A Feasibility Study for Modelling Tie Strength with the Facebook API

Diogo Pereira

Supervisor: Prof. Ian Oakley, PhD
Co-Supervisor: Tasos Spiliotopoulos, MSc

Funchal – Portugal

September 2012

ABSTRACT

This thesis examines the concept of tie strength and investigates how it can be determined on the fly in the Facebook Social Network Service (SNS) by a system constructed using the standard developer API. We analyze and compare two different models: the first one is an adaptation of previous literature (Gilbert & Karahalios, 2009), the second model is built from scratch and based on a dataset obtained from an online survey. This survey took the form of a Facebook application that collected subjective ratings of the strength of 1642 ties (friendships) from 85 different participants. The new tie strength model was built based on this dataset by using a multiple regression method. We saw that the new model performed slightly better than the original adapted model, plus it had the advantage of being easier to implement. In conclusion, this thesis has shown that tie strength models capable of serving as useful friendship predictors are easily implementable in a Facebook application via standard API calls. In addition to a new tie strength model, the methodology adopted in this work permitted observation of the weights of each predictive variable used in the model, increasing the visibility of the factors that affects peoples' relationships in online social networks.

Keywords

Social Network Sites, Tie Strength, Prediction, Facebook.

RESUMO

Esta tese tem como objectivo estudar o conceito *tie strength* (força de ligação) e investiga como pode ser implementado em tempo real na rede social online Facebook a partir de um sistema construído através da sua própria API. Foram analisados e comparados dois tie strength algoritmos: o primeiro é uma adaptação da literatura passada; o segundo modelo foi construído de raiz, baseado em dados colhidos a partir de um questionário online. Esse questionário está em forma de uma aplicação de Facebook, colhendo assim um total de 1642 amizades de 85 participantes distintos. O novo modelo foi então construído baseado nesse conjunto de dados e usando para o efeito o método de múltipla regressão. Vimos que o modelo recém-criado teve um desempenho ligeiramente melhor em relação ao modelo anterior, sendo ainda mais fácil de implementar. Em conclusão, esta tese demonstra que ambos os modelos são úteis em prever amizades no Facebook e ao mesmo tempo são fáceis de incorporar numa aplicação de Facebook usando para o efeito a sua própria API. Adicionalmente a um modelo de *tie strength*, a metodologia utilizada permitiu também compreender quais as variáveis que têm mais influência na predição de *tie strength*, tornando visível quais os factores que influenciam os relacionamentos das pessoas nas redes sociais online.

Palavras-Chave

Redes Socias, Força de Ligação, Predição, Facebook.

ACKNOWLEDGMENTS

First of all, I want to thank both my supervisors: Professor Ian Oakley and Tasos Spiliotopoulos for their guidance, teachings and will to help me throughout the journey of creating this thesis. I have learned many things from them.
Moreover, I want to express my immense gratitude to my family, especially my parents, my both grandmothers and my girlfriend, for all their support and encouragement to make this thesis.
Finally, I want to say that it was a pleasure to work along my friends and colleagues from University of Madeira, which also helped me testing my application, which I am very grateful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF CHARTS

1. INTRODUCTION

1.1. Goals and Overview

This thesis aims to explore the feasibility of calculating tie strength on the Facebook social network site live, in real time and using data captured directly from the Facebook API. To achieve this, a survey in form of a facebook application was built with two goals in mind: to validate and adapt a previous published offline tie strength model to an online scenario; and to collect interpersonal data. The data collected was then used to build a refined tie strength model; this step was achieved by performing a multiple regression analysis on the data. Finally, both models were analyzed and conclusions draw as to which factors can indeed predict a friendship between two persons – tie strength prediction. This same analysis outputs a prediction equation that can be easily incorporated into a Facebook application.

1.2. Motivation and Background

The exponential rise of social network sites (SNS) in the last years was the main motivation to engage in this work. SNSs like Facebook, Twitter, LinkedIn, etc. are huge repositories of information about their users. This information can be used to analyze and study various aspects of peoples' lives (Lazer et al., 2009). Today, more than ever, many valuable data are available to researchers, originating a broad range of research studies. These studies cover many topics, such as: social capital, influence, information propagation, trust, privacy, network features, and much more. From all these topics, tie strength was chosen as the main theme of this thesis.

Tie strength notion was first introduced by Granovetter in his landmark paper: *The strength of weak ties* in 1973. Tie strength is defined as a linear combination of the amount of time, the emotional intensity, the intimacy (or mutual confiding), and the reciprocal services between two persons. It can be simply defined as the bonding level between two persons. Tie strength can be characterized in three types: strong, weak or absent. Strong ties are the people one usually trust, family, or close friends. On the other hand, weak ties are loose relationships, i.e. acquaintances.

The possibility to measure friendships makes tie strength desirable to be studied and modeled. One of the main uses for a tie strength model would be for example categorizing friends on social networks, such as in Facebook smart lists or Google plus circles. In fact Google recently acquired a company named Katango that is uniquely devoted to develop social algorithms - developing automatic friend sorters (tie strength algorithm). Such technology makes the users' job much easier by automatically those monotonous tasks.

Considering another different perspective, one can use a tie strength model to identify strong and weak ties. Emotional support is provided almost exclusively from strong ties

contributing for the happiness of an individual. At the first glance, weak ties seems to be useless, however their importance is huge, for example they are said to be more helpful In tasks such as finding a new job, or channels through which information is propagated most efficiently in a social network (Zhao, Wu, & Feng, 2011). Studying tie strength can enhance such models as well improve understanding of their operation.

To illustrate tie strength importance, below are presented a list of some practical applications, some of them are just improvements of already existent services (Gilbert & Karahalios, 2009; Xiang, Neville, & Rogati, 2010):

- **Link prediction:** improved systems for suggesting new connections;
- **Item recommendation:** recommend items that coincide with those a person is strongly related;
- **Newsfeeds:** filter Newsfeeds, prioritizing the updates by relationship strength;
- **People Search:** when searching people in a SNS, rank the results according to the relationship strength of the query sender and the discovered people;
- **Visualization:** improved visualization tools with scaling/shading links according to the estimated relationship strength;
- **Privacy Controls:** system that make privacy choices, i.e. strong ties have less restrictedly privacy options than weak ties;
- **Broadcast Information:** a system that only update strong or weak ties. If an individual wants to share information and wants that it is only accessible for his best friends, then it choose to broadcast only via strong ties. One can choose to share information only with weak ties because it propagates more efficiently (Zhao et al., 2011; Zhao, Wu, & Xu, 2010).
- **Friend Recommendation:** if strong ties A-B and A-C exist, and if B and C are aware of one another, then according to Granovetter (1973), a “psychological strain” may exist between B and C. So, a system that understands tie strength can avoid those recommendations, and make better friend recommendations.

1.3. Contribution

The Strength of Weak Ties has been cited about 19.000 times, however, very few of those works did not try to make a quantitative or continuous measure of tie strength (Petróczi, Nepusz, & Bazsó, 2006). The most known and successful paper released about this specific subject was *Predicting Tie Strength with Social Media* by E. Gilbert and K. Karahalios in 2009. In this work, the authors collected personnel data from Facebook, extracting then a huge set of predictive variables that combined, effectively predict tie strength. All the data

was captured in a lab and using the Greasemonkey Firefox extension – this means the data was collected at the client side. Only a few predictive variables and their correspondent weights (beta coefficients) were made available on the paper. This motivated a validation process of their partial model in order to check its effectiveness. This validation proved to be insufficient therefore encouraging the construction of a new tie strength model.

In this thesis, as already mentioned, the data extraction process was made on the fly through a Facebook application. The data collected is then used to build a new tie strength model. This thesis offers the following contributions:

1. **A tie strength model that can be easily implemented on Facebook by using its own API.**

2. **Validate or test the Gilbert and Karahalios' partial tie strength model.**

This work intends to offer a model capable of predicting Facebook friendships as also to be easily used for social network developers and researchers. At the same time we make an interpretation and analysis of each predictive variable used in the model.

1.4. Structure

This thesis is divided in 5 main chapters; bellow follows summary of each chapter:

- **Literature Review** – In this chapter is made a review of the main literature. It contains a detailed introduction of Social Network Analysis and Computational Social Science; the fields which this thesis is embedded. Also, it covers relevant works done with tie strength, with a special focus in Gilbert and Karahalios paper.

- **Methods** – This chapter covers mainly the study design. It contains the decisions behind its design. It is described a pilot which motivated the following work, as well the participants' demographics. Other explanations regarding the process, application, algorithm/model, and hypothesis are described in this chapter.

- **Application Architecture** – Some technical aspects are covered here. Specifically the Facebook API technology and the Software Development Kit (SDK) used. The database and the file scheme organization are also presented here.

- **Results** – This chapter contains the statistical procedures and analysis of the data previously obtained from the study. It covers a quick introduction of the main method used: multiple regression. The tie strength models and their predictive variables are also analyzed. Some other statistical procedures not directly related with the models are also covered.

- **Discussion** – In this final chapter the results achieved in the previously chapter are interpreted and discussed. We also see if the hypothesis are met or rejected.

2. LITERATURE REVIEW

Social Network Sites (SNS) have gained great popularity in the recent years, attracting the attention of academic and industry researchers therefore originating diverse disciplinary and methodological studies, which addresses a range of topics, and builds on a large body of Computer Mediated Communication (CMC) research. Along with some definitions and concepts, some previous studies relevant to the current work are presented. It will be done an introduction of SNSs, as well a brief history, as well the reasons why people use them. It will also be presented the field that is responsible for studying SNSs as also the theoretical concepts of Social Network Analysis (SNA). After reviewing the basic principles of the field it is introduced the core notion of this work – tie strength. The main and more relevant works about this concept are covered. Still in this chapter, another related concept is covered - social capital, highlighting the relation between both. Finally, the last section covers some methods of data extraction, important knowledge for the generation of the GraphML files.

2.1. Social Network Sites

The usage of SNSs has increased exponentially in the last years, its popularity is so vast that it is easily noticeable even for people that do not use Internet neither computers. To illustrate this fact, in a recent survey-based study (K. Hampton & Goulet, 2011) – about 79% of American adults stated that they used the Internet, and almost half of American adults (47%), that would be 59% of Internet users, say they use at least one SNS. It gets even more astonishing if you compare it with the percentage of SNS users in 2008, 26% of adult Americans (34% of Internet users, back then); it basically doubled the amount of users in just three years. In this section, along with a formal definition of SNSs, I will make a brief history of SNSs, and try to explain why these sites attract so many people.

2.1.1. Definition

According to Boyd and Ellison (2007) SNSs are defined as a web-based services that allow individuals to create a public or semi-public profile in a bounded system, manage a list of others users with whom they share connections, and navigate through their list of connections and those whose profile is accessible within the system. These features define the skeleton of a SNS, and the nature of these connections varies according to the site.

2.1.2. Brief History of SNSs

The first SNS that was according to the definition above emerged was SixDegrees.com which emerged in 1997. SixDegrees.com allowed users to create profiles and list their friends and introduced a new feature - friend lists which were visible to other another users. SixDegrees.com ended in 2000 however during this period other sites adopted the same features, such as AsianAvenue, BlackPlanet, MiGente, LiveJournal, CyWorld, and

LunarStorm. A new generation of SNSs started with Ryze.com launch in 2001, with the main goal of helping people to increase their professional networks. Similar to Ryze.com, LinkedIn surged in 2003 and it is by now (2011) the most powerful business service used by professionals in the web. A year before the launch of LinkedIn, one of the most promising SNS was launched – Friendster. However, after its great popularity, Friendster encountered a quick decline, mainly because of some technical and social difficulties. Its servers were bad equipped to handle the fast growth and faltered several times. Friendster also limited the profiles visibility of users who had more than four degrees way. Some users, in order to view additional profiles started to add other users for the simply expand their reach. This situation originated a large number of fake profiles (especially fake celebrities) to gain more friends and consequently a higher reach (D.M. Boyd, 2006). With the decay of Friendster, MySpace took its place as the most popular SNS, until 2007, where was overtaken by Facebook. Today, Facebook counts with a total of 800 million unique users by July 2011, followed by Twitter who has a quarter of Facebook users (200 million).

There are many SNSs available nowadays some of them will be mentioned below according to their audiences (Danah M. Boyd & Ellison, 2007):

- Business networks - e.g. LinkedIn, Visible Path, and Xing.
- Photo sharing - e.g. Flickr.
- Video Sharing - e.g. YouTube.
- Friendship - e.g. Facebook, and Friendster.
- Music – e.g. LastFM, and MySpace.
- Religion – e.g. MyChurch.

2.1.3. Why people use SNSs

Adding the fact that SNSs are free and easy to use, Sledgianowski and Kulviwat (2009) advocate that enjoyment is the most important factor which affects the behavior of SNS's users. Also, SNSs provide a vast range of features that enrich the experience of using these types of sites, for example social games and applications (Lin & Lu, 2011). In some studies (Burke, Kraut, & Marlow, 2011; Ellison, Steinfield, & Lampe, 2007) the usage of Facebook increases social capital, especially bridging social capital. The work of Lampe and Ellison (2006) made a distinction between 'social browsing' – the use of the site to develop new connections with the purpose of making an offline interaction, and 'social searching' – finding information about known offline people. They found out that the primary use of Facebook it is for 'social searching' (maintaining offline relationships), and that 'social browsing' obtained a low score among the respondents. This reinforces the idea that SNSs do not necessarily serve the purpose of making new friends, but they are used to reinforce or maintain offline connections (Ellison et al., 2007). In 2008, Joinson used a 'uses and gratification' framework to study the uses of Facebook and his findings were similar from previous works, for example, Facebook was mostly used to 'keep in touch' with offline

friends. Also, Choi (2003) found out that 85% of the study's respondents listed maintenance and reinforcement of pre-existing social networks as their main motive for using Cyworld. Most of those relationships are acquaintances, however most of the times there is some common offline element that connects the two individuals. This is one of the main differences between SNSs and the earliest forms of CMCs such as newsgroups.

2.2. Social Network Analysis

SNSs made available on the web, a large amount of information about people's lives. This data can be fairly easily collected and analyzed to posteriorly be used to understand many social aspects of a particular individual or even an organization. For instance, it is possible to learn how to improve relationships to share knowledge by analyzing and discovering patterns in online social networks. This aspect of the work is performed by Social Network Analysis (SNA). The field that employs SNA as a tool is designated by Computational Social Science which combines Computer Science with SNA.

SNA is based on the simple idea of a network, which involves a set of points linked with each other, and that awareness lead SNA to the mathematical theory of graphs, with the purpose of find a formal model to represent social networks (Scott, 1988).

The inherent necessity of human beings to socialize required an organizational structure to represent people and their relationships. This social structure constituted by individuals (or even organizations), which are designated by nodes, are connected by invisible bonds which are linked together in a mesh of connections (Wasserman & Faust, 1994). An example of social network is illustrated in Figure 1.

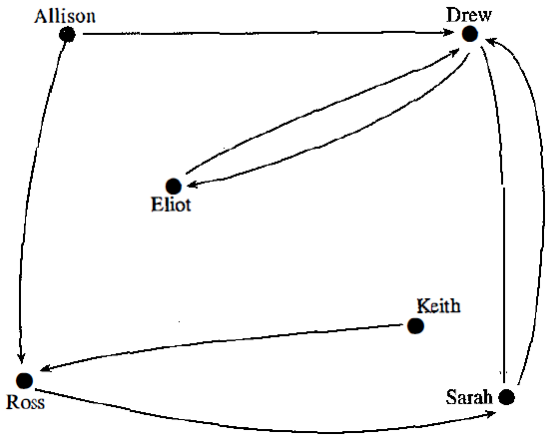

Figure 1 – Mathematical graph representing people (nodes) and their relationships (edges). This is a directed graph, as we can see there are only two reciprocal connections (Drew with Eliot, and Drew with Sarah) (Wasserman & Faust, 1994)

The connections between nodes may characterize specific types of interdependency, such as friendship, kinship, professional relationships, and common interest and so on. For example, Borgatti, Mehra, Brass, and Labianca (2009) divided dyadic relations into four

basic types - similarities, social relations, interactions, and flows. Those relations are represented in Table 1. Much of social network research can be seen as working out how these different kinds of ties affect each other. SNA tools make it easy to analyze, visualize and study the relationships between individuals, and then apply this information to strategically identify the key nodes or even improve the flow of information (I.Ehrlich et al., 2005).

