we begin to observe that **the narrower purpose OSNs** such as ResearchGate (mainly for researchers) or Couchsurfing (mainly for open minded travelers), **have a smaller number of registered users**, which is expected since the target audience is also smaller.

Other OSNs not listed in Table 1, but still worth mentioning include **Classmates** (helps users finding classmates form kindergarten, primary school, high school etc.) known for being one of the first OSNs, since it was launched in 1995, and **Ask.fm** (allows users to interact with other users asking and answering questions (revealing identity is optional)).

An important note on the listed OSNs in Table 1 is that only Qzone, Vine, Couchsurfing and ResearchGate don't provide any web APIs to fetch data or publish content, while all the others offer a wide variety of web services for developers to consume and use as they please, of course within the terms and policies of use of each OSN.

3.1   HISTORY OF ONLINE SOCIAL NETWORKS

Although the first platform possessing some of the main characteristics that define OSNs, according to Ellison et al. (2007), the first recognizable OSN launched in 1997 as we can observe in the Figure 1. *SixDegrees.com* allowed users to create personal profiles, connect with friends and consult friends of friends lists. The profile feature came from the online dating sites and online communities, while the surfing trough register users in the network and consulting friends was an existing feature in Classmates.com. *SixDegrees.com* was the first to combine these features.

SixDegrees promoted itself as a tool to help people to connect, but in 2000, it became an unsustainable business and the service closed. At the time the creators conclude that *SixDegrees* was a service that was very ahead of its time.

Until 2002 many OSNs have emerged, but still incapable of projecting themselves at a global scale. As we can observe in the timeline of Figure 1 from 2002 and 2005 the *big*
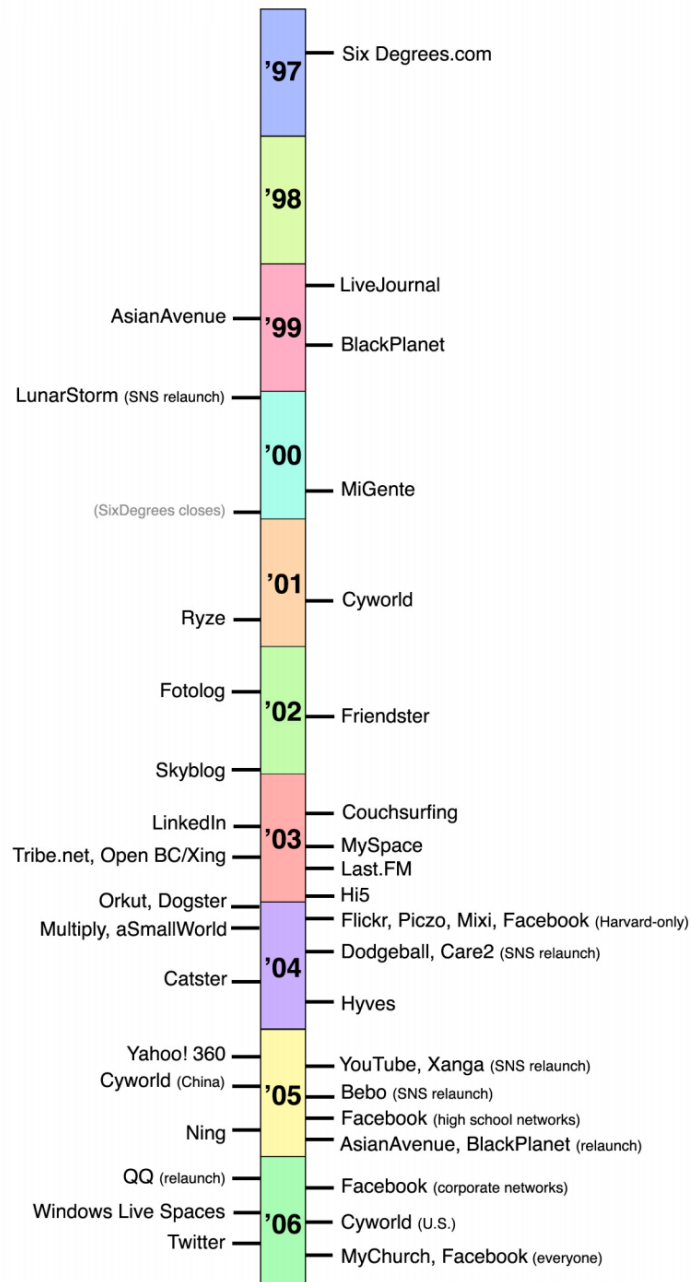
Figure 1: Launch dates of major OSNs. (Ellison et al. (2007))

players came to existence, in these period, OSN such as Friendster, LinkedIn, MySpace, Hi5, Facebook and Youtube were born, shaping the business, cultural and research landscape.

3.2   PORTUGUESE PEOPLE AND ONLINE SOCIAL NETWORKS

From Table 1, we get a good overview on OSNs usage among modern society. In this section we do a deep exploration of the most adopted OSNs by portuguese citizens, and get to compare then with the more global scenario presented in Table 1, also, other interesting facts will be revealed where appropriate.

A recent study, Marktest (2016), revels portuguese relationship with OSNs. This study, has been made by *Marktest Consulting* since 2011, with the goal of know the notoriety, utilization, opinion and habits of portuguese concerning social networks. The study information was collected trough online interviews. The sample was built from 819 interviews from individuals with age between 15 and 64 years, living in Portugal and using OSNs in a daily basis.

Some of the most interesting facts revealed in this study, relative to the participants are:

- 94% has a Facebook account and 43% a Youtube account;

- 21% has abandoned a social network in the past year;

- 27% considers that their dedicated time to social media has increased;

- 67% follows celebrities and 62% follows brands;

- 87% is used to watch videos in social networks.

These are indeed interesting conclusions, but what about the top used OSNs, **the most used are the following (by order): Facebook, Youtube, Google+, LinkedIn, Instagram and Twitter**.

Relatively to Marktest (2016) past studies, Facebook has maintain the top position, maintaining a grow tendency that has been standing out in the past years.

Going back to Table 1, we may now comment the usage of OSNs by portuguese people comparing it to the global scenario. As one may notice Facebook still rules users preferences within portuguese people.

Concerning to global time related usage statistics, according to Marktest (2016), **portuguese spend 91 minutes a day with social networks**, 68% considers that this is the ideal time to spent with social media, despite 1 in each 4 saying that in the past year has dedicated even more time to them. Even if people spent more than one hour and an half in this platforms, the study concluded that **67% of the users that visit OSNs several times a day only 41% does daily publications**.

The prime time for using OSNs is between 8pm and 10pm, being the smartphone the most used device in this time. Also in this short period the featured OSN is Facebook, the majority of the interviewed say that is the most credible site, the one that provides better and useful information, the most interesting and addictive.

3.3 EXPLORING SPECIFIC ONLINE SOCIAL NETWORKS

In this section we are going to explore in greater detail some of the OSNs presented in Table 1. The selection of the social networks was not aleatory, we are going to study deeply the OSNs that gather some important characteristics, that will be of use in the future when we design the system for analyzing and visualizing social networks. First, the OSN must be accessible, this said, one must be capable of extracting information from the platform in order to analyze it. Second, the OSNs should preferably be the most diversified as possible, so that we can draw different types of conclusions deriving from different kind of analysis, for then give proof of the adaptability of the system to different OSNs. Considering the previous comments, these are the following OSNs that we think that as a group, better represents the intentions previously mentioned, so we will cover them with more detail (with no particular order):

- Facebook;

- Instagram;

- LinkedIn;

- ResearchGate;

- Pinterest;

- Twitter.

3.3.1 *Facebook*

Facebook is an OSN, created by Mark Zuckerberg in 2004, which started out by being an exclusive social network for Harvard students, but came later to spread across the country and the globe, having today more than one billion users.

Before diving into details of Facebook's domain, one must first point out some of its general aspects. Facebook basically allows anyone with a valid email address to create a public and personalized profile, we say personalized in terms of displayed content or information such as profile photo, name, work, homeland, education etc. . The next fundamental step is connect with other users, by sending friendship requests to other Facebook users (this are bidirectional relations). The base entity of the network is the user, but entities such as brands, companies can also be part of the platform, appearing normally in the form of page, being a page a public place inside the network with marketing or business related purposes (celebrities, public institutions also use pages as form of appearing in Facebook).

The next parts of this section will clarify the roles of this entities and their way of interact with each other, also other important concepts will be presented.

Domain Model

In this section we explore the domain of Facebook represented in Figure 2 in detail, what are the pieces that conceptually build this platform, and how they relate. The schema in Figure 2 represents a macroscopic perspective among Facebook components and their organization.

There are two entities with bold labels in the schema, this are, **User** and **Post**, being *User* the base entity in the network (the node in the network graph basically), and *Post* the most basic unit of content sharing in Facebook.