Similarities			Social Relations				Interactions	Flows
Location	**Membership**	**Attribute**	**Kinship**	**Other role**	**Affective**	**Cognitive**	e.g.,	e.g.,
e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	Sex with	Information
Same spatial and temporal space	Same clubs	Same gender	Mother of	Friend of	Likes	Knows	Talked to	Beliefs
	Same events	Same attitude	Sibling of	Boss of	Hates	Knows about	Advice to	Personnel
	etc.	etc.		Student of	etc.	Sees as happy	Helped	Resources
				Competitor of		etc.	Harmed	etc.
							etc.	

Table 1 - Typology of ties studies in SNA, edges can be represented by the above 4 different categories (Borgatti et al., 2009)

2.2.1. Computational Social Science

SNA combined with the computer technology gave origin to a new paradigm – Computational Social Science (CSS). CSS may be defined as a computational facilitation of social studies and human social dynamics as well the design and use of Information and Communication Technologies (ICT) (Lazer et al., 2009). In other words it is a multi-disciplinary field which combines the usage of computer science techniques with SNA's methods. Social computing incorporates two main fields: computational and social sciences. Other fields are important as well, such as: human-computer interaction; communication; sociology; psychology; economic and anthropological (Wang, Carley, Zeng, & Mao, 2007). One of the main advantage of this field is that allows to collect and analyze large amounts of data, with relatively few efforts by the researcher, this was once impossible in past times. After the data is collected it is performed a special kind of data mining – "social mining", with the finality of retrieve and find interesting social patterns. These type of activities may be performed by "sociometers" - possible electronic devices that perform some kind of data analysis with the goal of finding patterns from how a population breaks down into groups, to which groups are most social and productive, to the personality traits of single individuals (Lazer et al., 2009). This novel science requires a distributed monitoring and permission for analyzing the data, which implies some privacy issues i.e. it requires permissions from people to analyze their private data. Since CSS is an interdisciplinary field, it has physical limitations - distance between social science departments and engineering/computer science departments.

2.2.2. Applications of Computational Social Science

According to Wang et al. (2007) there are 4 main applications areas of computational social sciences:

- Creation of Web 2.0 services and tools to support effective online communication among online social communities, e.g., blogs, wikis, social networks, RSS, collaborative filtering, and bookmarking.
- Entertainment software which focus on the development of intelligent entities that can interact with people, e.g., programs, agents, and robots.
- Business and public sector, e.g., various e-business, healthcare, economic, political, and digital government systems.
- Forecasting systems, which could be various predictive systems for planning, evaluation, and training in areas ranging from counter-terrorism to market analysis to pandemic and disaster response planning.

2.3. Tie strength

The notion of tie strength was first introduced by Granovetter in 1973, who defined it as: "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and reciprocal services which characterize the tie". Granovetter in his work differentiated two types of ties, strong and weak. Strong ties are the ones that a person really trust and often are the beloved ones, i.e., family, and close friends. On the other hand, weak ties represent acquaintances, and they often provide access to novel information.

2.3.1. Why Tie Strength is important?

Strong ties have been found to provide social support to individuals therefore increasing their mental health (Schaefer, Coyne, & Lazarus, 1981). It was also been found that they help organizational subunits surviving in time of crisis (D Krackhardt, 1988). Paradoxically those same ties are more likely to create crisis because they push organizations towards change (David Krackhardt, 1992).

Weak ties have been shown to be more helpful than strong ties when trying to get a job (Granovetter, 1973), because the information that flows from weak ties are novel to the individual, therefore contributing to more opportunities to find a new job. For the same reason, weak ties are more efficiently to propagate information in online social networks than strong ties (Zhao et al., 2011, 2010). Also, weak ties have been shown to be useful in conduit information in CMC (Burt, 2004; Petróczi et al., 2006). Duhan et al. (1997) found out that although strong ties were more likely to influence consumer's decision, weak ties were more likely than strong ties to facilitate word of mouth referral flows. Goldenberg et al. (2001) came to the same conclusion but with a different method: a cellular automata algorithm which estimated the propagation of the information of virtual individuals in a

simulated social system. Organizations whose employees are weakly tied tend to receive better performance reviews and produce more creative ideas. Also, weak ties have been shown to be useful in conduit information in CMC (Burt, 2004; Petróczi et al., 2006).

2.3.2. Tie Strength and Social Capital

Social Capital is "the actual or potential resources which are linked to a durable network of more or less institutionalized relationships of mutual acquaintance or recognition" (Bourdieu, 1986). This concept brings many positive social outcomes, for example better public health, lower crime rates or even increased psychological well-being (Adler & Kwon, 2002; Helliwell & Putnam, 2004). Despite some initial beliefs that Internet contributed to losses in social capital (Nie, 2001), many other works (Bargh & McKenna, 2004; Keith Hampton & Wellman, 2003; Kavanaugh, Carroll, Rosson, Zin, & Reese, 2005) proved otherwise, arguing that the Internet contributed to increases in social capital.

Putnam (2000) in his book makes a distinction between bridging and bonding social capital. The relation between social capital and tie strength is perceived when those two types of social capital are introduced. Bridging social capital is related to "weak ties", i.e. acquaintances or loose connections. On the other hand there is bonding social capital which corresponds to strong ties. Some researchers (Donath & Boyd, 2004; Ellison et al., 2007) argued that the usage of SNSs helps to maintain existing relationships and greatly increase the weak ties one could form and maintain. So, SNSs are a very accessible and cheap way that a person has to increase their bridging social capital. In a more recent study Burke et al. (2011) made a survey to Facebook users which contained standard scales for social capital (bonding and bridging, social communication skill, and self-esteem. The survey responses were after matched with the server logs of the participant's activity. Burke et al. reinforced the hypothesis that the usage of Facebook was associated with increases in bridging social capital.

2.3.3. How to measure Tie Strength

Many researchers have adopted tie strength as an analytic framework for studying individuals and organizations however, when compared with the number of studies that uses the tie strength notion; there is a lack of empirical studies that actually try to measure it (Mathews, White, & Long, 1998).

Marsden & Campbell (1984) performed a survey which asked respondents to recall their three closest friends along with ten characteristics of the friendship. This study permitted a clear definition of two types of variables when measuring tie strength: indicators and predictors. Indicators are actual components of tie-strength (closeness, duration and frequency, breadth of topics and mutual confiding), whereas contextual contingencies (neighborhood, affiliation, similar socio-economic status, workplace and occupation prestige) are predictors. Predictors are aspects of relationships that are related to, but not components of tie strength. In the same work, Marsden and Campbell (1984) showed that

many indicators are not free of contamination by predictors, except for closeness. In Table 2 Petróczi (2006) summarized a list of indicators of tie strength based on data collected in offline social groups.

Indicators	**References**
Frequency	Benassi et al., 1999; Blumstein & Kollock, 1988; Granovetter, 1974; Lin et al, 1981; Marsden & Campbell, 1984; Mathews et al, 1998; Mitchell, 1987, Perlman & Fehr, 1987
Intimacy/Closeness	Blumstein & Kollock, 1988; Marsden & Campbell, 1984; Mathews et al, 1998; Mitchell, 1987; Perlman & Fehr, 1987
Voluntary investment in the tie	Blumstein & Kollock, 1988; Perlman & Fehr, 1987
Advice given/received	Mathews et al, 1998
Desire for companionship	Blumstein & Kollock, 1988; Perlman & Fehr, 1987
Breadth of topics	Blumstein & Kollock, 1988; Granovetter, 1973; Marsden & Campbell, 1984; Perlman & Fehr, 1987
Duration	Blumstein & Kollock, 1988; Granovetter, 1973; Marsden & Campbell, 1984; Perlman & Fehr, 1987
Reciprocity	Blumstein & Kollock, 1988; Friedkin, 1980 Granovetter, 1973; Mathews et al, 1998; Perlman & Fehr, 1987
Provide support/emotional intensity	Blumstein & Kollock, 1988; Granovetter, 1973; Mitchell, 1987; Perlman & Fehr, 1987; Wellman, 1982; Wellman & Wortley, 1990
Mutual confiding (trust)	Granovetter, 1973; Marsden & Campbell, 1984; Mathews et al, 1998
Sociability	Mitchell, 1987

Table 2 - List of offline indicators along with their respective references (Petróczi et al., 2006)

2.3.4. Measuring Tie Strength in social media

The indicators and predictors mentioned above were collected in offline social groups. In the last decade some studies regarding tie strength on virtual communities emerged and they tried to estimate tie strength as also verify if tie strength in online communities have the same characteristics as offline communities. One of the first were Muncer et al. (2000), which simply estimated tie strength by the number of posts between two participants.

Paolillo (2001) made a different approach by analyzing the context of the messages and used informal speech as indicator of tie strength (e.g. using “u” instead of “you”). In 2003, L. Adamic obtained a social network by analyzing users’ homepages and mailing lists. Then she introduced a matchmaking algorithm to predict the similarities between users and found out that homepage links and mailing lists are poor predictors of relationships, whilst having mutual friends. Petróczi et al. (2006) conducted a survey in an online forum to analyze tie strength, each question corresponded to an indicator of tie strength, and he found out that indicators in virtual groups are similar to those in offline networks. Also, both Chun et al. (2008) and Petróczi et al. (2006) in their works concluded that reciprocity between individuals usually are indicators of tie strength. More recently Wuchty and Uzzi (2011) used self-reported human relations and email data from a typical company to investigate how email communication patterns map onto self-reported social network data. This reinforced the idea that although the e-communication lowered the cost of communication and barriers to communicate over long distances the fundamental patterns of human interaction did not changed. Still in the same work, it was tested time-resolved information on email responsiveness to determine whether a tie is a social or a professional connection. Despite the lower absolute volume along social ties, it was stated that social closeness is positively associated with response time.

2.3.5. Measuring Tie Strength in SNSs

The vast majority of the studies regarding tie strength in social media were not realized in SNSs. However there are two recent studies (Gilbert & Karahalios, 2009; Kahanda, 2009) that are particular interesting to the current work because they both were performed on Facebook and they use interaction data (wall posts, photos, etc.) to predict tie strength.

Kahanda developed a supervised learning approach to predict link strength from transactional information. He extracted the friendship, wall, photo, and group graph, and then constructed 50 features divided in 4 categories: attribute-based features (i.e., gender, relationship status, etc.); topological features; transactional features (i.e., wall postings, picture postings, and groups); and network-transactional features (same as transactional features but considers the context). He then compared his results to the results he got from the top 10 friend application (collected from the same respondents). He found out that transactional events are useful for link prediction, and that the wall and photo graph offers important information to determine tie strength.

Gilbert and Karahalios made a slightly different approach from Kahanda’s work. They developed a script containing a survey with 5 tie strength questions on the user’s Facebook page and at the same time they retrieved the results of the survey they also retrieved interaction data between the respondents. The survey’s answers had the same purpose of the top 10 friend nomination application in Kahanda’s work – proxy for the real tie strength value. Based on past literature Gilbert and Karahalios defined 7 dimensions of tie strength: intensity, intimacy, duration, reciprocal, structural, social distance, and emotional support. With these dimensions as guide they identified 74 Facebook variables

as potential predictors of tie strength. Some of these predictive variables are illustrated on table 3. The authors tried to o take advantage of Facebook's breadth while simultaneously selecting variables that could carry over to other social media. After that, they used a statistical model to determine the strength between two nodes, and compared it with the survey's results. They found out which dimensions and variables performed better predictions, for example, intimacy dimension makes the greatest contribution when measuring tie strength.

Despite some good insights to predict tie strength the above works require on supervised methods, which usually involves human annotation, e.g., top friend nomination or friendship rating (Xiang et al., 2010).

Dimensions of tie strength	**Predictive variables in Facebook**
Intensity	Wall words exchanged Participant-initiated wall posts Inbox thread depth Friend's photo comments
Intimacy	Days since last communication Wall intimacy words Inbox intimacy words Appearances together in photo
Duration	Days since first communication
Reciprocity	Links exchanged by wall post Applications in common
Structural	Number of mutual friends Groups in common
Emotional Support	Wall and inbox positive emotion words Wall and inbox negative emotion words
Social Distance	Age difference (days) Educational difference (degrees) Political difference (scale)

Table 3 - Gilbert and Karahalios defined 7 dimensions of tie strength based on past literature (left column). On the right column are defined some of the 74 Facebook predictive variables predictors of tie strength.

2.3.6. Practical implementation of a tie strength model

On gilbert's PhD thesis (Gilbert, 2010) he also presents a practical implementation of the tie strength model described above, on Twitter (OSN which allows a microblogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets"). In order to understand the capabilities and limitations of the tie strength model Gilbert developed a Twitter application, We Meddle, using the own Twitter API. With such application it was possible to understand how a computational tie strength

model generalizes to another SNS. Also it was able to see if the model can attack the collapsed context problem, i.e. instead of relying on time as the central design axis it would put relationships as the center of social design.

It was necessary to adapt all the tie strength Facebook predictive variables to Twitter's predictive variables. Most variables were drawn directly from the Facebook experiment, with the exception of a few, for instance difference in education variable, which is absent on Twitter. The nonexistence of this variable would complete exclude the Social Distance dimension from the model. To avoid this, it was added a new variable: "fame differential" – log of the difference in follower counts.

We Meddle is a web application that creates lists for the Twitter's user. These lists cover strong ties, weak ties and 6 more communities' lists (e.g. university, golf club, etc.). This research focus in the strong and weak ties lists. The other six lists are purely for functionalities purposes. The user has the choice to correct the lists if he feels that their lists are not fully customized

In general We Meddle experiment was successful where only 1,105 out of 14,075 potential ties were subjected to corrections by the user. Corrections in this context, is seen as changing a person of list to other list, for example: from the circle of strong ties to the circle of weak ties. The ratio it is quite good, however it must be accounted that the We Meddle usage were not supervised in the lab, so participants that did not made any corrections could not perceive that it was possible to correct lists or did not feel that it was not worth the effort to make the extra clicks.

2.4. Data Extraction

This last section covers some aspects of the collecting data process. This topic is separated from SNA however it is important to have some notions of data extraction concepts for a better and full understanding of a SNA study, as also having the foundations in how to generate a GraphML file. Those files are important to study social network characteristics and might offer some insights when studying tie strength.

Gjoka et al. (2010) distinguished two categories of crawling techniques: graph traversal techniques and random walks. In graph traversal techniques each node is visited one time, if the process is run until completion. These methods vary in the order in which they visit the nodes, the most known is the Breadth-Search-First (BFS) which are used in the work of (Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007; Ye, Lang, & Wu, 2010). Random walks are different because they allow node re-visiting. Gjoka itself have sampled and analyzed the Facebook friendship graph with different visiting algorithms (BFS, RandomWalk and Metropolis-Hastings RandomWalks).

Independently of the methodology used, the process of data collection can be performed as it follows (Catanese, De Meo, Ferrara, Fiumara, & Provetti, 2011):

1. Preparation for the execution of the agent.
2. Starting the process of data extraction.
3. The crawler execution extracts the friend lists, cyclically.
4. Raw data are collected until the extraction process is over.
5. Data cleaning and de-duplication of information.
6. Data structured in appropriate format (e.g. GraphML).

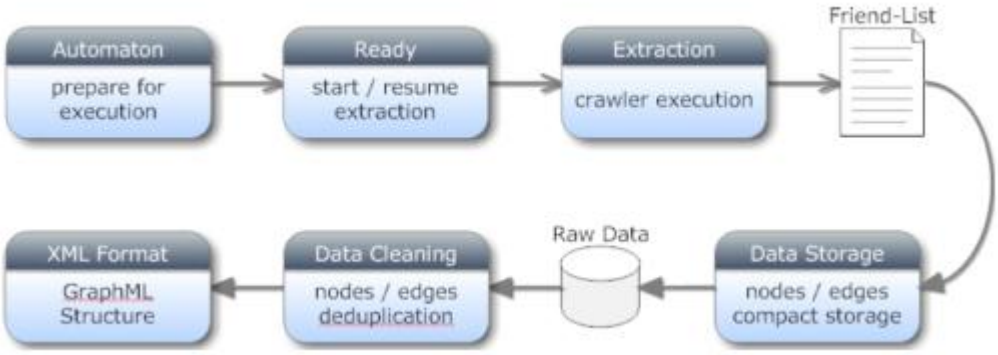

Figure 2 - Data collection process adopted in Catanese et al., 2011

3. METHODS

3.1. Study Design

Tie strength survey is an online survey in a form of a facebook application. Only Facebook users can perform the survey. The facebook application developed allowed participants to rate as much friendships they wanted to. The goal of the survey is to make participants answer several questions and at the same time capturing their interaction history about their friendships. By doing this I gathered a comprehensive dataset and I could perform a multiple regression analysis for better understanding which online social network's factors contributes to a friendship.

3.2. Theoretical Approach

The main research question addressed on this thesis is: it is possible to model tie strength on Social Network Sites, in this particular case, Facebook by using its own API? Much of the literature analysis implies that tie strength can be indeed modeled on online environments as is on offline communities. The work of Gilbert and Karahalios, 2009 approaches this subject eloquently by capturing a huge number of facebook predictive variables, using for this end a script on the user's browser. With the data gathered, Gilbert and Karahalios performed a multiple regression analysis producing an accomplished equation, which could predict tie strength with great accuracy.

This study, in the likeness of Gilbert and Karahalios work, has the goal to produce an equation capable to predict tie strength between two particular users by adopting a multiple regression equation by using the Facebook API. This method also permits determining the weight of the different variables when estimating tie strength.

The main difference and emphasis are situated in the technical implementation as well a different approach. In the previous research (Gilbert & Karahalios, 2009) of tie strength it was a used a Firefox extension named Greasemonkey which allows JavaScript scripts inject on the Firefox browser making changes in the webpage that it is displayed to the user. Another approach can be used, for example the Facebook's download your data feature (Panovich, Miller, & Karger, 2012). This last method lacks the ability to interact directly to the user, a step necessary to make the survey and collect the answers on the fly.

In this study it was used the Facebook API capabilities to perform the work. By using the Facebook API it was possible to develop a survey in form of a Facebook application and at the same time capture the data. This methodology did not require the presence of the participants in the lab as did past work.

At the same time the application collected the data necessary for the purpose described above and it also tested a tie strength algorithm on the fly. This tie strength algorithm is

based on the beta coefficients of the predictive variables obtained Gilbert and Karahalios' work. A similar algorithm based in those beta coefficients were used by Panovich, Miller and Karger, 2012.

3.2.1. Pilots Test

Two pilots were addressed to test the tie strength algorithm based on the beta coefficients got by Gilbert and Karahalios in 2009. The secondary goal of the pilots was to detect any technical bugs related with the application itself in order to solve it in time to the main study. The survey presented in the pilots had the following five tie strength questions:

- *How strong is your relationship with this person?*
- *How would you feel asking this friend to loan you $100 or more?*
- *How helpful will this person be if you were looking for a job?*
- *How upset will you be if this person unfriended you?*
- *If you left Facebook for another social site, how important will it be to bring this person with you?*

Both pilots were performed on 10 participants and it was asked to each one to rate no less and no more than 20 friendships, performing a total of 200 friendships for each pilot. The first pilot served essentially to detect and solve technical errors. Errors were found in the storing process consequently compromising some of the data collected. So, the first pilot served essentially to correct some bugs as also some changes in the application design.

The second pilot conducted was free from errors and it was aimed to make a validation of Gilbert and Karahalios partial algorithm by making a correlation between the respondent's answers and the tie strength algorithm scores.

	Mean	**Std. Deviation**	**N**
model	.2241	.23748	200
average	.3366	.25284	200
Q1.HowStrong	.3939	.30396	200
Q2.Loan100	.2102	.29949	200
Q3.HelpfulJob	.3668	.29339	200
Q4.Upset	.3541	.28554	200
Q5.LeftFB	.3579	.26921	200

Table 4 - Descriptive statistics for the tie strength model, the five questions and the mean of the five tie strength questions

The real tie strength score is indicated by the participant's answers. It was assigned by the five questions' mean as also by each question individually.

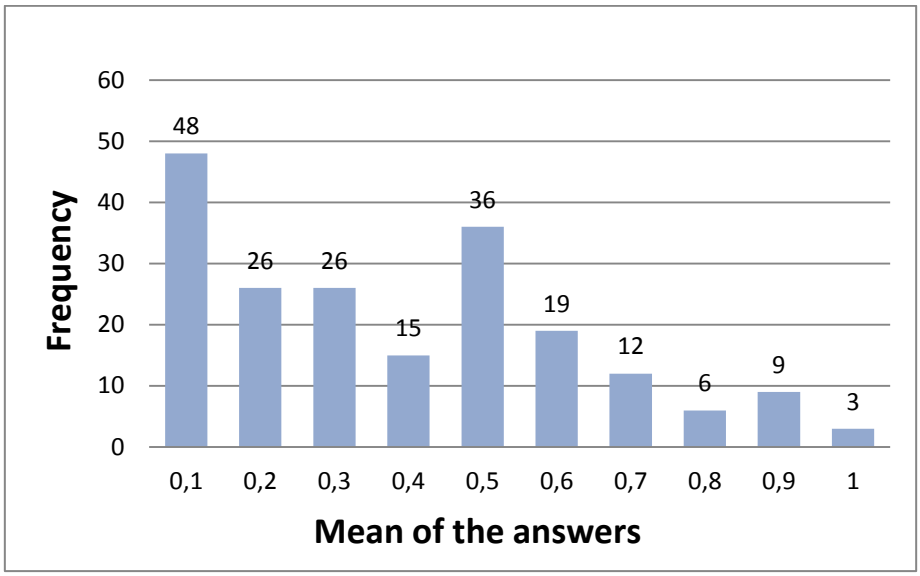

Chart 1 - Distribution of the mean of the five tie strength questions

In Chart 1 it is represented the mean of all the tie strength questions. As expected, low tie strength values are more frequent than high ones.

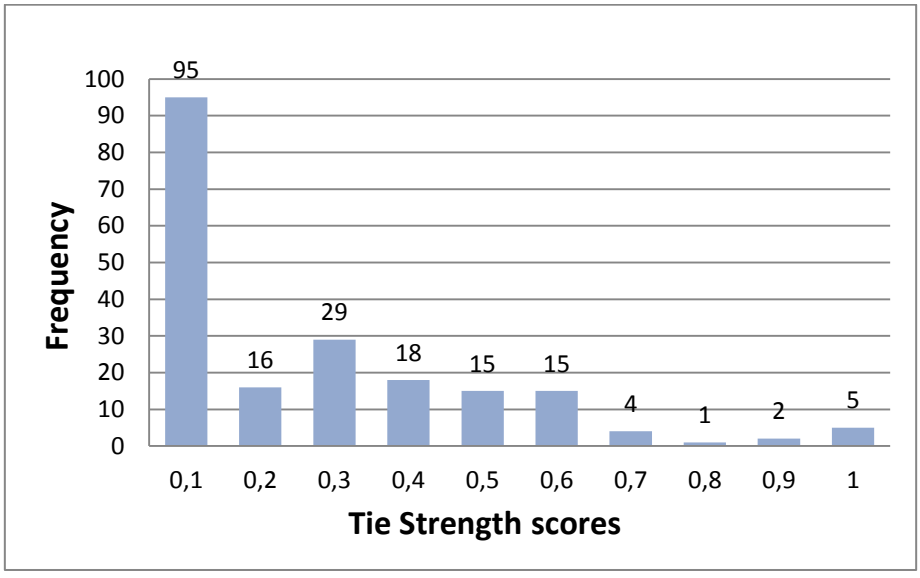

Chart 2 - Distribution of the tie strength algorithm for the corresponding 200 friendships rated

Most of the scores attributed were between 0 and 0.1, mostly due the fact that a lot of people do not interact at all (acquaintances) on Facebook.

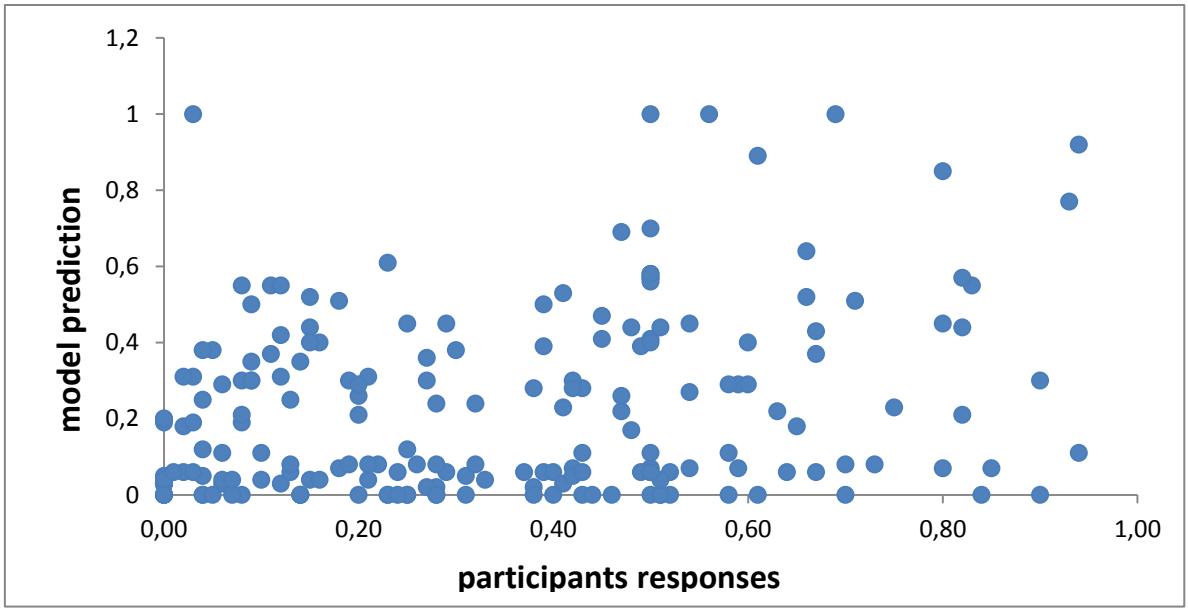

Chart 3 - Scatter plot for the respondent's answers and tie strength values

		model	average	Q1.HowStrong	Q2.Loan100	Q3.HelpfulJob	Q4.Upset	Q5.LeftFB
model	Pearson Correlation	1	.273**	.270**	.269**	.148*	.261**	.241**
	Sig. (2-tailed)		.000	.000	.000	.037	.000	.001
average	Pearson Correlation	.273**	1	.931**	.826**	.763**	.934**	.899**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
Q1.HowStrong	Pearson Correlation	.270**	.931**	1	.761**	.645**	.854**	.786**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
Q2.Loan100	Pearson Correlation	.269**	.826**	.761**	1	.413**	.778**	.630**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
Q3.HelpfulJob	Pearson Correlation	.148*	.763**	.645**	.413**	1	.594**	.672**
	Sig. (2-tailed)	.037	.000	.000	.000		.000	.000
Q4.Upset	Pearson Correlation	.261**	.934**	.854**	.778**	.594**	1	.844**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
Q5.LeftFB	Pearson Correlation	.241**	.899**	.786**	.630**	.672**	.844**	1
	Sig. (2-tailed)	.001	.000	.000	.000	.000	.000	

Table 5 - Correlation between the model and the five participant's answers, including their mean

The scatter plot shows a positive correlation yet low between the model's prediction and the survey's answers. In Table 5 is possible to see that the highest correlation occurs when using the mean of the 5 questions. However the difference is not significant when compared with others questions individually. With the exception of the question: "*How helpful will this person be if you were looking for a job?*" which presents a very low correlation. This low correlation is expected, since past literature supports that weak ties

are better in finding a new job, so getting a job is almost unrelated to how strong a relationship is between two persons, but more an indicator of the helpfulness of the friend when searching for a job.

These results supported the decision to keep just the first question – "*How strong is your relationship with this person*?" in a future survey as single indicator of the real tie strength value. Excluding all the remaining 4 questions from the survey due its redundancy and replacing it with other questions. Most important, these results motivated a quest for a better tie strength model because Gilbert and Karahalios made available only a small set of beta coefficients and these same coefficients were associated to predictive variables whose extraction was very difficult to get or because of some technical limitations of the Facebook API.

3.3. Participants

This survey gathered responses of 85 participants performing a total of 1642 friendships rated. Not all participants rated friendships equally, i.e. not all participants answered the questions for 20 friends, some respondents rated more than others. This was an expected situation because these participants were not lab controlled. This means that most participants did the survey where and when they wanted to do. In Chart 4 we can see the distribution of participants that rated 20, less than 20, and more than 20 friendships.

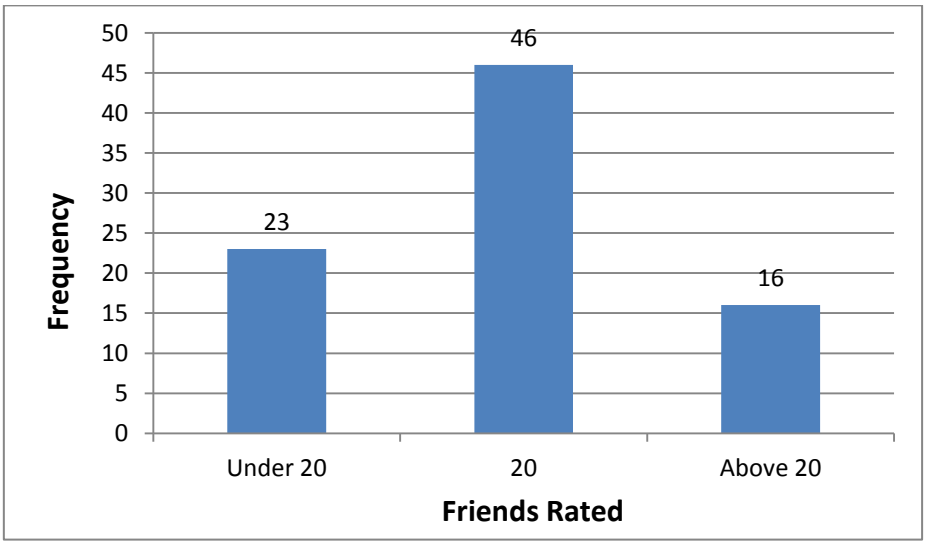

Chart 4 - Friends rating distribution

Most participants rated 20 friendships because of the way the survey was designed. Some did more than 20 because they wanted to contribute more or wanted to unlock more funny messages (funny messages appeared in the end of the 20 first friends and then for each 10 increment). Some did less probably due to a lack of interest in completing it.

A great part of the participants were recruited through Facebook itself. Others participants were got through submitting the survey to research websites. Of these 85 participants,

40% were females. The mean age was 27 years old, with a median of 26, ranging between 12 to 60 years old. The participants ages falls slightly low from the mean of 38 years obtained in 2010 by KN Hampton and Goulet, 2011.

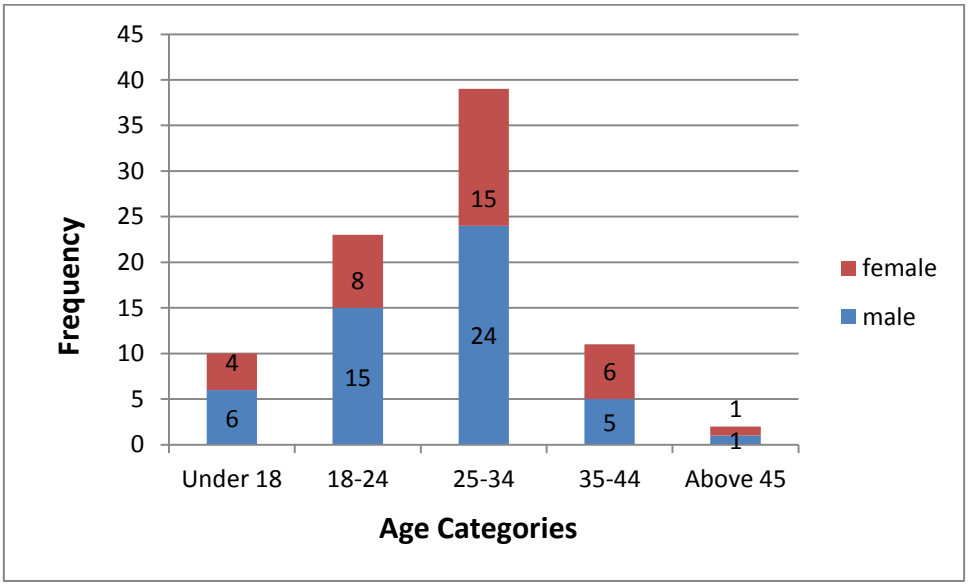

Chart 5 - Participant's age and gender distribution

The mean number of friends of this dataset is 360.1 friends, with a median of 324 ranging between 28 to 872 friends. These statistics slightly exceeds the mean of 190 and 229 friends acquired by KN Hampton and Goulet, 2011; Ugander, Karrer, Backstrom, and Marlow, 2011 respectively. In terms of number of friends, the results obtained could have been larger since there was a restriction of a minimum of 20 and a maximum of 1000 friends in order to participate in the survey. The survey's design itself implies a minimum of 20 friends. The 1000 friends limit was imposed by two reasons: technical limitations – for users with a very large number of friends (more than 1000) the Facebook API would cut off connection, probably due very high traffic of data; and for design issues – with an increase of the number of friends comes a low probability of capturing relevant friendships since the study's design suggests that participants would only rate 20 friends randomly selected from their entire list of friends.