Facebook is interesting in terms of data gathering, because despite offering users' basic information and to whom that users are related (*Friends* box), it has a collection of other interesting data such as the family relationships (*Family* box), geographical locations where the user lives, or visited locations (*Locations* and *VisitedPlaces* boxes respectively), and among other things, user information may contain the personal interests that were explicitly inputed by the user (*Likes* box).

In what concerns to user activity in the platform, the *Timeline*, provides all the user Posts chronologically ordered, this is where Facebook dynamism takes place, users are constantly adding content to their timeline, it may be life related events or simply sharing other users posts linking content. The user feed (*Feed* box) represents a global timeline where the user can consult all the posts on his network (this is by default the user's landing page on the platform).

Facebook has, with time, become more then a user profile centralized network, it has invested in expand its horizons, becoming the place where pages of brands, companies, organizations (media, political, non-profitable etc.), or places (cities, monuments, bars etc.) live (*Page/Local* box). This entities that are now cohabiting with users in the Facebook ecosystem, take advantage of the platform and its range to get their updates to most people as possible. The profile for these pages are in many ways different form the user's profile, it also has a timeline, but the about information and other details represent a smaller part of page's profiles, the most important metric for pages is its number of *likes* (*Likes* box), it represents the number of users in the network that follow the page, it might be users that simply have a certain relation with the entity or simply want to keep in touch by regularly receiving these entities updates in their Facebook walls [1].

Other Facebook entities not yet mentioned, are events (*Event* box). These are events inputed in the platform that allow users to keep updated about relevant events happening mainly in their area. Users can tag the event as *interested in*, showing their friends the will of participating in some event, or they can simply reject the event. Users also can confirm participation on events showing their network that they will be present. Events keep three

---

[1] Facebook wall an area where users can see the posts of their friends and/or liked pages, in a chronological order
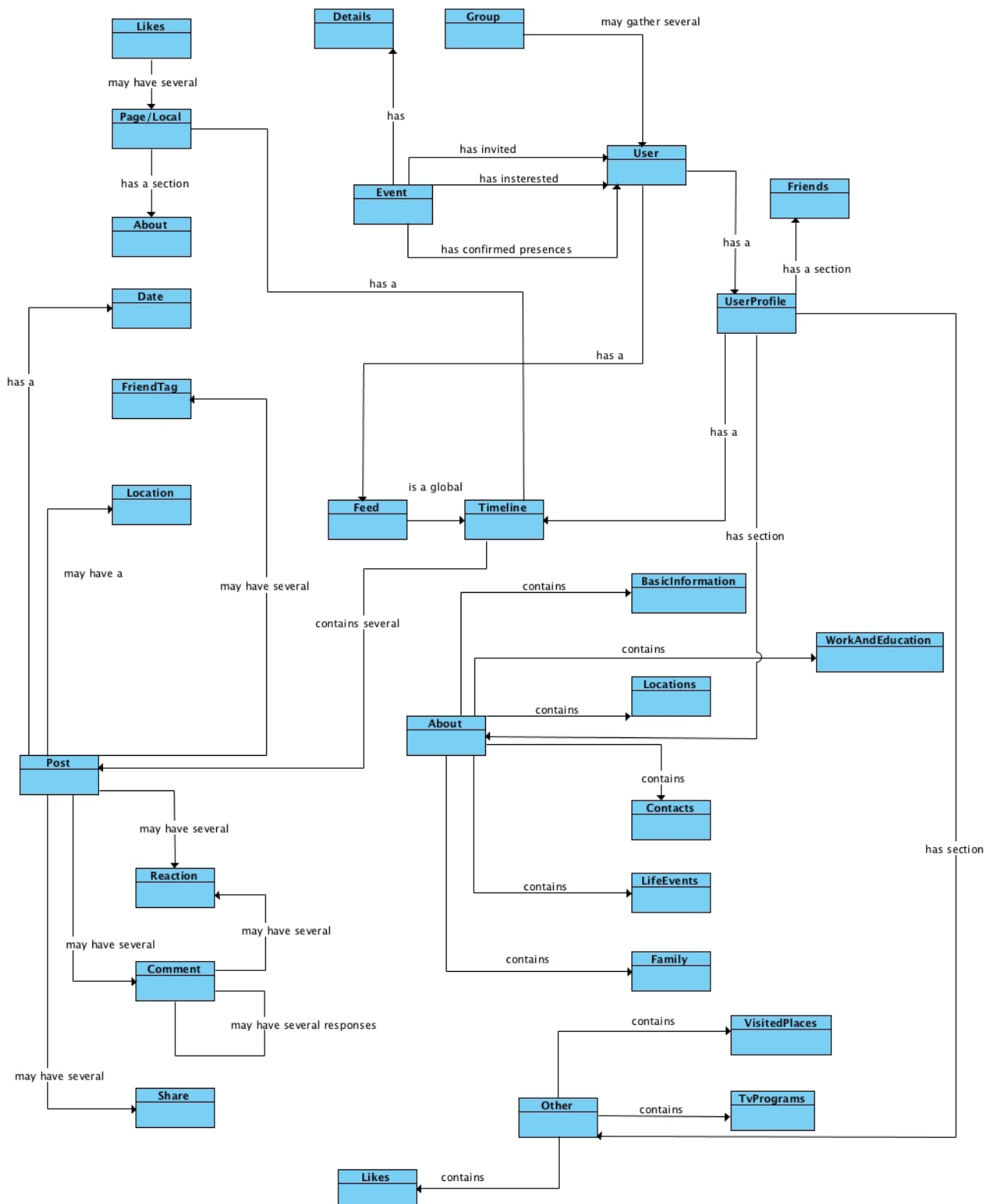
Figure 2: Facebook domain model schema.

separated counters for users, they count the number of invited users, number of interested users and number of confirmed users (these relations are expressed as links between the *Event* box and the *User* box).

In Facebook is also possible to join groups of users, this groups may be public or private, and they generally are focused on a specific matter, or gather users from one same institution or organization (e.g. Facebook group of students of the University of Minho). Having this feature of groups, clustering users by they interests one may say that groups, some way, transform Facebook in a *"multi interest-based OSN"*.

Facebook Graph API

Facebook has today several software *kits* for developers to interact with the platform in the most diversified and imaginable ways. Facebook developers offers a range of variated software products that vary from monetization programs, that focus on how to make users profit from Facebook, Analytics to developers who have their apps embedded in the Facebook platform understand their audience and the performance of their apps, etc. (developers.facebook.com/products/ (2016)).

In this master's thesis context, the relevant software that Facebook has available is the Facebook Graph API. This API basically allows developers to collect information from Facebook such as posts, photos, videos, pages etc. According to developers.facebook.com/docs/graph-api/common scenarios (2016), the common scenarios for using the Graph API are the following: determine whether two people are friends on Facebook; publishing new status and updates, uploading content (photos, video etc.); sharing links. But in this project what we seek is build the most biggest and detailed network as possible, with analysis and visualization purposes in mind.

For building the network fetching users friends information is crucial, this was possible until Facebook Graph API v2.0 (trough the router */me/friends*), one could actually retrieve friends information and build a network from there. From v2.0 on, to achieve what was explained before, one must request a special permission called **user_friends** from each user. The permission **user_friends** is no longer included by default in every login. This change breaks down the possibility of gather Facebook information via its Graph API, this said, we need in the future to look up alternative paths to extract data from Facebook.

3.3.2   *Instagram*

> *"Since the beginning, Kevin has focused on simplicity and inspiring creativity through solving problems with thoughtful product design. As a result, Instagram has become the home for visual storytelling for everyone from celebrities, newsrooms and brands, to*

teens, musicians and anyone with a creative passion." https://www.instagram.com/about/us/ (2016)

Similarly to Facebook we are going to explore Instagram in the same way. Instagram was originally developed by Kevin Systrom and Mike Krieger, and launched in 2010, only for iPhone devices. Within a year Instagram was able to gather around 10 million of users. Later, in 2012 Facebook acquire Instagram for approximately 1 billion dollars.

As already mentioned in Table 1, Instagram does not belong to the group of general purpose OSNs, instead, Instagram specially focused on photo and video sharing, building a global community that shares more than 95 million photos every day.

According to https://www.instagram.com/about/us/ (2016), since the very beginning Instagram was a very simplistic platform, being this characteristic reflected on its domain model.

Domain Model

Figure 3 represents the domain model of Instagram, and as we can observe, simplicity its the essence of this platform, since this diagram is far more a realistic representation of Instagram than Figure 2 is a representation of Facebook, and this may be why Instagram is so massively adopted by users on the Internet, because it goes directly to the point, focusing mainly on sharing activity, offering a real easy and simple user experience.