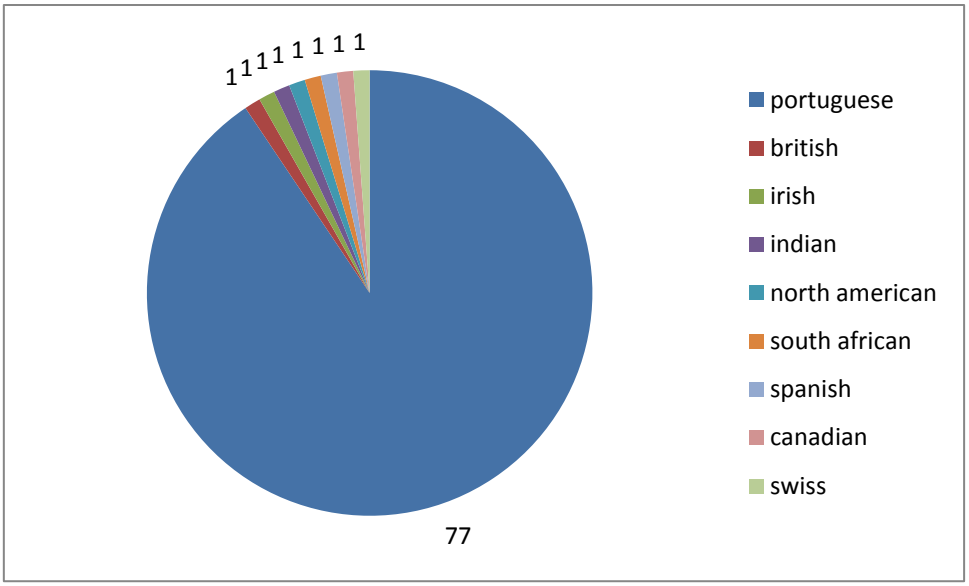

Chart 6 - Participants' nationalities

Most participants were Portuguese (77), others eight participants were from different nationalities (Chart 6). The lack of non-Portuguese participants is explained due the fact that the survey was made available on online websites very late and to worse things both websites posted the study after several days have been passed. The survey was available in English and Portuguese.

It was asked the participants to enter data about their facebook usage by asking how many hours they spent on Facebook on average per week. The results find a mean of 13.9 hours a week and a median of nine hours. According to Burbary, 2011 the typical Facebook user spent on average 15 hours per month on Facebook, so we can assume the participants used Facebook on a daily basis.

3.4. Measures

3.4.1. Gilbert and Karahalios model's predictive variables

The tie strength algorithm (Gilbert and Karahalios' version) used is calculated according to seven facebook variables. These variables are drawn directly from Gilbert's paper. It is taken the beta coefficients for each variable previously calculated using a multiple regression method.

Some of the variables could be easily implemented, i.e. without any extensive data computation. For example, the participant's friends initiated wall posts variable can be easily extracted from the JSON object by making a simple cycle to get all posts made by one particular friend ID and so on. On the other hand, some other variables took more complex steps in order to be collected. The general procedure to implement each variable is described on Table 6.

Predictive Variables	Procedure
Days since last communication	This variable was got by comparing all interactions between two particular users in order to find the most recent interaction.
Days since first communication	Ideally would be the date that two users became friends. Since this information is only partially available on facebook, it is attributed the most recent interaction between two users using a similar process like described above.
Wall words exchanged	Counting the number of words exchanged in the wall of the participant, as well the words that the participant did in the friends' wall.
Educational difference	This variable was get by analyzing the friends' educational history, a function to identify keywords that would correspond to the degree of education, next the difference was calculated.
Participant-initiated wall posts	It is simply the number of wall posts made by one particular friend on the participant's wall.
Inbox positive emotion words	Number of inbox messages words that counts as positive words. These words are compared with words of a small dictionary – short dictionary based in Bradley and Lang, 1999 .
Wall intimacy words	Number of wall words that are considered intimacy words. These words are compared with words of a small dictionary– short dictionary based in Nielsen, 2011.

Table 6 - Predictive variables and how they were retrieved

The educational difference variable was not available for some participants because they did not enter this type of information on Facebook. LIWC (Linguistic Inquiry and Word Count) was used by both Gilbert and Karahalios, 2009 and Panovich et al., 2012 to compute the inbox positive emotion words and also the wall intimacy words. For this thesis the use of LIWC software was not possible because it is not an online service, so buying it would not help with a live Facebook application. The alternative was to

summarize a set of positive emotion words and intimacy words from the papers Bradley and Lang, 1999 and Nielsen, 2011, respectively.

One must be aware of the limitations of adapting Gilbert's model; on his paper and thesis he only made available the top 15 predictive variables with the highest beta coefficients. In his experiment it were used a total of 67 individual predictive variables; some pairwise interactions variables between the seven dimensions; and some network structure variables. All these variables were not considered in the model for two simple reasons: only 15 variables had their beta coefficients published; most of these variables, especially the pairwise interactions and the network structure variables are very difficult to get and even not possible to collect using the Facebook API.

The existent literature suggests seven tie strength dimensions; five of these seven dimensions are represented on Table 6, which are: intensity (wall words exchanged and participants initiated posts); intimacy (wall intimacy words and days since last communication); duration (days since first communication); emotional (inbox positive words); and social distance (educational difference). The structural and reciprocal dimensions were not included in the algorithm.

3.4.2. Predictive variables for the new model

For the new model it was collected 13 different predictive variables from 3 different dimensions: intimacy, intensity and structural. These variables were not necessarily selected from a huge set of variables, but from the existent data available through the Facebook API.

Predictive Variable	**Dimension**
Wall posts exchanged	Intensity
Mutual friends	Structural
Common groups	Structural
Common events	Intimacy
Inbox messages exchanged	Intensity
Degree of relatedness	Intimacy
Comments on participant's wall	Intensity
Comments on participant's photos	Intensity
Comments on photos where the participant is tagged	Intensity
Likes on participant's wall	Intensity
Likes on participant's photos	Intensity
Likes on photos where the participant is tagged	Intensity
Appearances together in photo	Intimacy

Table 7 - Facebook predictive variables used to build the new model

The predictive variables chosen for the new model do not match the variables used in Gilbert's model; meaning that the variables used in the new model. Table 7 lists these variables as well their correspondent dimension.

From this set of variables it is important to make some observations. The common events variable considers only the events that the participant is taking part i.e. the participant is "attending"; the following statuses are discarded: "unsure", "declined", and "not replied". The inbox messages exchanged variable is the only variable that can be annulled by the participant. This variable as the name implies needs the inbox messages permission, and for obvious reasons some participants did not want to concede it.

As shown on Table 8, degree of relatedness is divided in four different levels. For each level is attributed a number. For example, for the level one is attributed the number one, level two the number two, and so on. The higher levels correspond to a higher degree of relatedness. Zero is the lowest level, which indicates a non-existing familiar relationship between the participant and the friend.

Level	**Degree of relatedness**
Zero	Not defined
One	cousin; niece; aunt; uncle; stepson; stepdaughter; stepsister; stepbrother; stepmother; stepfather; sister-in-law; brother-in-law; mother-in-law; daughter-in-law; son-in-law; father-in-law;
Two	brother; sister; grandmother; granddaughter; grandfather; grandson;
Three	mother; father; wife; husband;

Table 8 - Levels and correspondent degree of relatedness

3.5. Materials

3.5.1. The Survey

As soon the user read the description of the study in the first page (Figure 3), and presses the start survey button the application presents a loading page with a please wait message. After that, the application starts collecting data and at the same time executes the Gilbert's partial algorithm for all the participant's facebook friends. After finishing these two steps, the user must enter some demographic data (age, nationality, gender and average time spent on Facebook per week). Next the user is presented with the survey itself, in which one of their friends is randomly selected, for each friend it is displayed their basic information, eight questions and the next button. The eight questions are in the form of horizontal sliders and the handle must be moved to the desired value. When the next button is submitted the data is saved to a database and the next friend appears. This process is repeated for 20 randomly selected friends. When the user finishes rating these

20 friends a page is shown with all the friends rated so far and for each one a funny message. The funny message is a way to captivate the participant's interest and it is not randomly attributed, instead it considers the tie strength algorithm scores as also the respondent's answers. At the bottom of this page there are three buttons: a button to rate ten more friends (same process as 20 friends); a button to see their 10 best friends according to the algorithm, as well the rankings of their survey's responses; and an exit button. A user can rate as many friends they want to. It starts with 20 and then with increments of 10.

In an ideally scenario each participant would answer the eight questions not only for their 20 friends, but for all their friends. Unfortunately these conditions are impractical to be met. Considering the average number of friends that a typical Facebook user has – 190 friends in 2011 according to Ugander, Karrer, Backstrom, and Marlow, 2011, as well the natural tendency this number has to growth. It would be saturating to the participants to answer the questions for all their friends even more unreasonable if we want that this application is to be willingly installed by the participant. So, the participants could rate 20 friends at first and then rate more 10 if they want to and so on. At the end of each set of ratings the user is prompt with some funny messages as reward.

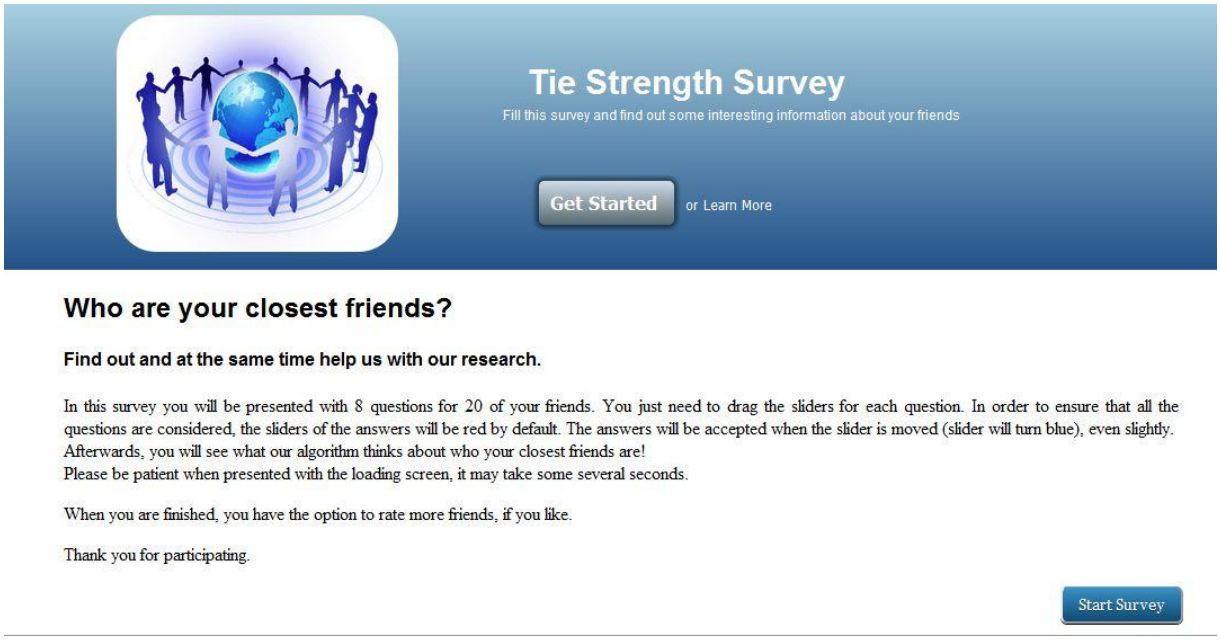

Figure 3 – Survey's start page with a short description of the study.

In order to make the survey faster, when the user pressed the next button a parallel process was executed in the background. This process is responsible for some of the data collection, generation of the GraphML files and getting the mutual friends for each friend. These processes are time consuming, so doing this way saves time and intends to not frustrate participants by waiting too long. For each participant is generated a GraphML file of their entire network – relationships between their mutual friends.

GraphML files and getting mutual friends are intrinsically connected since those files are basically the relationships between mutual friends. Obtaining the mutual friends between the participant and a friend is time consuming because one must do an API call for each friend (note that the mutual friends gathered are for all facebook friends, not only for the friend that is being rated at the time).

The decision of making these questions in the format of a horizontal slider form, as shown in Figure 4, were based in the assumption that tie strength can in fact be continuous (Granovetter, 1973). This assumption was not proven in the literature, as result the decision of how many discrete tie strength levels exist is left to be specified. Choosing a continuum scale avoids this problem. The answers as well the tie strength score are modeled as continuous 0-1 scale, where 0 is the weakest and 1 the strongest.

Figure 4 - Survey's main page displaying 8 questions for one random friend

When the page containing the questions first appears the sliders' handles are colored red and will stay this way until they are moved (Figure 4), when they are moved it turns blue. The user is prompted with a warning message when trying to press the next button while there are at least one red handle. By this design the user is forced to move the sliders even if they want to leave the handle in its initial position. This decision was adopted after the pilot was made and noticing an abnormal frequency of 0.5 scores in the distribution (score attributed by leaving the handle in its default position). It was noted that participants had tendency to leave the handler in the middle biasing this way the results.

3.5.2. Tie Strength Algorithm implementation

The tie strength generation according to Gilbert's work, was performed to all the participant's friends. The purpose of this process was to compare these tie strength values with the respective survey's responses. While the loading screen is shown a PHP script containing the algorithm is executed. The script's time execution depends majority on the number of friends a participant has, as also all their Facebook information.

The first step to calculate tie strength is getting all the data necessary as input by making about 15 API calls. This number is relatively very low for the intended tie strength generation. Despite doing very few API calls each call returns a relatively large JSON object. For example I made an API call to all the stream (participant's wall) data in one single turn, this returns a large amount of data. After getting all the necessary data it is performed several cycles to correctly analyze the data. This approach is more difficult and more prominent to errors, however is much faster. The slower tactic would not be viable due to its very long execution time. In the other hand the adopted approach despite being faster as said before more error prone because one must carefully analyze each field. The main problem would be the inexistence of some fields in some profiles. For example if a person has declared relationship in Facebook it would have an extra field pointing for the ID of the loved one. If a person is single, or only said to have a relationship but not declaring who is the other person, then this extra field is inexistent. Those types of problems could be easily avoided if this information would be available in detail on Facebook Developer Site which is not.

After refining the variables the model is applied by multiplying each variable with their respective beta coefficients which were drawn directly from Gilbert and Karahalios. Those beta coefficients are standardized coefficients, so for the correct tie strength calculation the variables must be first unstandardized, and then multiply it by the respective beta coefficients.

3.6. Research Questions

This section has the goal to formalize the research questions this thesis intends to answer. Below are presented the research questions:

RQ1 – It is *possible to model Tie strength by using a live Facebook application for data collection purposes?*

Past literature focus their experiments in building tie strength models from no-live environments – by collecting data passively. The software built for this experiment is a Facebook application with two main goals: test Gilbert's partial model and collect data for the purpose of constructing a new tie strength model. This model differs from the ones of past research in the sense that the data used to build the model is collected by a live Facebook application.

RQ2 – *Which variables are more important in predicting tie strength?*

Tie strength can be modeled by a various set of components – predictive variables. This research question looks to highlight the variables that have greater impact to the model built.

RQ3 – Can the new model *built from the data collected by the Facebook API perform better than the partial Gilbert and Karahalios' model?*

This question aims to answer which model performs better: the new model or the Gilbert and Karahalios' partial model. It is said partial model instead of just model because many variables are left over as mentioned on 3.3.1 section.

4. APPLICATION ARCHITECTURE

4.1. Iframe applications

Facebook allows the creation of two distinct types of applications: Facebook for Websites and Canvas Applications. The first one consists in using the Facebook API to add features such as the like button and the share button to a website. The second ones are the actual applications which can only be used from within Facebook.

The application created for this thesis is a Canvas application. Those applications are constituted by two different parts: a container that is defined on Facebook itself and an external web application that is connected with the container, as shown in Figure 5. The canvas page is basically an ordinary web page (HTML, JavaScript, CSS, etc.) which is loaded within an iframe on Facebook. Iframe's applications will only reload the content inside the actual iframe, instead reloading the entire facebook page.

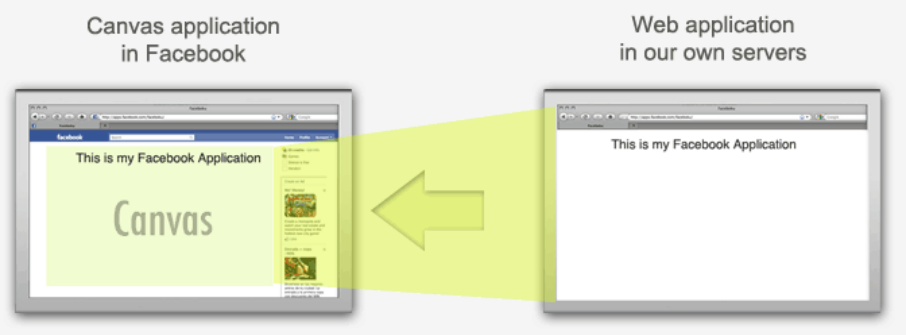

Figure 5 - A Facebook App is a Web application inside Facebook, i.e. an iframe

4.2. Security

All Facebook applications require a valid SSL certificate (was not mandatory until very recently) to work. Despite the Facebook login is secured, the application or page where the application is hosted may not be, leaving the user susceptible to attacks. For this reason the canvas URL provided must be SSL secured. This policy reinforces security by protecting the privacy of the user as well preventing attacks against the Facebook account.

4.3. Authentication and Permissions

When a user authenticates into an application this gives it the ability to determine the identity of the Facebook user, as well read and write data using the Facebook API. The protocol that serves this authentication is called OAuth 2.0. This protocol enables a third-party application (iframe) to obtain limited access to an HTTP service (Hammer & Hardt,

2012). OAuth 2.0 accomplishes this by introducing an authorization layer and separating the role of the client from the resource owner.

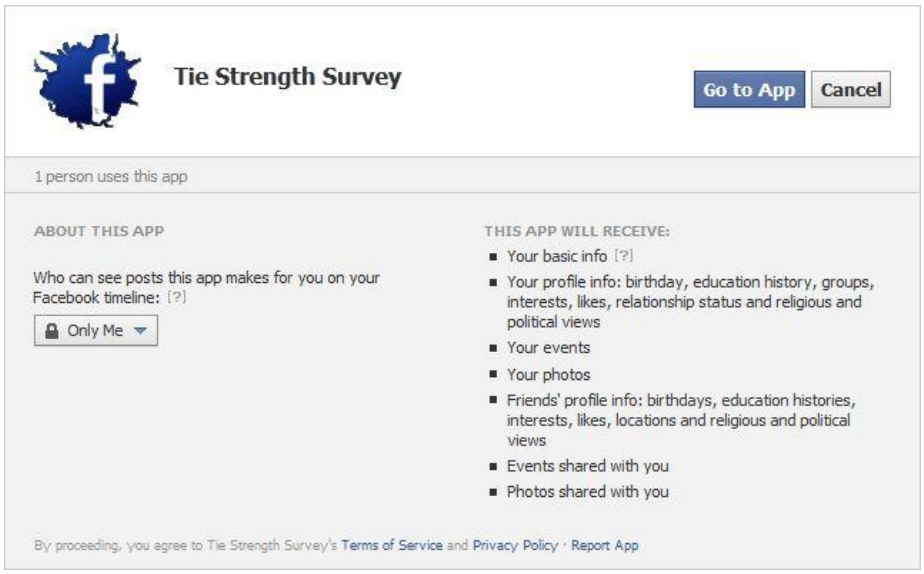

Figure 6 - Application authorization dialog

More specifically, instead of using the resource owner's credentials to access all protected resources the third-party application uses an access token – a string representing an authorization issued to the client and composed a scope, lifetime and other access attributes. The client (third-party application) now uses a different and more restricted credential to access the protected resources hosted by the resource server (Facebook).

By default, when a user authorizes an application, it only grants access to their basic information such as id, name, picture, gender, and locale. Also, some information about friends may be available, this depends which data is publicly available. The access of additional information is allowed by asking the user for extra permissions. These types of permissions are asked when the user is about to authorize the application. There are three categories of permissions: user permissions – information about the user; friends' permission – information about the user's friends (must be made public by the user's friends); and extended permissions – are presented on the second page of the dialog and can be individually discarded (Figure 7). On the other and, user and friends' permissions cannot be revoked, as is seen is Figure 6.

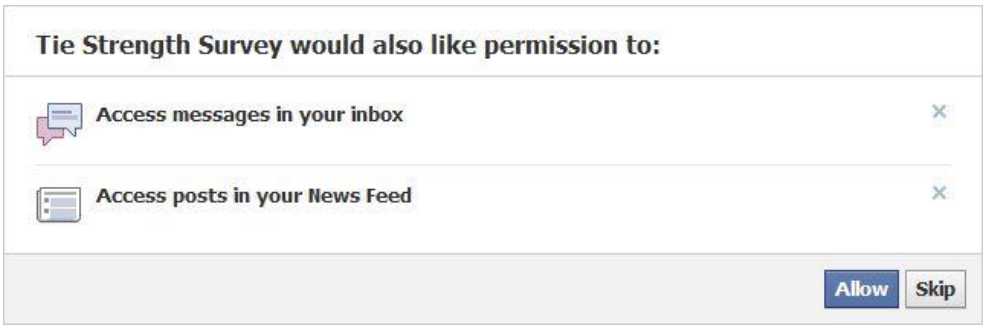

Figure 7 - Dialog box asking for extended permissions

```
$location = "". $facebook->getLoginUrl(array(
   'scope' => 'read_stream,
      read_mailbox,
      user_likes,
      user_education_history,
      user_birthday,
      user_relationships,
      user_photos,
      user_interests,
      user_groups,
      user_events,
      friends_events,
      friends_likes,
      friends_birthday,
      friends_photos,
      friends_education_history,
      friends_location,
      friends_interests',
   'redirect_uri' => 'https://apps.facebook.com/tie_strength/'
   ));
```

Figure 8 – Snippet for asking Facebook permissions

There are no distinctions when asking for different types of permissions. It is achieved by adding the permissions required to the scope parameter in *getLoginUrl* method of the PHP SDK, like shown in the code bellow. The *redirect_uri* parameter indicates the applications' URL which is redirected as soon the user grants access to the application.

4.4. Facebook SDK for PHP

The main programming language used to develop this application was PHP. This decision was made considering the existence of a Facebook SDK for PHP and also a vast documentation available on the web.

The SDK for PHP provides many server-side functionalities for accessing Facebook's server-side API calls. The integration of Facebook SDK for PHP is made by providing the ID and Secret of the Facebook application, as it shown in the following snippet (Figure 9).

```
require_once("facebook.php");

$config = array();
$config[appId] = 'YOUR_APP_ID';
$config[secret] = 'YOUR_APP_SECRET';
$config[fileUpload] = false; // optional
```

Figure 9 – Initialization of the Facebook object

As we see in the snippet above, after the installation of the PHP SDK libraries, the SDK is used to instantiate a new Facebook object, and it needs an app id and app secret. These parameters are provided in the Facebook developer page. The recent created variable *$facebook* serves as access point of all Facebook methods. Some of the most important are described below:

1. **api** is the most used method, it allows calling the Graph API or using a FQL (Facebook Query Language) query. This method was used to get all the data needed for the study from the participant. For instance, the following code (Figure 10) gets the entire participant's friends data in a JSON object.

```
$friends = $facebook->api('me/friends');
```

Figure 10 - Code to get all participant's friends

2. **getLoginUrl** method returns a URL that when clicked will redirect the user to login into Facebook to authorize the application, this only happen if the application has not been authorized before. The process behind this method it is described in the Authentication section.

3. **getUser** method returns the Facebook ID of the current logged in user.

In conjunction with PHP SDK, the Facebook JavaScript SDK provides a smooth integration between the client and server-side of the application. If a user has already authorized the

application, the JS SDK can pick up the user session which was previously saved by the PHP SDK.

```
<script type="text/javascript">
   FB.init({
   appId : 'APP_ID',
   status : true, // check login status
   cookie : true, // enable cookies   xfbml : true // parse XFBML
   });
</script>
```

Figure 11 - JavaScript SDK initialization

JavaScript for Facebook needs to be initialized in every page where it is used by providing the application ID. The JavaScript SDK provides a method which can access Facebook server side APIs. However in the scope of the application developed, all API calls were made by PHP. JavaScript was solely used for resizing the Canvas iframe and for supporting communication with the parent facebook.com page.

4.5. Graph API

The main concept of the Facebook API is the social graph. The graph represents all the objects (e.g. people, photos, events, and comments) as well the relationships (friendship, photo tags, and shared content) between these objects. Every object in the social graph has a unique ID and the properties of an object can be accessed making the following http request: *https://graph.facebook.com/ID*. Additionally, people and pages can also be accessed by their usernames.

The social graph makes possible fetching all kind of data available on Facebook. This feature is essential to this application, since it is used to capture all kind of Facebook information. However, when querying a particular object, the query returns only the public information. To get additional information it is necessary as mentioned before get the proper permissions.

4.6. Facebook Query Language (FQL)

FQL is a SQL-style language used to query the data exposed by the Graph API. Despite being a more complex method when compared with the Graph API it has more advanced features, for example it can access some information not available through the Graph API, it is more customizable and permits batching multiple queries into a single call. Data returned from a FQL query is in the JSON format. FQL is a more limited version of SQL this is easily visible because: the FORM clause can contain only a single table; subqueries cannot reference variables in the outer query's scope; and the query must contain

properties that are indexable (they are marked in the documentation). In the figure it is presented one query used in the application.

```
$fql_events = "SELECT eid, uid
        FROM event_member
        WHERE eid IN (
        SELECT eid
        FROM event_member
        WHERE uid = me()
        AND rsvp_status = 'attending')
        AND rsvp_status = 'attending'
        AND uid IN (
        SELECT uid2
        FROM friend
        WHERE uid1 = me())";
```

Figure 12 - Query that returns a JSON object with all the friends that will attend an event that the current user will participate.

4.7. Batch Requests

The typical way of querying the Graph API consists in getting data for individual objects in a single go by doing multiple individual HTTP requests. Batch requests are a performance-enhancing feature for making API calls that allow packing multiple transactions in one single http request (batch). Each request is processed as independent operation and executed in parallel. In practical terms, if 20 independent http requests are made, it will take a very long execution time because they are executed sequentially. In the other hand, using batch requests will make only one HTTP request instead of 20 since it is performed in parallel. This brings an enormous performance advantage because we get rid of 19 HTTP requests. The application built takes advantage of this feature to improve the data collection speed otherwise would be impractically to gather some types of data. At the moment this was writing, it was possible to make 50 API calls as one single batch request (increases the speed in 50 times).

For implementation purposes it was created a Facebook Batch Requests PHP Class. This class replaced the use of the PHP API method to get Facebook data. The process is divided in four simple steps:

1. Create an instance of the class:

```
$batch = new facebook_batch();
```

2. Add the queries one by one:

```
$friends_key = $batch->add('/me/friends', 'get');
$photos_key = $batch->add('/me/photos', 'get');
```

3. Execute the queries:

```
$batch->execute();
```

4. Get back the response using the keys:

```
$friends = $batch->response($friends_key);
$photos = $batch->response($photos_key);
```

4.8. Data collection

There are two processes executed when the loading screen is on: the tie strength algorithm; and data collection. Based on facebook API documentation it was possible to verify the types of personal information possible to gather. The facebook API calls necessary by the tie strength algorithm are the same used to capture the participant's data. Obviously some additionally API calls were made with the purpose of gathering information.

All the data was stored in a MySQL database which we can see in Figure 13. In the storing task it is important to state that both participants and participant's data are stored in one single table called user. The distinction between participants is made by one Boolean field. The database design adopted was built with flexibility in mind with the drawback of adding some complexity to the data analysis step. Flexibility was a concern because the data is in their original format, i.e. raw data. This way one can make different types of data analysis.

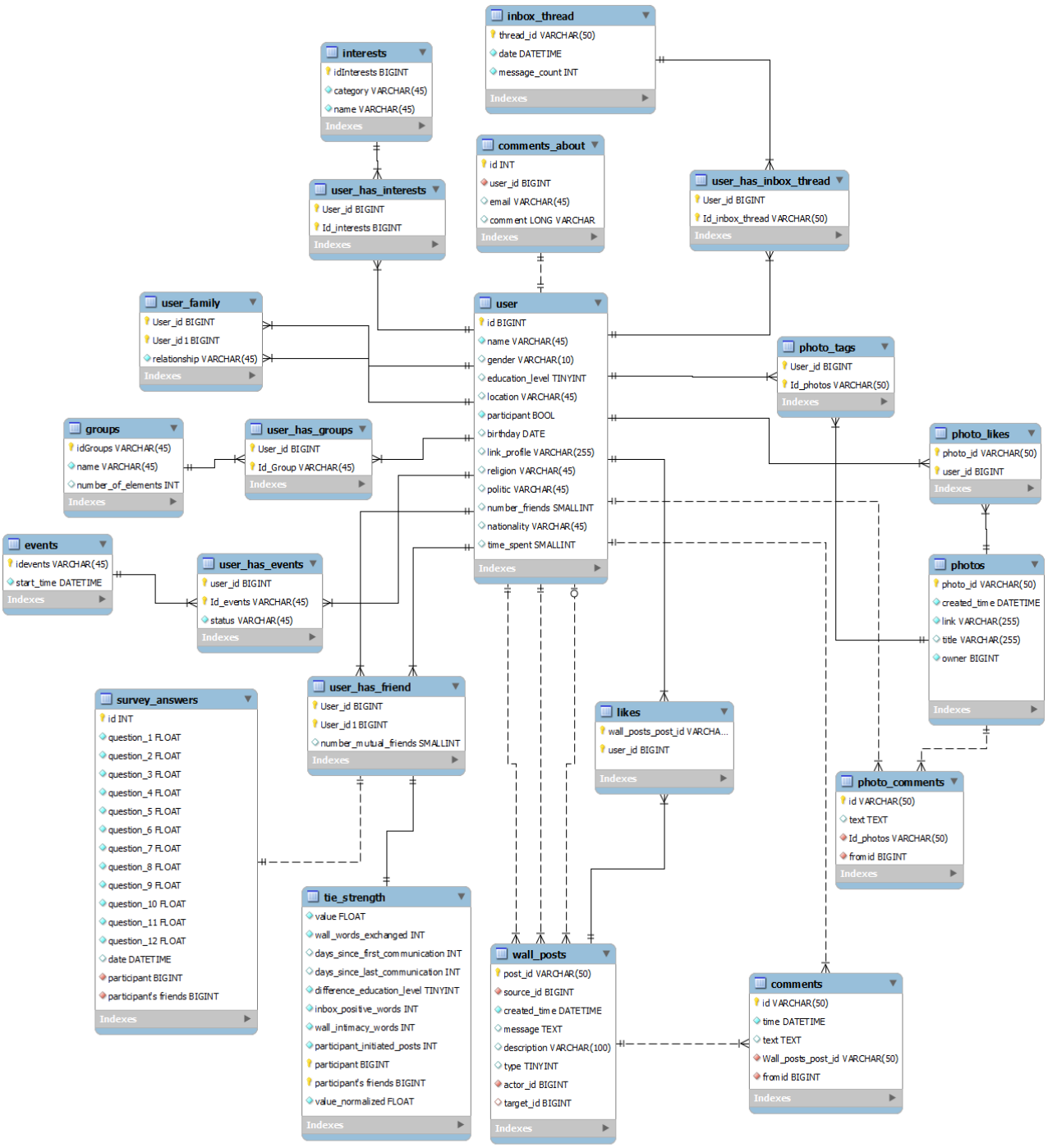

Figure 13 - EER diagram of the database used, note that table user plays a central role

4.9. File organization

Figure 14 shows the directory scheme of this application, it contains all the files responsible for the application operation. The main structure of the application is constituted with 10 PHP scripts, seen as 10 first PHP files in the figure. The files concerning the survey's introduction are index.php and start_page.php. The core file is the algorithm.php which contains the tie strength algorithm as well all the data retrieval methods and the respective insertions in the database. All the following pages, with the exception of the background.php are responsible with the survey itself. The background.php is a UNIX background process responsible for inserting the respondents' answers in the database as well the GraphML files generation.