Now concerning to the domain model, we can see that a user and its profile (*User* and *UserProfile* boxes) are very simple entities, because a user's profile is only its biography (*Biography* box), relationships (*Followers* and *Following* boxes) and the user's posts, that despite being chronologically ordered, do not intend to form any kind of timeline such as Facebook, instead it represents more the concept of a wall with frames hanged on it.

In Instagram the landing page, represents a timeline (*Timeline* box) with posts from users we follow. Regarding to posts (*Post* box), one can comment posts (*Comment* box), but one cannot react or respond to comments (this preserves simplicity even more, for nested comments represent a complex part of OSN such as Facebook), and react to them by the *like* reaction (*Like* box).

Instagram API Platform

In consequence of a simple domain, Instagram API Platform, provides simple and useful end points for programmatic publishing, and for network discovering, as far as concerning to this project, the late utility is more of interest. Instagram allows to get users, their relationships and also the media shared content (posts).

Similarly when exploring Facebook Graph API, we now found also very intimidating
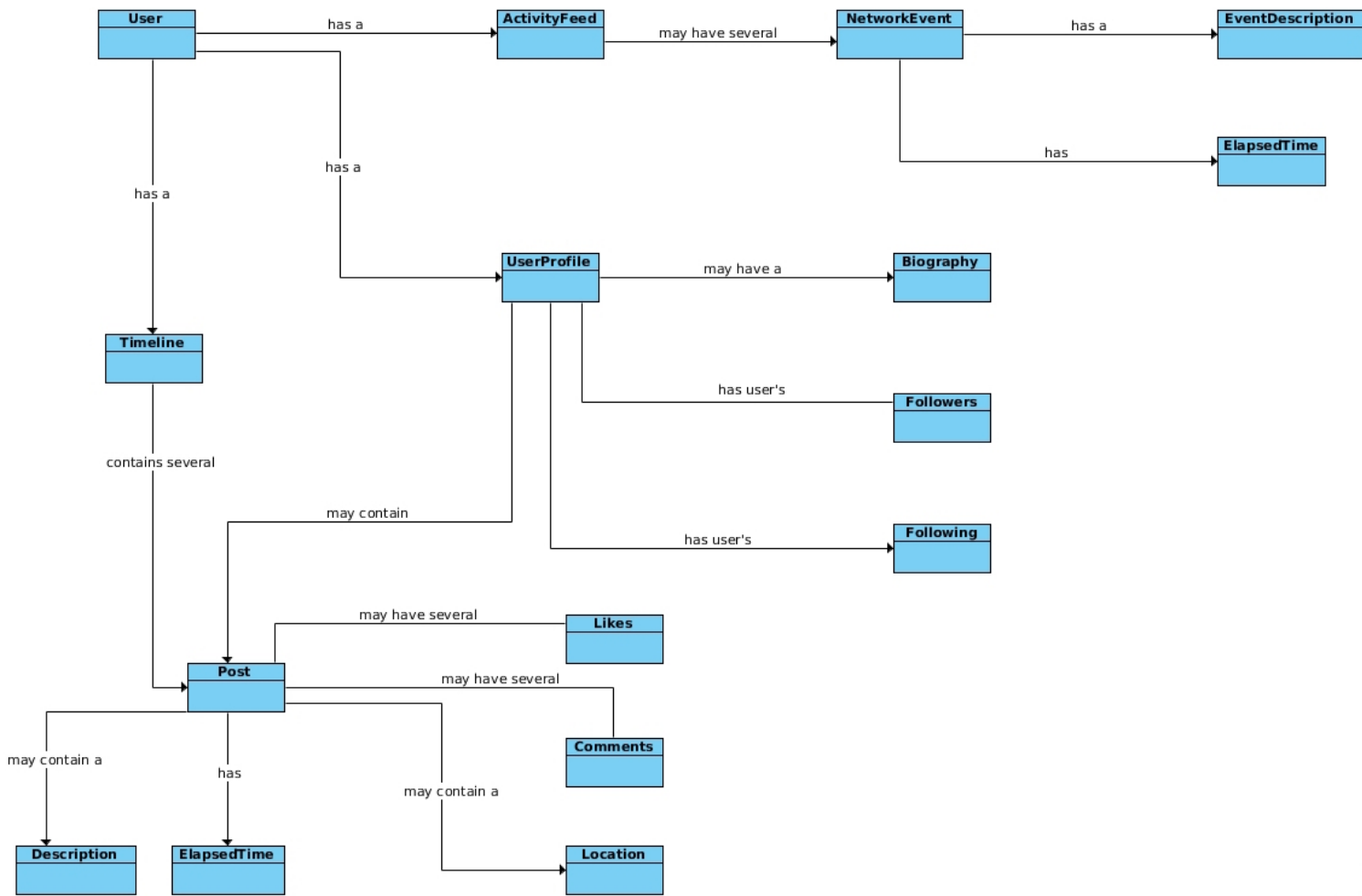
Figure 3: Instagram domain model schema.

restrictions for the purpose of this project, this restrictions include limited rate of 500 API requests per hour, and end point specific limitations that allow only to perform 30 requests per hour to getting users' relationships data. (https://www.instagram.com/developer/limits/ (2016))

3.3.3 *LinkedIn*

Moving on to the next OSN we now have LinkedIn. According to https://press.linkedin.com/about linkedin (2016), LinkedIn was launched officially on May 5 of 2003, and by the end of that month, the network had already more than 4500 members. In 13 June of 2016 LinkedIn was acquired by Microsoft in an all-cash transaction valued at $26.2 billion (Guardian (2016)).

LinkedIn is an OSN that has a very narrow purpose, which is connecting professionals around the globe to make them more productive and successful.
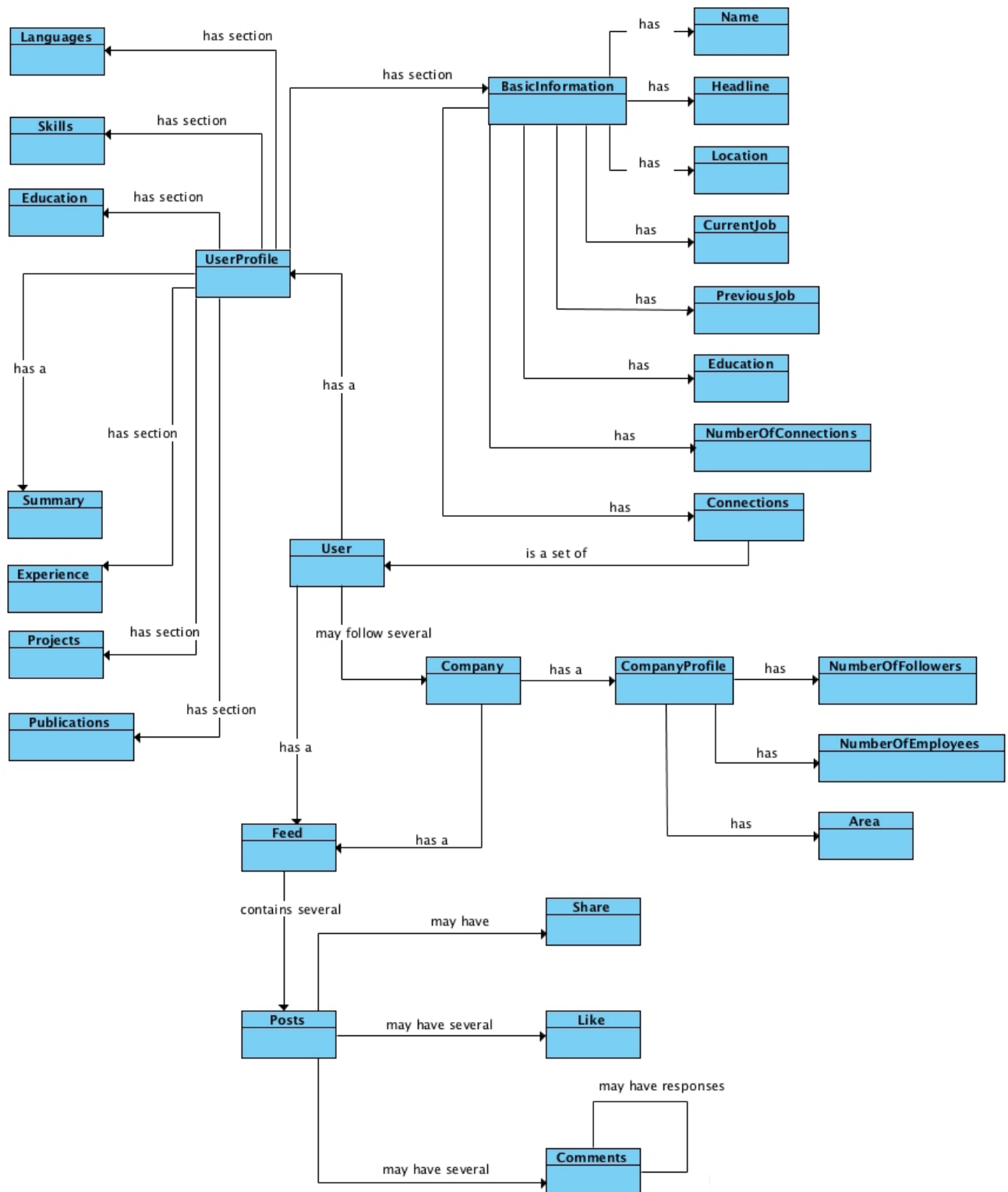
Domain Model

Figure 4: LinkedIn domain model schema.

Being a more purpose oriented OSN and focused on the professional world, makes LinkedIn platform more complex, even with a simplified representation of the domain model, as we can observe in Figure 4 it is schema [2] far more complex that Instagram, having more or a similar complexity comparing to Facebook.

In LinkedIn the user profile (*UserProfile* box) is very rich in terms of what is important for building an individual professional image (profile), starting by one individual's basic information (*BasicInformation* box) that has information like name, location and current and/or previous jobs. Then the user profile has several sections with very specific purposes such as professional experience (*Experience* box), languages (*Languages* box) or education (*Education* box), all this summed up give a very precise perspective of an individual's "professional appearence". At the bottom of the profile we have along with the professional recommendations and connections, the skills or expertise section (*Skills* box), this is one of the most attractive features in the LinkedIn platform. Skills in LinkedIn are a tagging system that allow user's to expose their expertise trough their public profile and then receive feedback on them according to their ability on that specific skill, this is a very important and promising feature for matching user's profiles with job positions requirements.

LinkedIn's main entities are not only users, the industry is massively represented in this network too. Companies may have a company profile (*Company* and *CompanyProfile* boxes) where they present the company, containing basic information such as number of people following the company number of employees (giving the idea of the company dimension) and the area where the company fits (pharmaceuticals, technology etc.) (*NumberOfFollowers*, *NumberOfEmployees* and *Area* boxes respectively).

Other important concept of LinkedIn is the user feed where the user can chronologically consult a series of posts produced by their connections or by companies that their follow.

LinkedIn API

LinkedIn provides a REST API (https://developer.linkedin.com/docs/rest api (2016)), but still similarly to the OSNs we been studying its very limited. In what concerns to data retrieval LinkedIn only allows the consult of basic profile data, this is the data retried from the LinkedIn interactive REST console:

```
{
"firstName": "Daniel",
"headline": "Graduate Front-end Developer at Blip.pt",
"id": "k_yk8W37WH",
```

---

2 In the schema presented on Figure 4, much of the platform complexity was simplified in order to produce a simple domain, and to narrow down this analysis to the core components and concepts of LinkedIn.

```
"lastName": "Caldas",
"siteStandardProfileRequest":  {
   "url": "https://www.linkedin.com/profile/..."
}
```

As we can see from the above data sample, we only could fetch some data properties, that would not bring value in terms of network analysis.

3.3.4   *ResearchGate*

> *"Founded in 2008 by physicians Dr. Ijad Madisch and Dr. Sren Hofmayer, and computer scientist Horst Fickenscher, ResearchGate today has more than 11+ million members. We strive to help them make progress happen faster."* researchgate.net (2016)

ResearchGate is an OSN built specifically for scientists, with the goal of easing the task of collaborative research around the globe. ResearchGate strikes to connect the world of science and make research open to all.
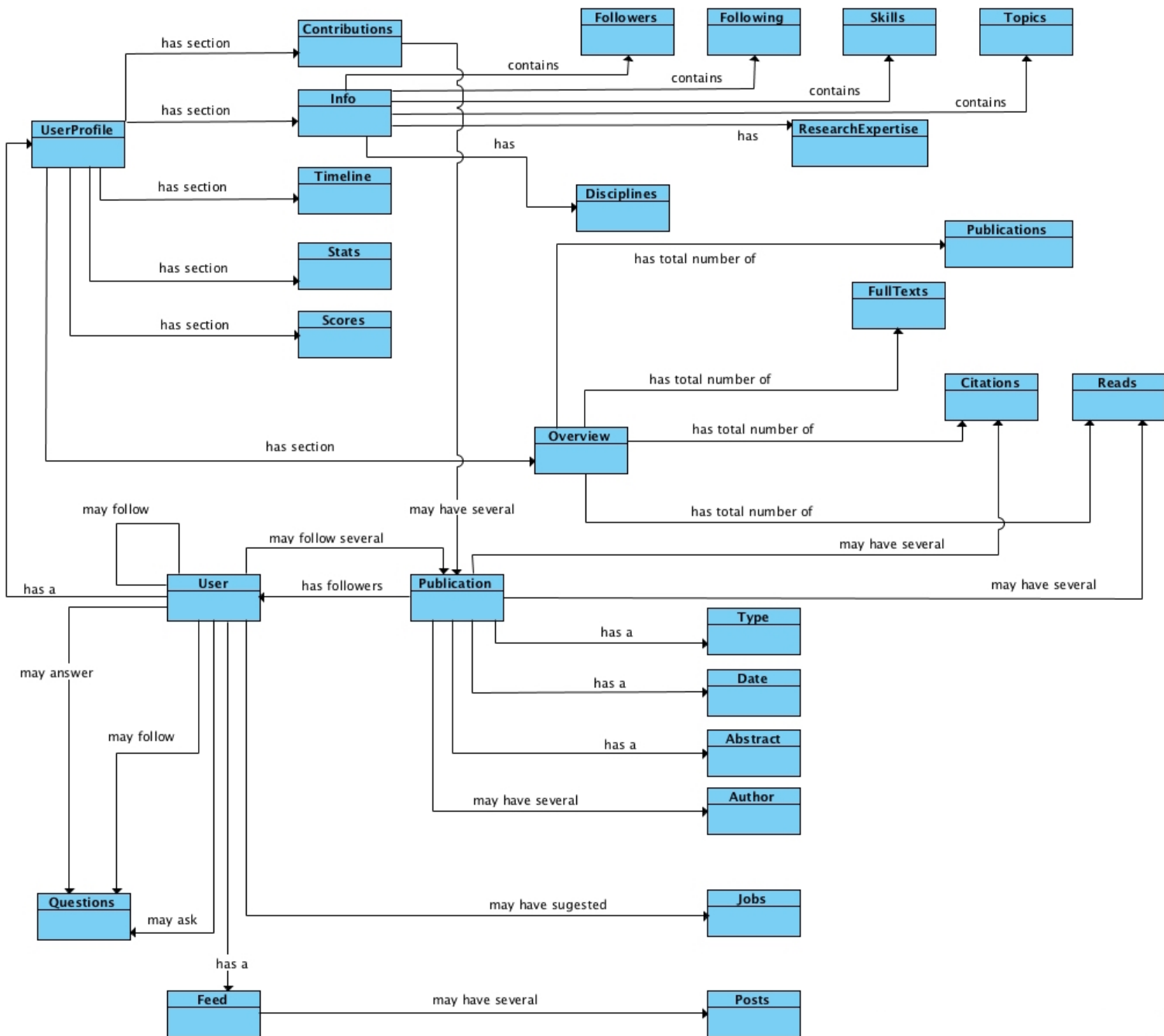
Domain Model

Figure 5: ResearchGate domain model schema.

Data Dictionary

Some terms on the schema presented in Figure 5 may be quite ambiguous due to the the specificity that they represent. In order to make the schema fully legible and before diving

into the domain model analysis, we present first, a small data dictionary detailing the terms one may found more ambiguous:

- **Scores** - This term represents a collection of metrics that evaluate the performance of a user based on his contributions and research experience. The user has also associated a global score;

- **Topics** - Topics represent the user's scientific areas of interest, ResearchGate uses topics to provide personalized suggestions;

- **Disciplines** - Represent more broad areas of the user education, expertise and interest;

- **Type** (*Type* box connected with the *Publication* box in Figure 5) - A type classifies a publication, this said, a publication may be an article, a book, a thesis a conference paper etc. .

Domain Model Analysis

ResearchGate is a peculiar OSN that despite having connections between individuals, it has alongside connections between individuals and scientific publications, making the publication (*Publication* box) a social object, playing the same role that videos play in Youtube for example.

Like LinkedIn the user profile (*UserProfile* box), is very detailed and builds up a very clear image of the researches work, positions and areas of interest. The relations among users are bidirectional, following the followers/following (*Followers* and *Following* box) strategy like other OSNs such as Instagram or Twitter. Very simillarlly to LinkedIn, a user's profile has a skills (*Skills* box) section, where skills are expressed in the form of tags, the tag description is far more specific than LinkedIn tags, that may some times acquire very abstract or high level descriptions (e.g. Technology Information). In ResearchGate tags have are very specific and are normally related with the user topics (*Topics* box) or disciplines.