All the sub-folders are necessary for a correct application's operation. The configuration folder deals with the database's connection and contains the facebook PHP SDK which has two PHP files and a certificate file, these files ensures the connection between the Facebook API and the webserver that hosts the application. The functions folder contains all the PHP functions built for the application – the Facebook batch file allows multiple API calls in one single HTTP request; the translation file is responsible for supporting multi-language of the survey (English and Portuguese); the funny functions file is responsible for the funny phrases generated in the end of the survey; and all the remaining functions are embedded in functions.php file.

For building the survey's horizontal slides, the loading page, alert messages and the calendar it was used jQuery. The files responsible for this are confined in the jQuery folder. Not required for the functioning of the application is the GraphML files folder. This folder serves as a repository for all the GraphML files created.

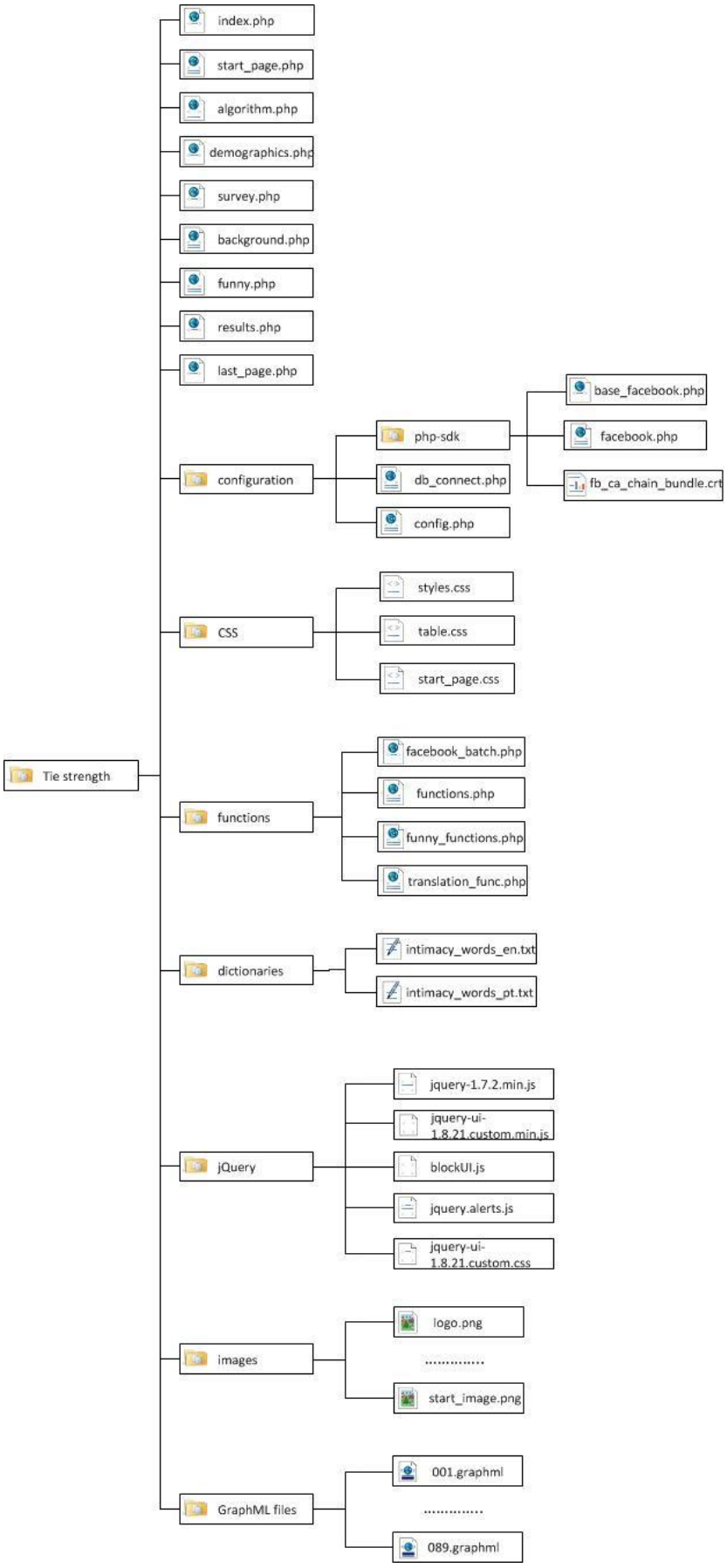

Figure 14 - Application's file scheme

5. RESULTS

5.1. Why Multiple Regression?

Multiple Regression is a statistical method for studying the relationship between a single dependent variable and one or more independent variables. It is one of the most widely used statistical techniques in the social sciences (Allison, 1998). Multiple regression is very desirable for two main reasons. The first one is for prediction purposes; by using it is possible to combine many variables to produce an optimized equation to predict the value of the dependent variable. The second reason is that it separates the effect of independent variables from the dependent variable, making it possible to examine the unique contribution of each variable. A multiple regression analysis outputs a linear predictive equation as shown in the following equation (Lane, 1998).

$$\hat{\gamma} = \alpha + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \quad (1)$$

In equation 1, $\hat{\gamma}$ is the dependent variable and the x_n are the independent variables. The letters a and b represent constant numbers which represents the intercept and slope respectively. The intercept is the line on the vertical axis that intercepts the line. The slope indicates how big a change we get in $\hat{\gamma}$ from one unit increase in x; the so called unstandardized coefficients are represented by the slopes.

By applying this concept to the tie strength model it is quickly visible that the dependent variable would represent the tie strength score and the independent variables represent the various facebook predictive variables. It is possible to build an equation that helps predicting tie strength and at the same time the contribution of each variable to the model is known.

5.2. Variables entered in the model

The dependent variable will correspond to the real tie strength value; in this case it corresponds to the scores obtained in the "*How strong is your relationship with this person?*" question. As stated before, tie strength scores are represented in an interval of 0 to 1 with two decimal places; being 0 the weakest and 1 the strongest.

Chart 7 shows the answers' distribution, as expected there is a higher concentration in the 0 – 0.1 interval. These friendships are predominantly acquaintances; people in social network sites are more willingly to accept or make friend requests than in an offline environment. The mean of the all the 1642 first question responses is 0.29 with a median of 0.20. The tie strength value (dependent variable) will be our $\hat{\gamma}$. There will be a set of 13 independent variables which would be ours x_n.

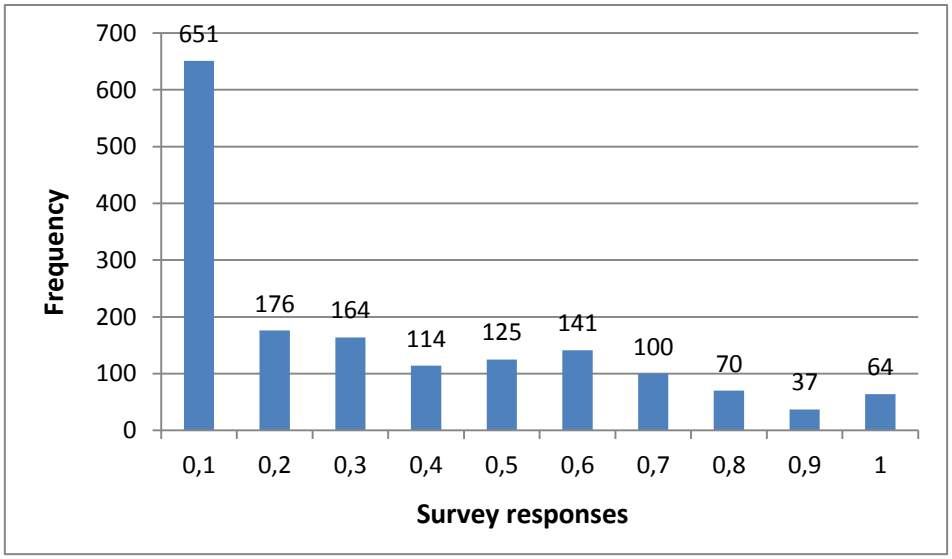

Chart 7 - Frequency of the tie strength question - "How strong is your relationship with this person?"

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation
How strong is your relationship with this person?	1642	1.00	.00	1.00	472.11	.2875	.28720
Valid N (listwise)	1642						

Table 9 - Descriptive statistics for the "How strong is your relationship with this person?" question

Table 10 shows some descriptive statistics for the 13 independent variables entered in the model. Most variables have a small range. These variables occur with a very low frequency due its own nature and due to some limitations regarding the Facebook API which limits the access of some old data. Not all participants contributed equally in terms of variables, the inbox messages authorization was denied by some participants, excluding this way the inbox messages exchanged variable for these friendships. Approximately 20 percent of the participants did not grant access to the inbox messages permission.

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
wall_posts_exchanged	1642	0	12	391	,24	,726
wall_post_comments	1642	0	14	149	,09	,539
photo_comments_of_participant	1642	0	16	87	,05	,692
comments_participant_tagged	1642	0	19	444	,27	1,466
likes_wall_posts	1642	0	142	594	,36	3,809
likes_participants_appearences_photos	1642	0	24	389	,24	1,396
likes_participants_photos	1642	0	12	78	,05	,450
mutual_friends	1627	0	326	57247	35,19	43,903
groups_in_common	1642	0	13	1102	,67	1,252
events_in_common	1642	0	3	48	,03	,204
inbox_messages_exchanged	1341	0	21900	44600	33,26	654,173
family	1642	0	3	48	,03	,210
appearences_together_photos	1642	0	23	565	,34	1,633
Valid N (listwise)	1328					

Table 10 - Descriptive statistics for the 13 predictive variables

5.3. The new model

In this section I have entered all the independent variables and the dependent variable as in SPSS then I executed a multiple regression. The *enter* method was chosen to run this regression. This method means that all independent variables enter in the method ignoring their actual contribution to the model; this way it is possible to see the impact of each variable to the equation created, or see if exist or not a relationship between tie strength and the variables.

R	R Square	Adjusted R Square	Std. Error of the Estimate
,398	,158	,150	.26425

Table 11 – Overall Model Fitt

Table 11 summarizes the completeness of the model. The letter R represents the regression (correlation) coefficient; it is the correlation coefficient between the observed values of Y and the predicted values of Y. The interpretation of R is similar to the interpretation of the correlation coefficient, the closer the value of R to one, the greater the linear relationship between the independent variables and the dependent variable. The R value is 0.398, meaning that exists a positive correlation between the tie strength score and the predictive variables. Nevertheless, the correlation coefficient has a relatively low moderate value.

On the next column there is the R-squared (coefficient of determination) value. This coefficient is the ratio of the sum of squared errors produced by the least squares equation that is being evaluated by the sum of squared errors for a least squares equation with no independent variables (just the intercept). The R-squared score of 0.158 means that using those 13 predictive variables to predict tie strength yields a 15.8 percent reduction in the prediction errors when compared with using only the mean. In other words, it can be said that those 13 predictive variables "explain" 15.8 percent of the variation in tie strength.

A R-squared of 15.8 percent may appear low; however the R-squared is only an indicator of the completeness of the regression model (Haynes, 2010). Only the p-value in the regression ANOVA (Analysis of Variance, Table 12) and the p-value of the coefficients (Table 13) should be used to determine the goodness of a regression. So, if the p-value is less than five percent then it may be possible that the regression has found a significant relationship.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	17,278	13	1,329	19,034	,000[b]
Residual	91,752	1314	,070		
Total	109,030	1327			

Table 12 - ANOVA table

In this experiment we consider a significance level of 0.05 (five percent); the null hypothesis is that the independent variables do not aid the prediction of tie strength – none of the independent variables has a relationship with the dependent variable. The column Sig. in Table 12 states a p-value of .000 which is smaller than the confidence level. This means that at least one of the coefficients values is not zero, which means that at least one independent variable has a significant relationship with the dependent variable; by rejecting the null hypothesis we conclude that the model is useful.

Looking at Table 13 it is visible the weight of each variable to the model created. Considering a significance level of five percent, the following variables are not statistically significant: mutual friends, events in common, comments in participant's photos, wall posts' likes, likes in photos where the participant is tagged, and likes of the participants' photos. The remaining variables are significant related to the dependent variable; therefore they have effect in the predicting tie strength. By watching the beta standardized coefficients column we can make a direct comparison of the independent variables. The variables wall posts exchanged, family and wall posts comments are considerable significant when estimating tie strength. The unstandardized coefficients may seem to have a very small value, but one must take in account that tie strength is modeled between 0 and 1. So for a valid comparison, one must look into the standardized coefficients.

New model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	,242	,010		24,175	,000
wall_posts_exchanged	,080	,012	,180	6,629	,000
mutual_friends	,000	,000	-,045	-1,605	,109
common_groups	,016	,006	,073	2,591	,010
common_events	-,019	,034	-,015	-,561	,575
inbox_messages_exchanged	3,018E-005	,000	,069	2,713	,007
wall_post_comments	,084	,019	,131	4,464	,000
photo_comments_of_participant	-,003	,017	-,007	-,188	,851
comments_participant_tagged	,012	,007	,060	1,741	,082
likes_wall_posts	,002	,002	,025	,908	,364
likes_participants_appearences	-,005	,007	-,026	-,692	,489
likes_participants_photos	,024	,021	,042	1,147	,252
appearences_together_photos	,017	,006	,105	2,674	,008
family	,202	,034	,155	5,880	,000

Table 13 - List of the predictive variables and their respective coefficients and significance level

By excluding the not significant variables from the equation we can come with the equation (2) which helps predicting tie strength. Only six of the 13 predictive variables are inserted in the model, the others are discarded because they are not significant.

$$\widehat{y} = 0.242 + 0.08X_1 + 0.016X_2 + 3.018^{-5}\,X_3 + 0.202X_4 + 0.084X_5 + 0.017X_6 \quad (2)$$

Where:

- X_1 is the number of wall posts exchanged between the participant and the friend.
- X_2 is the number of groups in common between the participant and the friend.
- X_3 is the inbox messages exchanged between the participant and the friend.
- X_4 is the family rank between the participant and the friend.
- X_5 is the wall posts comments of the friend on the participant's wall.
- X_6 is the number of appearances together in photos .

The 6 terms equation has a correlation coefficient (R) of 0.394 for the training dataset. To achieve the maximum correlation coefficient (0.398), all independent variables must be entered; this implies entering the not significant independent variables in the equation.

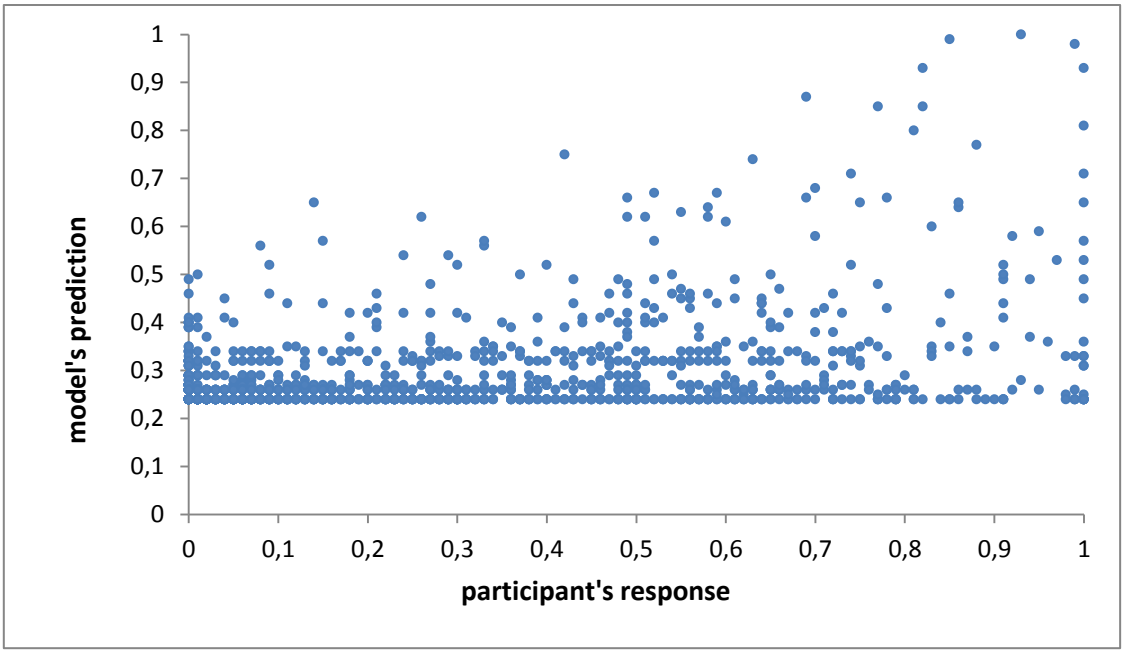

Chart 8 - Scatter plot for the respondent's answers and correspondent new model scores

The intercept for this model as seen in Equation 2 is 0.242 so the model basically does not make predictions below the 0.2 level (chart 12). However the Pearson correlation coefficient is 0.394.

		How strong is your relationship with this person?	New model
How strong is your relationship with this person?	Pearson Correlation	1	,394**
	Sig. (2-tailed)		,000
	N	1642	1341
New model	Pearson Correlation	,394**	1
	Sig. (2-tailed)	,000	
	N	1341	1341

Table 14 – Correlation between the tie strength question and the new model (six terms)

5.4. Tie strength model using Gilbert's beta coefficients

As indicated earlier on this thesis, the Facebook application tested a tie strength algorithm using some of the beta coefficients obtained on the Gilbert and Karahalios paper. It were used a total of seven predictive variables whose beta coefficients were available. These variables and their beta coefficient are listed in Table 15.

Predictive Variables	$\boldsymbol{\beta}$
Wall words exchanged	0.299
Days since first communication	0.755
Days since last communication	- 0.76
Inbox positive emotion words	0.135
Wall intimacy words	0.111
Participant initiated posts	0.146
Educational difference	-0.22

Table 15 - Predictive variables and their respective beta coefficients. The coefficients were drawn directly from Gilbert and Karahalios.

Since we are dealing with standardized coefficients and not unstandardized coefficients; the prediction equation is formulated as follows:

$$\hat{\gamma} = \beta_1 \frac{x_1 - m_1}{s_1} + \beta_2 \frac{x_2 - m_2}{s_2} + \beta_3 \frac{x_3 - m_3}{s_3} + \cdots + \beta_n \frac{x_n - m_n}{s_n} \quad (3)$$

Where:

- β is the beta coefficient of the independent variables.
- x is the independent variable.
- m is the mean of the respective independent variable.
- s is the standard deviation of the respective independent variable.

The values are then normalized between 0 and 1.

A total of 35708 friendships were analyzed using the equation above. The distribution chart is as follows. The mean of the score is 0.15 and the median is 0.06.

The high concentration at the 0.1 level indicates that most of these variables were attributed a very low value probably due a lack interaction between the participant and the friend. These distributions are similar to the Panovich et al., 2012 experiment, which used six Gilbert and Karahalios' beta coefficients to generate tie strength.

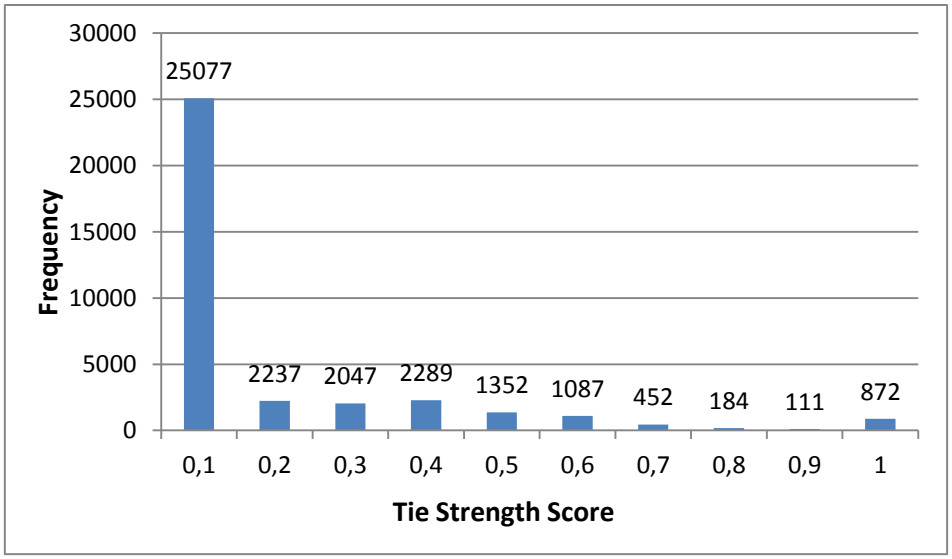

Chart 9 - Distribution of the tie strength scores for all the participant's friendships

Chart 10 illustrates the frequency distribution of the tie strength scores for only the friendships that the participant rated. It is important to keep in mind that Chart 11 contains 1341 data points, not 1642 (total number of friendships rated). These 301 friendships were removed from the scatter plot because the participants from these ratings did not allowed access to the inbox messages, removing this way the variable *inbox positive emotion words*. So, for a better comparison with the new model developed these 301 ties were left out. This exclusion in the sample improved the correlation coefficient from 0.276 to 0.311 (Table 16).

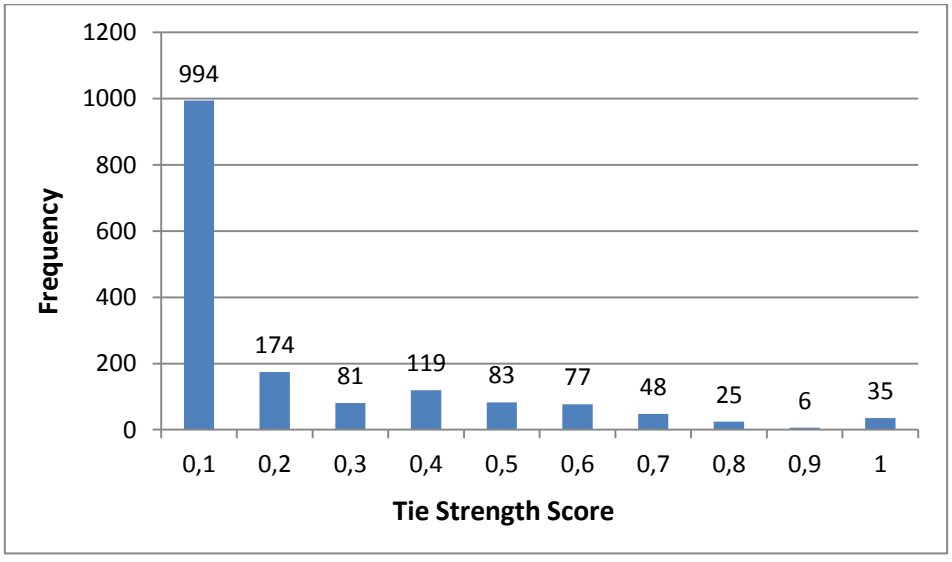

Chart 10 - Distribution of the tie strength scores for only participant's friendships that were rated

The mean of these scores is 0.19 with a median of 0.08; slightly above of the mean and median of all the friendships analyzed (chart 12).

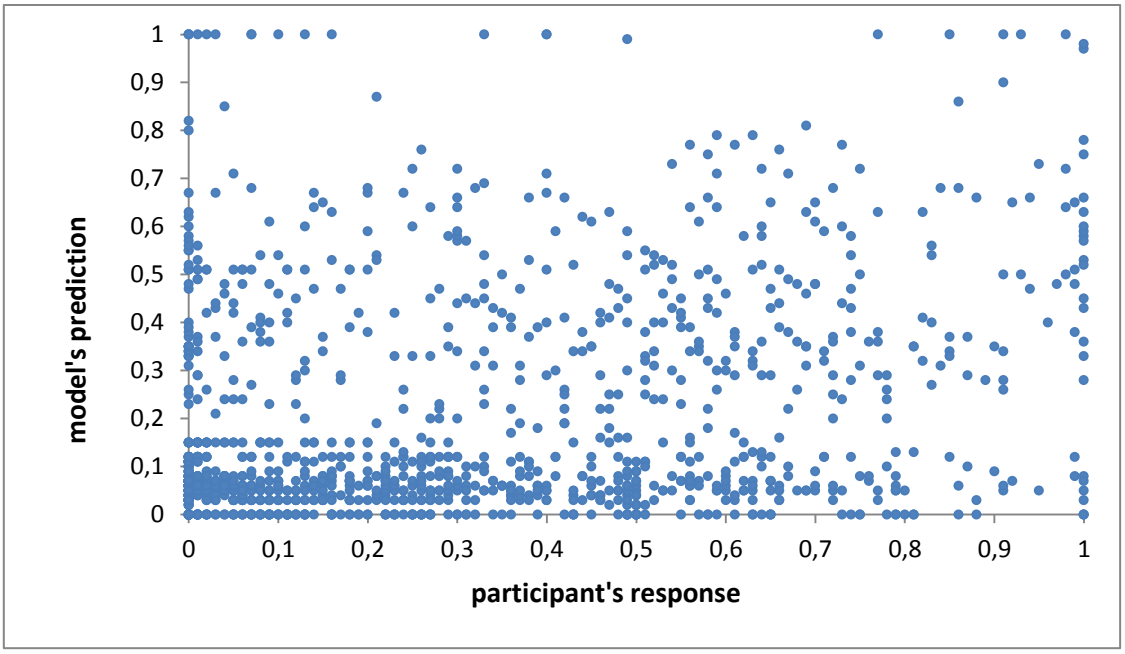

Chart 11 - Scatter plot for the respondent's answers and tie strength values

Table 16 shows a low moderate correlation of 0.311 between the tie strength algorithm and the tie strength question. The Correlation is significant at the 0.01 level (2-tailed).

		How strong is your relationship with this person?	Gilbert's
How strong is your relationship with this person?	Pearson Correlation	1	,311**
	Sig. (2-tailed)		,000
	N	1642	1341
Gilbert's	Pearson Correlation	,311**	1
	Sig. (2-tailed)	,000	
	N	1341	1341

Table 16 - Correlation between the tie strength question and the tie strength scores obtained by the algorithm

5.5. Small Validation Study

A small pilot was conducted to answer the following questions:

1. *Does the sampling method influence the tie strength model?*
2. *Could the new tie strength model perform well in another dataset?*

This validation study counted with only two participants: one with 264 friends and another with 182 friends. It was asked the participants to rate all their friends; creating a dataset of 446 ties. Rating all friends is an exhausting process, turning difficult to find participants that were willing to participate.

The mean of the survey's answers was 0.31 and a median of 0.22, which are higher than the main dataset. Chart 12 shows the distribution of the survey's answers; this dataset distribution is similar to the dataset of the 1640 friendships of the main study.

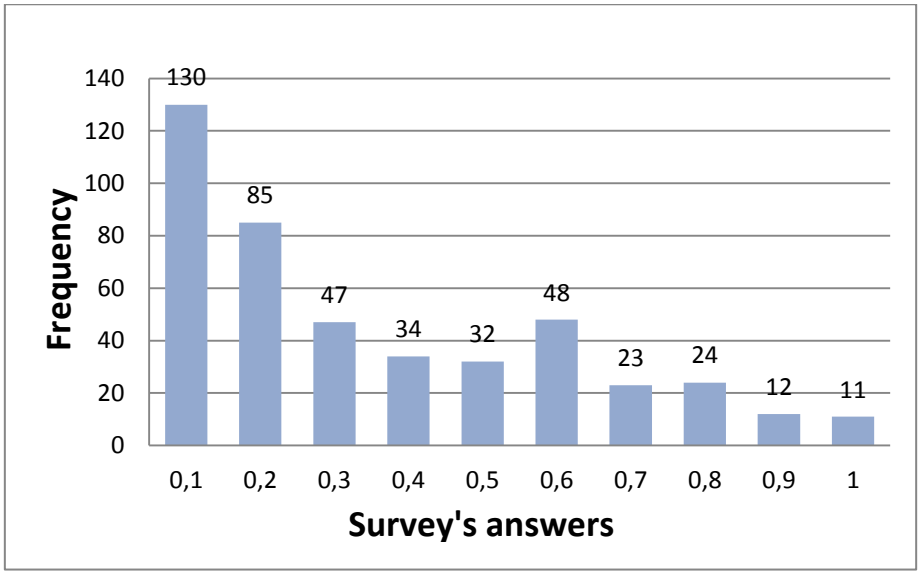

Chart 12 - Distribution of the survey's answers

The Chart 13 and Chart 14 show the scatter plots between the survey's answers and the new model, and the survey's answers and the Gilbert's model respectively. A small correlation is visible in both charts. The new model does not rate friendships below the 0.242 value; 0.242 is the intercept value, and one must keep in mind that the new model considered has only 6 predictive variables and none of them have negative coefficients.

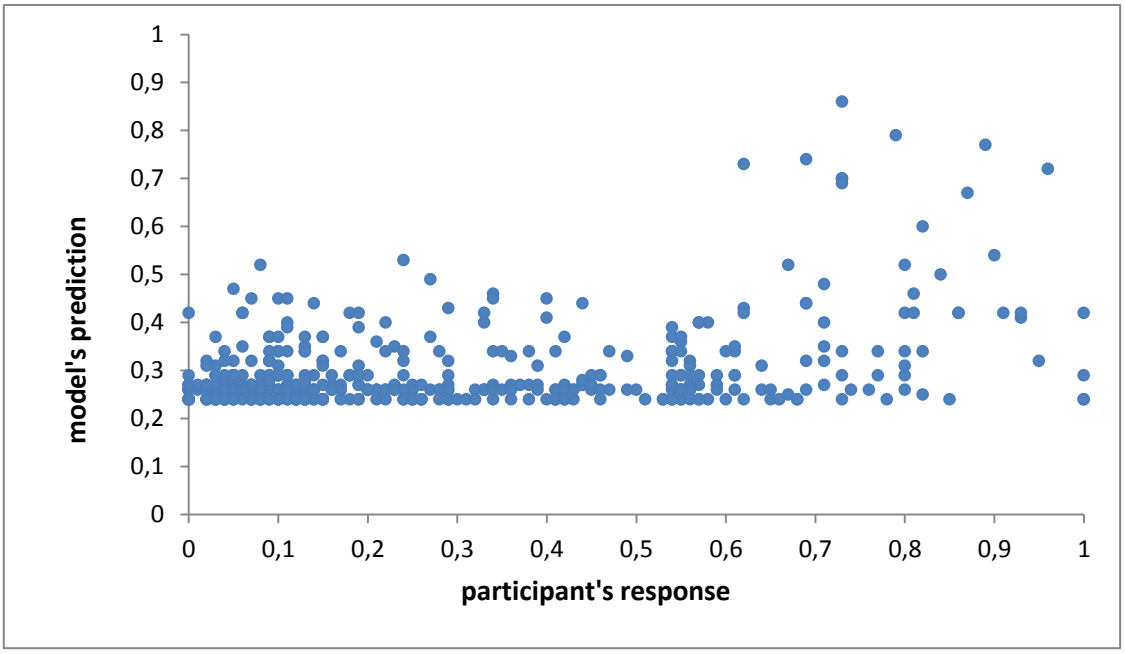

Chart 13 - Scatter plot for the survey's answers and the new model predictions

In table 13 we can see the Pearson's correlation coefficient for Chart 14 is 0.409; a low moderate value. It is even higher than the Pearson's correlation coefficient obtained when correlating the new model on the main study's dataset.