Publications play along with the user a main role in ResearchGate. Normally publications have associated a type (already explained in the data dictionary section), a date, an abstract and may have one or more authors. The main metrics for Publications rating are the number of reads (*Reads* box) and the number of citations (*Citations* box) of that publication. The publications may also be followed by users that may have interest on particular publications.

Other concept of ResearchGate that raises the collaborative spirit among users, living up to the values that originated the platform, is the questioning system (*Question* box). Users may ask each other specific questions and have them answered by an expert on a specific scientific area, this opens up the possibility of having the best experts on a specific matter giving their opinion, thus the possibility of obtaining the *"best possible answer in the globe"*.

ResearchGate users' receive open jobs suggestions based on their profile, also user's have a post where they receive activity notifications of the people or publications that they are following.

API

Today ResearchGate does not provide any API for accessing its data or for any kind of interaction with the platform.

3.3.5 *Pinterest*

According to Pinterest (2016), Pinterest **is the world's catalog of ideas**. Created by Ben Silbermann, Paul Sciarra and Evan Sharp and launched in 2010, Pinterest is a simple but yet very original OSN, instead of aiming for connecting people like Facebook or LinkedIn, it aims for inspire people trough new ideas.

Domain Model

Data Dictionary

As one may notice from Figure 6, Pinterest introduces very particular concepts that may lack explanation, that is why we present first a small data dictionay before going trough the analysis, as we did with ResearchGate on a previous section:

- **Pin** - A Pin is the basic unit of Pinterest, it represents an idea of some user, presented in some context (the board context), and it is presented to us with a picture;

- **Board** - As the name suggests, a board is a collection of pins. Boards are created from users to other users, and normally present pins within some context (e.g. travels, technology, food etc.). In Pinterest boards may be followed by other users;

- **NumberOfPinedTimes** - This entity is not entirely a Pinterest entity, instead it represents a relevant metric introduced to measure pins popularity, and it refers to the act of saving pins. Pins that are presented to the users may be saved (or *"pinned"*), and the number of times that users have saved a particular pin is expressed in Figure 6 by the box *NumberOfPinnedTimes*;

Domain Model Analysis

Pinterest introduces new concepts forming a very original OSN, because it's very different from others that we analyzed previously. Just as we seen in ResearchGate, where the domain model is build around a social object (the scientific publication), with Pinterest we
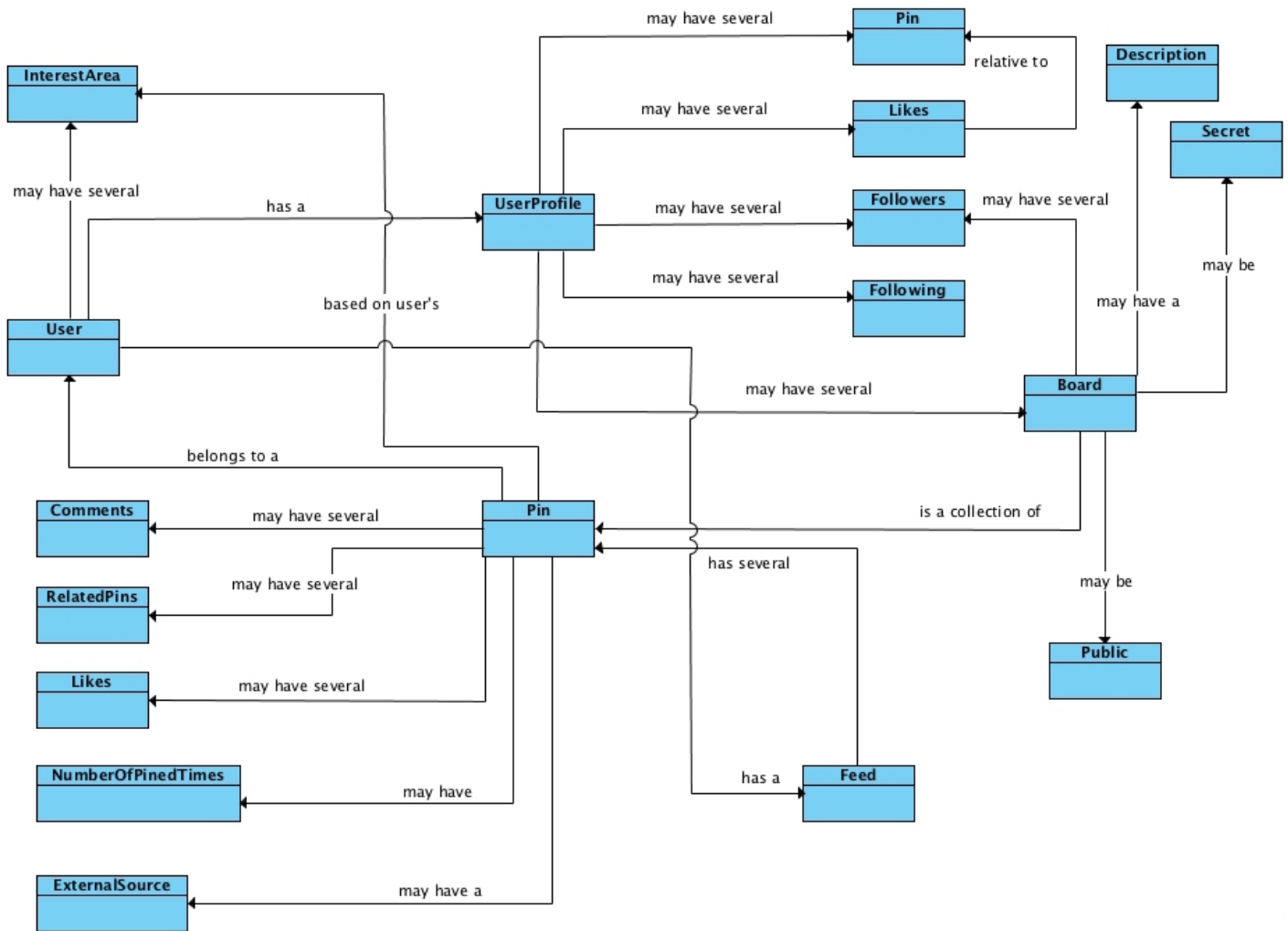
Figure 6: Pinterest domain model schema.

have a similar scenario, where the concept of the platform is built around a different social object the Pin (*Pin* box), which also as a grouped perspective introduced by a group or collection of pins that are the boards (*Board* pin). Pinterest is basically a set of pins aggregated in boards that are explored in the platform accordingly to the user's interests.

Simmilarly to other networks (e.g. Instagram) Pinterest also has direct unidirectional connections between users that adopt the concept of *"follow/following"* (*Followers* and *Following* boxes). As user's can follow publications in ResearchGate, Pinterest users may follow boards, being then notified if some pin is added to that specific board.

In what concerns to Pins, they may be commented by users (*Comments* box), they also may be targeted by likes as posts in Facebook (*Likes* box). A particular point concerning to Pins is that they can have a explicit external reference, for instance, if some image is extracted by some other web site or from other OSN they can be explicitly referenced, and that same reference appears at the top of the pin along with its title (*ExternalSource* box).

Pinterest was the traditional concept of feed, but in this case, the feed represents a completely different concept compared to other OSN. First the content of the feed (pins) is not related with users we follow on the network, is instead related is our personal interests (*InterestArea* box) and second, they are not presented according to a chronological order, and visually they do not follow the standards of typical timeline/feed design, instead the different pins displayed on some user's feed, form some kind of board or catalog, like the ones people use to hang in walls and pin post-its on it.

Pinterest API

According to Developers (2016), Pinterest provides a REST API for developers interact with the platform. The data restrictions follow Facebook politics, where the application that integrates Pinterest API can only fetch data for authenticated users. Pinterest provides endpoints to interact with users, boards and pins. Concerning to the requests limitation, Pinterest offers a 60 minute sliding window where 1000 requests can be made by unique user token.

3.3.6 *Twitter*

One OSNs that frequently is bring to discussion for being more of a *"news content generator"* is Twitter. Twitter is another of must most used OSN listed in Table 1, is basically a social networking microblogging service that allows their users to broadcast short posts (short because they're maximum size cannot exceed the 140 characters) called tweets. Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams, gaining fast worldwide popularity, Twitter has today more then 300 Million users according to Table 1.

Unlike many other social networks that have private or semi-public profiles with restrict policies concerning to exterior access to information within the network (examples of this kind of OSNs may be LinkedIn or Facebook), Twitter default settings are public, making tweets spread more effectively across all social media, this particularity makes Twitter one of the most *"barrier-free"* OSNs. Of course that despite unregistered people may read tweets they cannot interact with them as Twitter users by linking, comment or *"retweet"*[3] them.

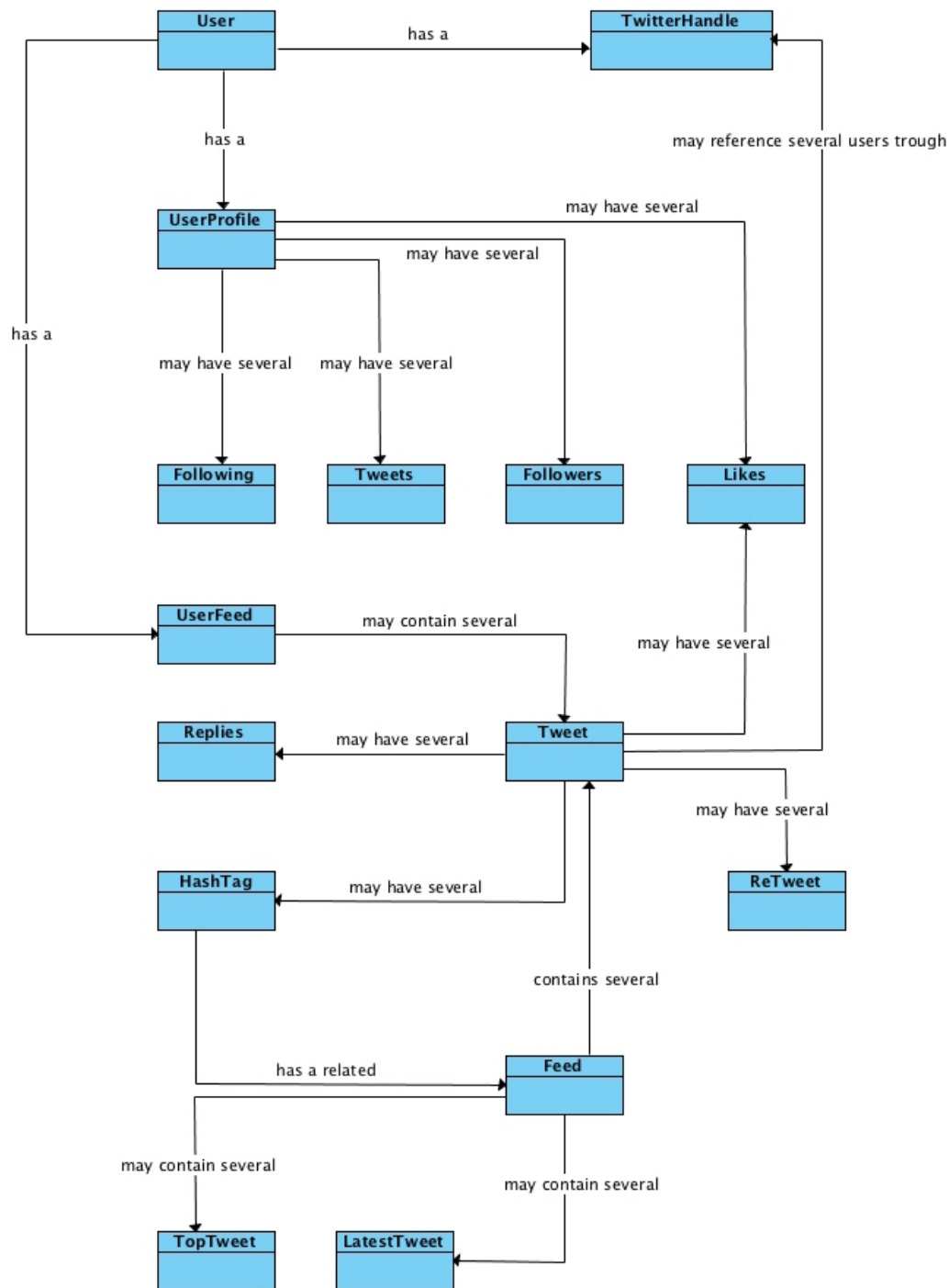3 The act of retweet consists and share some existent tweet originated by another user.

Domain Model

Figure 7: Twitter domain model schema.

Domain Model Analysis

In Figure 7 on may observe a very concise representation of the Twitter domain. Despite being a very minimalist OSN concerning to data properties and relationship complexity, Twitter has some very semantically strong features that brand the platform, being those features used among other well known OSNs such as Facebook. We are referring specifically the hashtag ((*HashTag* box) and the (*TwitterHandle* box), but let us first introduce some of the more basic features.

As usual in these kind of platforms, Twitter's users posses a user profile (*UserProfile* box) that as attached to it some properties and user metrics such as number of followers, number of tweets, number of other users the user is following, and number of likes the user has obtained across all his or her tweets (*Followers*, *Tweets*, *Following*, *Likes* boxes respectively).

The tweet is the basic unit of Twitter, is trough tweets that information flows in Twitter. Also tweets may have several comments, and they may be *retweeted*. Now back to hashtags and twitter handles. Hashtags is in some way how these chaos of unstructured tweets gain some semantic value, in order to a group tweets according to a specific matter. Hashtags may in many cases be misleading, because users start to adopt hashtag to express sentiments or simply describing a tweet pushing back to the already mentioned unstructured chaos. Twitter handles are the same as tags in Facebook they serve as a mean to mark a specific twitter user, a twitter handle may be used in a comment or in a tweet leading directly to the respective user's twitter profile.

Twitter API

Similarly to other OSNs, twitter provides a REST API to fetch users data such as profile data, tweets or user's followers. Restrictions are felt again this API is very limited providing a 15 minutes window for making http requests.

Other data sources

As more and more of OSN appear to be closing doors to data availability even when for research purposes one may want to search for alternative data sources to feed the system that we intend to build along this project. Projects as KONECT (Kunegis (2013)) provide large data sets of networks collected online so that researchers may perform all types of operations and experiments on that data. This kind of alternatives are very valuable in terms of network analysis even for real time data analysis systems that may in a more immature phase of the project benefit from this data sets.

"KONECT (the Koblenz Network Collection) is a project to collect large network datasets of all types in order to perform research in network science and related fields, collected by the Institute of Web Science and Technologies at the University of KoblenzLandau."
Kunegis (2013)

3.3.7 *Summary*

In this section we have explored with some detail six of the OSNs listed in the Table 1. In this analysis we followed a similar approach for analyzing each OSN, adding only an additional step for the more domain specific OSNs, that were ResearchGate and Pinterest, which was building a small data dictionary in order to ease the interpretation of the domain model schema.

From the analysis we may draw some generic conclusions concerning the domain of each OSN. Despite the differences and specifics of each platform, all them sum up to the basic primitive concepts of social networks, that are **actors** and **relational ties** between them, which form **subgroups** originating **groups** that build the network. This being the high level conclusion for our analysis, there are other patterns that emerge when analyzing different OSNs like the **user profile** that is a key element characteristic of this platforms and **feeds** (or **timelines**) that represent a standardized way of communicating events within a OSN.

3.4 how online social networks have changed the world

Social media have clearly shifted the way we communicate and we perceived the world, simply putting it, nowadays with social media one can say that social media is responsible for *"everyone talking to everyone about everything all of the time"*.

According to Duggan (2015), 62% of the entire adult population in on social media. As an example of events that were clearly influenced by social media, we have the presidential campaign of Barack Obama in the United States, started in 2007 and ended in 2009, Barack Obama had as his campaign technological adviser Chris Hughes, co-funder of Facebook, who played a crucial role in the camping trough online social media. The outcome of the election of 2009 could have been very different without the online social media.

According to Farida Vis (2016), a very interesting reflection is made on how social media impact the world, and the six major drawn conclusions are the following: **across industries, social media is going from a *nice to have* to an essential component of any business strategy**; **social media platforms may be the banks of the future**, as example we have the bank customer profiling trough social media in order to get a loan; **social media is shaking up healthcare and public health**, because information is spreader *at the speed of light* trough social media, this means less struggle to achieve public health and well

being awareness; **social media is changing how we govern and are governed**, with OSNs public participation has grown and everyone can participate in their opinion making people voices louder, bringing more credibility to the democratically system implemented by many governments across the planet; **social media is helping us better respond to disasters**, as the health public awareness improved trough social media information propagation speed, so did improved the response of governments and institutions to disasters such as natural disasters, in countries that may have not the services or infrastructure to respond to some catastrophes, making social media and crucial component to raise awareness across the globe, that have impact in help mobility, or fund raising for supports the damages made by certain disaster; **social media is helping us tackle some of the worlds biggest challenges, from human rights violations to climate change**.

If we look particularly to the most globally used OSN, as reported by Elgot (2015), there are pointed out *"seven ways Facebook has changed the world"*, we are going to point and comment out some of the more relevant. **Facebook has changed the definition of friend**, if back there having a dozen of friends was already a very large number of relationships, with Facebook the new limit was raised up to the hundreds or thousands of friends, the concept was given a completely new meaning, since we don't need to know a person face to face so that one becomes friend with the other, one simply needs to click the *"add friend"* button, and it does not matter if it is one's neighbor or some other person on the another side of the planet; **We care less about privacy**, *"if you are not paying for it, you are the product"*, means that we are not paying for using Facebook or any other OSNs, this said we must retain that these online platform profit from our information and from our interactions, but even being the major of the users aware of this situation, that doesn't seems to bother anyone; **Facebook has created millions of jobs but not in its own offices**, for example the marketing industry suffer a revolution since the raise of the social media, there are jobs for people to manage business and brands profiles on OSNs it's also a new way to approach customers, as we have seen previously with banks; **Facebook has been the tool to organize revolutions**, protests and awareness campaigns are raised inside facebook, this is related to the political influence and awareness capacity that we previously have pointed out in this same section.

Now switching to the negative aspects of not only Facebook but OSNs and social media in general. Very strong campaigns were raised against social media, for instance, *"The Anti-Social Network"* a short film depicting a life of an adult which as become obsessed with social networking at the point he starts to break boundaries between his real life and his virtual one. Strategically or ironically this campaigns use social media to spread the word.

We have seen that social media had a great deal of impact in society, what about our bodies? There are numeral studies on this matter, focusing on finding the true negative impacts of OSNs on our personal health. According to Lin et al. (2012), scans to brains of

people how excessively use social media, point out that there is a clear degradation of white matter similar to people who are addicted to substances such as drugs or alcohol, in the regions that control emotional processing, attention and decision making, because social media immediate reward (instant feedback) with very small effort, this causes the brain rewire itself make us to desire this stimulations Berridge and Robinson (1998). Another common situation among OSNs users is the idea of multitasking, the felling that one is able to being productive in some task while browsing on social media. Well according Ophir et al. (2009) users who heavily use social media are more susceptible to interference from irrelevant environmental stimuli, leading this users to perform worse on a test of task-switching ability, because they were not able to filter out interferences.

SOCIAL NETWORK ANALYSIS

SNA is the study of how people are connected to each other, basically it studies a set of relations among a set of entities, these entities may be individuals, organizations, or even countries.

The common analysis procedure consists in mapping the network and then create metrics to characterize the network. Then one tries to figure what is the structure of the network and why does it have that structure. SNA is also about look at the individuals inside the network and where are those individuals located.

4.1 FUNDAMENTAL CONCEPTS FOR NETWORK ANALYSIS

The concepts listed below are of key importance to understand SNA.Wasserman and Faust (1994)

- *Actor* - SNA is concerned with understanding the linkages among social entities and the implications of these linkages, these social entities are described as actors. Actors are are discrete individual, corporate, or collective social units.

- *Relational Tie* - Actors are linked to one another trough *social ties*. The type of ties may be extensive, and it describes the nature of the connection. Some example of ties:

  - **Evaluation** of one person by another;

  - **Transference** of resources (business transactions);

  - **Association** (to social event or cause);

  - **Behavioural** interactions (communicating);

  - **Moving** between places or statuses (migration, social or physical mobility);

  - Others may be: physical connection (roads, rivers), formal relations (authority), biological relationship;

- *Dyad* - The most basic relationship that can be established is a dyad, a connection between two actors.

- *Triad* - A relation established between three actors. Many studies included breaking SNs down to small groups (triads), this allowed a more clear conclusion about the transitivity of the connections.

- *Subgroup* - It defines any subset of actors in a SN (conceptually, subgroups come after dyads and triads).

- *Group* - A finite set of actors who for conceptual, theoretical or empirical reasons are treated as a finite set of individuals in which network measurements are made.

- *Relation* - A collection of ties of a specific kind among members of a group is called a **relation** (e.g. a connection in *LinkedIn* is a relation while evaluating our connections of sending them messages are ties).

- *SN* - With the definitions of actor, group and relation, a SN consists of a finite set or sets of actors and and the relation or relations defined on them. The presence of relation information is critical and defining feature of a SN.

4.2 GRAPHS THEORY

Graphs are typically the base of representation of social structures. This mathematical approach maps with extreme convenience social networks. Nodes are individuals, and edges are relationships. Despite looking a quite simple approach, there is a very strong theoretical background that is of basilar importance for interpreting social networks.

4.3 NETWORK ANALYSIS

In this section we intent to explore the scientific concepts behind network analysis, always trying to map them to reality, so only the core and applicable concepts will be explored in this section, namely:

- *Power Laws*;

- *Random Graphs*;

- *Centrality Measures*;

- *Link Analysis*;

- *Community Detection*;

- *Spread of Information*;

- *Social Learning*.

4.4 SMALL WORLD PROBLEM, STANLEY MILGRAM'S EXPERIMENT

This principle the *small-world phenomenon*, is based on the idea that all human beings are connected by **short chains of acquaintances**. The pioneer of this work was Stanley Milgram.

4.4.1 *Six Degrees of Separation*

This idea is an extension of the small world problem. In the sequence of what we state before, the six degrees of separation materialize the previous concept in six interconnections to some individual reach some any other one. The concept a strong particular relevance when a play was written in the 90's portraying this concept.

4.5 NETWORK VISUALIZATION

Network visualization may be consider as a science by itself, in this section we will explore some relevant techniques for network representation and visualization. We will look in particular to STAR

4.6 SOCIAL NETWORK ANALYSIS SOFTWARE

> *"(...) more sophisticated graphics capabilities should make exploratory studies using visual displays of networks more fruitful. One should be able to display actor attributes and nodal or subgroup properties (such as expansiveness, centrality, or clique membership) along with the graph. (...)"* Wasserman and Faust (1994)

4.6.1 *Structure*

The program Structure Pritchard Lab (2000) is a free software package for using multilocus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed.

4.6.2 *Gephi*

Gephi Bastian et al. (2009) is a tool for keen data analysts and scientists who want to explore and understand graphs. Like Photoshop but for graph data, the user interacts with the representation, manipulates the structures, shapes and colors to reveal hidden patterns. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research. https://gephi.org/features/

4.6.3 *UCINET*

UCINET 6 Lin Freeman (2002) for Windows is a software package for the analysis of social network data. It was developed by Lin Freeman, Martin Everett and Steve Borgatti. It comes with the **NetDraw ?** network visualization tool.

4.6.4 *SocNetV*

Social Network Visualizer Kalamaras (2004) is a cross-platform, user-friendly application for the analysis and visualization of Social Networks in the form of mathematical graphs, where vertices depict actors/agents and edges represent their relations.

With SocNetV you can construct social networks with a few clicks on a virtual canvas or load field data from various social network file formats such as GraphML, GraphViz, Adjacency, Pajek, UCINET, etc.

Furthermore, you can create random networks using various random models.

4.6.5 *NetworkX*

NetworkX Developers (2005) is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. NetworkX relevant features are listed below:

- Python language data structures for graphs, digraphs, and multigraphs;

- Many standard graph algorithms;

- Network structure and analysis measures;

- Generators for classic graphs, random graphs, and synthetic networks;

- Nodes can be "anything" (e.g. text, images, XML records);

- Edges can hold arbitrary data (e.g. weights, time-series).

4.7 REAL WORLD APPLICATIONS

Here we will present some of the real world applications of social networks analyzes with special attention for developed projects with a similar focus to this master's dissertation, and other tools of OSNs that cross many fields of studies. Two examples of projects with similar focus to this master's dissertation are:

- *Vizster* Heer and danah boyd (2005) - Visualizing online social networks. A visualization system for playful end-user exploration and navigation of large-scale online social networks;

- *Project Palantir* Facebook (2008) - This is an impressive tool that displays the rate of interactions on Facebook across the globe.
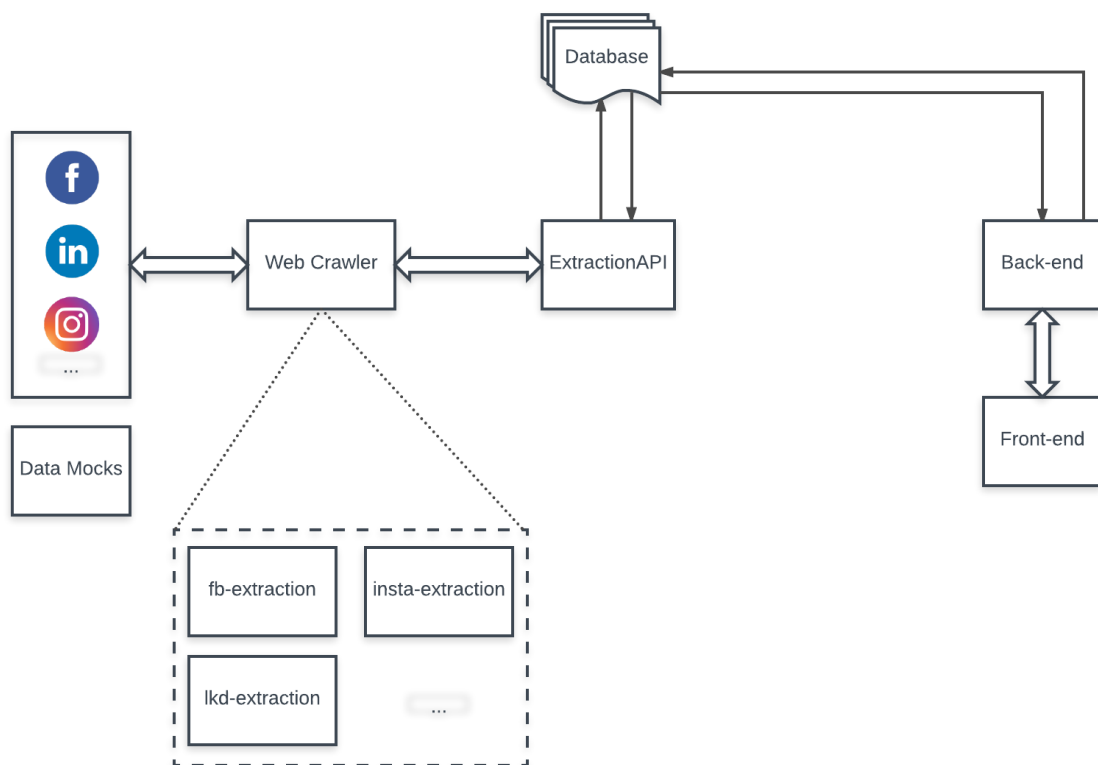
5

SYSTEM ARCHITECTURE PROPOSAL

Figure 8: System architecture proposal.

After exposing all the necessary theory and context on SNs and OSNs, we know present a more concrete image of the overall system. In Figure 8 we present an abstract system architecture.

5.0.1 *General overview*

As the interaction of the software components may be clear from the diagram, the role of each module its not clear by simple diagram observation, an underlying explanation of each component is needed in order to understand the system.

We will follow a *top down* approach for explaining the system architecture. First let us be clear about the two main and distinct parts of the system:

- *Application/Visualization and Analysis Tool* - The tool that directly interacts with the end user its composed by a **Back-end** that fetches data from a database, runs calculations and algorithms on top of stored networks as the user requests by interacting with a **Front-end** that provides the visualization and interaction features;

- *Information extraction and data mining* - All the other components are built for extract information from existent databases, or from OSNs (trough the **Web Crawler**) and store information being information properly treated before stored.

5.0.2 *Detailed Components Description*

Some of the components require a more detailed explanation, next we look more carefully into some of the components in Figure 8.

- *Data Mocks* - This represents the strong hypothesis that we feed some data trough data mocks, instead of crawling data from OSN. This data may be databases from projects that we already mentioned in this document, such as Kunegis (2013). This data would be accessed trough the **Extraction API**, or a new module could be constructed exclusively to feed this data to our database;

- *Web Crawler* - The **Web Crawler** consists in a set of modules for crawling each one of the OSNs (*fb-extraction* and other modules);

- *Database* - The database is where we store our data. It is not represented by the *classical CILINDRO* because it resembles relational databases, and the possibility of using non relational databases such as document data bases it grows strongly within the project, and the reason is the unstructured data that we will be storing into our database.

CONCLUSION

Until now, the worked developed under this master thesis project was manly research, not only we walked trough concepts in sociology that are of basilar importance for fundament and building the context to the problem we presented. Then we start focusing in information technology and presented the OSNs and explore them with some depth. Then with we present some theory of SNAs with focus in what we want to implement.

6.1 WORKING PLAN

In this section is defined the working plan for this master's thesis project, with the aim of structuring the project approach. Next, follows a list of the steps for this project.

- 1st **Research** - This first phase will be a exploratory overview on the theory of social networks, social network analysis and graphs visualization. It will also be a phase to study the state of the art of social networks analysis and visualization tools.

- 2nd **Requirements Specification** - Here, the system must be specified, among the specification, the proposed of the tool must be well defined, as well as the features that the system will provide to the end user. The requirements should also establish a range of social networks that the system could support.

- 3rd **System Modelling** - In this phase the focus goes to the system architecture design, conceptual data model and high level graphical interface mock ups.

- 4th **Choice of Technologies** - In the 4th phase the technologies to implement the previous specified system must be chosen. It will be an opportunity to explore the must promising technologies implementing some small prototypes, merely as proof of concept.

- 5th **System Implementation** - The implementation phase should be iterative, a divide and conquer strategy will be followed to break the system into small pieces so it grows the more naturally allowing coherence and robustness in the software.

- 6th **Evaluation and Conclusions** - For last the system must be evaluated and analyzed to see if it fits the purpose for which was build and if it fills the pre-established requirements. Concluding, a retrospective of the work must be done, such as a global overview on the developed system and some final comments on the obtained results.

6.2   WHERE WE ARE AND FUTURE WORK

At this point, the first phase is almost completed, missing only some more depth in the study of some concepts of SNAs in chapter 4. The requirements specification was already ramped up by the study and proposal of the architecture in the previous chapter.

The next step will consist in analyzing an choosing the best technologies suited for building the system, work that is already in progress. In parallel the system modeling and requirements will be defined.

BIBLIOGRAPHY

John Arundel Barnes. *Class and committees in a Norwegian island parish*. Plenum New York, 1954.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Vangie Beal. Webopedia definition for social network, 2016.

Kent C Berridge and Terry E Robinson. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369, 1998.

couchsurfing.com. Couchsurfing about page, 2016.

NetworkX Developers. Networkx. http://networkx.github.io/, 2005.

Pinterest Developers. Pinterest developers page. https://developers.pinterest.com/, 2016. Online accessed 29 October 2016.

developers.facebook.com/docs/graph-api/common scenarios. Facebook developer graph api, 2016.

developers.facebook.com/products/. Facebook developer products, 2016.

Cambridge Dictionary. Cambridge dictionaries online, 2002.

Maeve Duggan. The Demographics of Social Media Users. http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/, 2015. Online accessed 29 October 2016.

Jessica Elgot. From relationships to revolutions: seven ways Facebook has changed the world. https://www.theguardian.com/technology/2015/aug/28/from-relationships-to-revolutions-seven-ways-facebook-has-changed-the-world, 2015. Online accessed 29 October 2016.

Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

expandedramblings.com. Social media statistics, 2016.

Facebook. Project palantir. https://www.facebook.com/video/video.php?v=37403547074&ref=nf, 2008. Online accessed 5 December 2016.

Alejandra Guzman Farida Vis. 6 ways social media is changing the world. https://www.weforum.org/agenda/2016/04/6-ways-social-media-is-changing-the-world/, 2016. Online accessed 29 October 2016.

Andrew T Fiore and Judith S Donath. Homophily in online dating: when do you like someone like yourself? In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1371–1374. ACM, 2005.

The Guardian. LinkedIn bought by Microsoft for \$26.2bn in cash. https://www.theguardian.com/technology/2016/jun/13/linkedin-bought-by-microsoft-for-262bn-in-cash, 2016. Online accessed 22 October 2016.

Jeffrey Heer and danah boyd. Vizster: Visualizing online social networks. In *IEEE Information Visualization (InfoVis)*, pages 32–39, 2005. URL http://vis.stanford.edu/papers/vizster.

https://developer.linkedin.com/docs/rest api. Linkedin developer docs, 2016.

https://press.linkedin.com/about linkedin. Linkedin about page, 2016.

https://www.instagram.com/about/us/. Instagram about page, 2016.

https://www.instagram.com/developer/limits/. Instagram developer page, 2016.

Dimitris V. Kalamaras. Socnetv. http://socnetv.org/, 2004.

Martin Kilduff and Wenpin Tsai. *Social networks and organizations*. Sage, 2003.

Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM, 2013.

Fuchun Lin, Yan Zhou, Yasong Du, Lindi Qin, Zhimin Zhao, Jianrong Xu, and Hao Lei. Abnormal white matter integrity in adolescents with internet addiction disorder: a tract-based spatial statistics study. *PloS one*, 7(1):e30253, 2012.

Steve Borgatti Lin Freeman, Bruce MacEvoy. Ucinet software. https://sites.google.com/site/ucinetsoftware/home, 2002.

Marktest. Os portugueses e as redes sociais, 2016.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

meetup.com. Meetup about page, 2016.

Eyal Ophir, Clifford Nass, and Anthony D Wagner. Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106(37):15583–15587, 2009.

Pinterest. Pinterest about page. https://about.pinterest.com/en, 2016. Online accessed 29 October 2016.

Stanford Pritchard Lab. Structure software. http://pritchardlab.stanford.edu/structure.html, 2000. Online accessed 5 December 2016.

researchgate.net. Researchgate about page, 2016.

Aaron Retica. Homophily. http://www.nytimes.com/2006/12/10/magazine/10Section2a.t-4.html, 2006. Online accessed 5 November 2016.

statista.com. statista, global social media ranking, 2016.

Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.