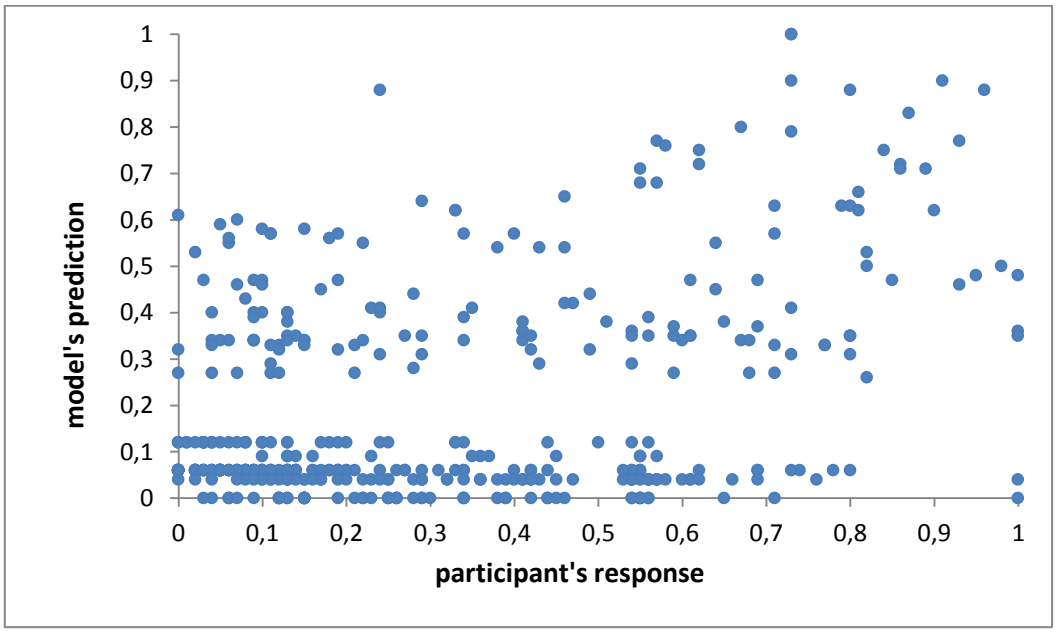

Chart 14 - Scatter plot for the survey's answers and Gilbert's partial algorithm

The Pearson's correlation coefficient for Chart 14 is 0.394; very similar to the new model correlation coefficient. Both correlations are significant at the 0.01 level (2 tailed).

		Question	New Model	Gilbert's Model
Question	Pearson Correlation	1	,409**	,394**
	Sig. (2-tailed)		,000	,000
	N	446	446	446
New Model	Pearson Correlation	,409**	1	,619**
	Sig. (2-tailed)	,000		,000
	N	446	446	446
Gilbert's Model	Pearson Correlation	,394**	,619**	1
	Sig. (2-tailed)	,000	,000	
	N	446	446	446

Table 17 - Correlations between the main question, new model predictions and Gilbert's model predictions for the new dataset

For this new dataset it was also performed a new multiple regression analysis in order to see the fitness of the model. It was used the same 13 independent variables as in the previous model.

R	R Square	Adjusted R Square	Std. Error of the Estimate
,473	,224	,204	.23415

Table 18 - Model summary**.**

The regression coefficient is relatively high and the r-squared of 0.205 states that the variables inserted in the model are responsible for 20.5% of the variation of tie strength.

5.6. Survey's Questions

Apart the survey's main question – "*How strong is your relationship with this person?*" which is the indicator of the real tie strength, it was made another seven questions, and these questions are listed on Table 19. The initial idea was to study in more detail these questions accompanied with a more specific literature review as well. However, due to lack of time to analyze the respective literature, the research of these questions was discontinued; still some basic statistics - descriptive statistics and correlations were performed on them.

From this table we can see that questions two and three, four and five, six and seven are related. Question number eight regards trust and is not related with the other questions. Questions two, four, and six are questions with intend to capture the "interest" of the participant on the friend that is showing up.

Number	Questions	Mean	Std. Deviation
1	*How strong is your relationship with this person?*	.2875	.28720
2	*How much are you looking forward to receiving updates from this person?*	.2954	.28771
3	*How much do you think this person looks forward to receiving updates from you?*	.2538	.26679
4	*Imagine this friend went on a trip. How much are you looking forward to seeing, liking, or commenting on their photos?*	.2963	.29573
5	*Imagine you went on a trip. How much are you looking forward to posting your photos from the trip so that this person can see, like or comment on them?*	.2516	.27422
6	*How interested are you in knowing exactly where this person is right now?*	.2274	.28850
7	*How much do you think this person is interested in knowing your location right now?*	.2107	.26955
8	*How much do you trust this person?*	.3175	.28810

Table 19 - Descriptive statistics for the seven remaining questions

In the other hand, the questions three, five, and seven tries to capture the "interest" that the participant thinks the friend have on him. Questions two, four, and six have higher averages than questions three, five, and seven. This could mean that most relationships captured are not mutual, or more likely, people usually tend to believe that the person they are rating is slightly less interest in them, than vice-versa.

Table 20 shows us the Pearson correlation for all the survey's answers. At a first glance we can see that the first question (tie strength question) is highly correlated with question eight (trust question). There is a high correlation (0.910) between question six and seven (Location's questions). As expected all the correlations are high and positive indicating coherence in the answers by the participants.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Q1	1	,812	,819	,756	,793	,625	,676	,845
Q2	,812	1	,850	,884	,823	,743	,715	,796
Q3	,819	,850	1	,791	,881	,713	,786	,775
Q4	,756	,884	,791	1	,862	,763	,733	,751
Q5	,793	,823	,881	,862	1	,742	,788	,764
Q6	,625	,743	,713	,763	,742	1	,910	,607
Q7	,676	,715	,786	,733	,788	,910	1	,643
Q8	,845	,796	,775	,751	,764	,607	,643	1

Table 20 - Correlation between the survey's answers. Correlation is significant at the 0.01 level (2 tailed)

6. DISCUSSION

This chapter analyzes and interprets the results described in the previous chapter by answering the three research questions separately, which were formulated previously. Still in this chapter, an informal analysis is conducted to evaluate the sampling method. This is achieved by comparing the main dataset with a small but complete dataset of two particular participants. This same dataset is still used to test both models making an independent comparison of both models. Finally, some limitations of the current work are pointed in the last section of this chapter. Bellow follows the research questions and their analysis and interpretation.

RQ1 – It is *possible to model Tie strength by using a live Facebook application for data collection purposes?*

Very few works address the question if tie strength can be modeled using a real time facebook application as data collector. In fact there are a huge number of Facebook applications that use tie strength, for instance: my top friends; top 10 friends; who cares?; etc. Those applications rely on simple metrics to estimate tie strength such as number of posts, number of comments, and so on. Estimation of tie strength in those applications is very vague, i.e. no empirical work has been made to answer these questions and consequently no work has been done to model tie strength in those types of environments. The regression coefficient R of 0.398 (Table 11) between the 13 independent variables and the tie strength question shows a medium low correlation. The R-squared of 0.158 indicates that 15.8% of the variation in tie strength is explained by these 13 variables. From those 13 variables, six of them have p-values less than 0.05 (Table 13), demonstrating statistical significance. These results support the assumption that is possible to model tie strength via a Facebook application by using a system using the standard developer API.

RQ2 – *Which variables are more important in predicting tie strength?*

When interpreting Table 13 it is important to highlight two important aspects, the standardized coefficients and their respective p-values. Standardized coefficients allow checking the individual importance for each predictive variable. The p-values allow us to know if the predictors are statistically significant. With this in mind we can identify the variables that most contributed for the model. The predictors contributed to the model in the following order: wall posts exchanged; family; wall post comments; appearances together in photos; common groups; inbox messages exchanged. The statistical analysis supports that these variables do not occur by chance and can indeed predict tie strength. The stronger predictor is the wall posts exchanged, This finding are supported by the work of Kahanda, 2009, where he states that wall content and photo tagging are important in link prediction. Common groups have a very small beta coefficient (0.073) however is statistically significant, meaning that group members have a higher chance to be really friends than just acquaintances. The family variable scores well, it was expected to do so

because usually people include only real family members in their family lists. Appearances together in photos implies physical proximity between the actors, the findings support this affirmation. The results suggest that the usage of facebook chat occurs more between strong ties than weak ties. In the other hand, seven of the 13 variables were not statistically significant. Common events do not show a small negative beta coefficient, however the p-value is very high (0.575), so most likely the beta coefficient occurred by chance, discarding this way common events from the equation. Surprisingly, photo comments did not contributed to the model, neither did likes. Three types of likes were considered: wall posts, participant's photos, tagged photos. In none of them were found any relationship with tie strength. This suggests that people do make likes for the content of the posts or photos itself than the affinity for the content's owner. This assumption is easily accepted in the wall posts scenario, however in the photos context this situation seems unlikely to happen. Since photos implies the participant presence, is expected that likes serve as a kind of intimacy between the participant and the "liker". At last, mutual friends did not show any relationship with tie strength itself therefore it does not add strength to the model. In fact, mutual friends do not necessary indicate higher tie strength, a participant who has many mutual friends with other person can feel "obliged" to accept or made a friend request to this person. No mutual friends in a friendship could mean a more genuine friendship since it is not "forced" by other elements – mutual friends. These results support previous findings that mutual friends are negatively correlated with a closeness measure (Bapna, Gupta, Rice, & Sundararajan, 2011).

RQ3 – Can the new model *built from the data collected by the Facebook API perform better than the partial Gilbert and Karahalios' model?*

In the pilot performed before the main study, the correlation between the tie strength question and Gilbert and Karahalios' partial model was 0.27. This value was somehow low than expected therefore motivating a new study with a larger dataset to collect enough data to build a new model and at the same time executing Gilbert's algorithm. A total of 1640 friendships were successfully rated, however only 1329 friendships were entered in the model, after excluding participants who did not granted the inbox message access permission and some outliers. This permission added an extra variable in both the Gilbert's model and the new model, and in both models this permission contributed positively. As said before the multiple regression coefficient R (Table 11) for the new model construction, scored 0.398 which shows a medium low correlation. The not significant variables were omitted from the prediction equation (only 6 terms remained), then performing a correlation on the same dataset. The result was a Pearson's' correlation coefficient of 0.383; very similar when considering all the independent variables in the equation. To comparison purposes the tie strength question was correlated with Gilbert's model. It scored 0.311, slightly higher than what it was gotten in the pilot experiment. Despite that, this comparison has a weakness – the correlation coefficient of the new model is calculated on the same dataset that it was used the build the new model. In order to make a more reliable comparison it was conducted a small validation experiment; this same study had as secondary goal to check the effectiveness of the sampling method. This

last experiment, as already said on this chapter counted with only two participants, in which they rated all their friendships. Three kinds of statistical analysis were performed on this data: a correlation between the tie strength question and Gilbert's partial model; a correlation between the tie strength question and the new model built; and a new multiple regression analysis. The latter will be discussed in the next section. Both models performed relatively well on this dataset. Gilbert's partial algorithm correlated at the 0.398 value and the new model at the 0.409 (Table 17). These values do not show a significant difference between the two models in terms of effectiveness. It is very important to emphasize the fact that Gilbert's partial model uses seven variables and the new model counts with six variables. This research question cannot be answered, a more comprehensive dataset would be necessary to see which model can predict better. With the experiments conducted so far, one can assume that both models perform equally fine.

In their paper, Gilbert and Karahalios made available their top 15 predictive variables and their respective beta coefficients. Some similarities and differences can be found between their and this new model predictive variables. For instance, wall words exchanged is a good predictor; this variable may not be directly, but indirectly related with the wall posts exchanged and number of wall posts comments variables contained in the new model; both these variables are also good predictors. Regarding to the number of inbox messages exchanged variable, the findings suggests that it is positive correlated with tie strength. This assumption clashes with Gilbert and Karahalios work, their findings claims that inbox depth is negatively related with tie strength. However our findings are supported by the work of Whittaker, Terveen, Hill, and Cherny, 1998 which states that familiarity between Usenet users increases with inbox depth.

6.1. Evaluation of the sampling method

Sampling method is the way that the answers and data were collected from the participants. The application aimed to gather an average of 20 questions for each participant, prioritizing the ratio: number of participants by number of answers than vice-versa. It is not practical to ask participants to rate all their Facebook friends. Sampling decision could have other implications in the tie strength's calculation. So an informal analysis has been made to address the following assumption: *A tie strength model could perform better if the dataset was formed by participants that rated all their friends.*

The small validation study cited in the previous chapter had as secondary goal to validate the sampling method adopted. The sampling method chosen consisted in select more or less 20 friendships randomly for each participant. The dataset constitutes partial friendships from a lot of participants (mostly 20 friendships, some less, some more). This hypothesis states that a dataset composed by ties from participants that rated most of their friends would result in higher performance by the tie strength models. The high mean (360) of participants' number of friends motivated this assumption. A high number of friends could signify that a great probability of valuable friendships would be left over. This could be solved by making the "valuable" friendships occur more often by creating a

simple metric which could indicate tie strength, however this action would biased the study. So, a totally random selection was imposed, but the high mean number of friends may have lowered the models' performance. The two participants recruited had a mean of 223 friends. In the last section we saw correlations of 0.398 and 0.409 for the Gilbert's partial model and the new model respectively. Higher values from the ones obtained in the main study. A multiple regression analysis was also conducted in this new study, outputting a multiple regression coefficient R of 0.473 and a R-squared of 0.224 (Table 18). Again these values were higher than the R and R-squared from the main study. This indicates that the sampling method chosen may have not the better. This limitation could have been avoided, by keeping just one question (tie strength question) and removing the other questions, and at the same time making the application faster. Doing this would make participants rate more friends, therefore getting more tie strength data. However due the reasons cited on the 5.6 section the extra questions and the survey design had been kept.

6.2. Study limitations

Some limitations have been associated to this study. Firstly, the inherent restraint of using the facebook API for data collection, although many data were made available through the facebook API, some of them would return null because sometimes participants had declared their personal data as private in their facebook settings. Limiting to some degree the information gathered. Secondly, some participants did not grant access to their inbox private message count (no content was stored neither used in the model, although it was possible). Inbox messages count revealed to be an important predictor of tie strength.

Some technical difficulties had been the source of some participant's dropouts. The main technical problem was failures in some API facebook calls. Without apparent reason some API calls were lost, most of the times if the participant reloaded again the application it started working. Participants who made the survey without supervision, was not aware of the problem and probably quitted the study after the first error. This was more frequent in participants with a great number of friends.

The participants may have experienced some fatigue with the progression of the survey, some of the answers might be biased, especially the latter ones. The Facebook application was totally translated to Portuguese; as well the dictionaries used in the Gilbert's partial model to get the positive emotion words and intimacy words. Some complications may have existed, not in the translation itself but in the semantic context of some words in the dictionaries from English to Portuguese. Fortunately, these problems are believed to have a small effect or even none in the study.

7. CONCLUSION

The main contribution of this thesis is a prediction model of tie strength which can be implemented using the Facebook API. Another tie strength model based in past work (by Gilbert and Karahalios) was also explored; this model was also adapted so that it could be implemented using the Facebook API. A comparative analysis was made to verify the effectiveness of both models. The results suggest a modest improvement of the new model over the existing model, however the new model had the advantage of being simpler to implement as, among other factors, it lacked variables based on complex lexical analysis. Consequently, the new model does not require searching for intimacy or positive emotion words in the participant's wall or inbox messages. This step requires dictionaries of intimacy or positive emotion words, something can be challenging to collect for a wide range of languages.

This thesis has also uncovers the unique contribution of each predictive variable to the final model. Knowing the single weight of each predictive variable is important because it can help understand which variables should be focused on in order to improve future tie strength models. For example, this information can help select the strongest predictors for future refinement. For instance, as the analysis indicates that wall posts exchanged is a good predictor of tie strength future work might refine this measure by analyzing their nature (funny, motivational, personal, etc.). These types of refinements may lead to improvements in the power of the tie strength model.

Tie strength models have many practical applications. For example they could help the monotonous task of automating and sorting friends lists, or more accurately setup automatic narrowcasting services. Depending in the complexity of the models, tie strength calculations could also be adapted to be used in other SNSs. Another interesting aspect of studying tie strength is that it can help us better understand human relations. Since SNSs provide huge repositories of data, this data can be used to better understand the factors which affect people'S friendships in real life. In this scenario, SNSs would work as proxy of real human relationships. This method would be more practical since it would be easier to get data without disturbing people with observations and interviews.

The model built in this thesis was intended to be used through the Facebook API as a tie strength algorithm. It could be used by other developers for example as a friendship predictor when developing Facebook applications. It could also be used by researchers in order to study tie strength or even use this model as a means to study other subjects.

Facebook growth has been remarkable in these last few years, with 840 million active users by the end of 2011 and a total of 10.5 billion minutes spent on the site per day, excluding mobile use (Protalinski, 2011). Indeed, Facebook is not simply growing but is also eliminating competition by attracted users from others SNSs. In June of 2009 there were 17 SNSs in the top 100 sites (according to Alexa and Google trends), but this number has been reduced to just six in just two years (Protalinski, 2012). As the dominance of

Facebook grows this provide increased opportunities to study and apply online tie strength models.

These facts increase the relevance of studying social networks. However, tie strength modeling is not an exact science. Even with all the large amounts of data available it is still difficult to build a model that makes perfect predictions. Human nature is hard to predict and many aspects of human life are still excluded from SNSs. Regardless of this limitation, research on this subject remains important as there is a huge amount of data available to be mined and studied.

7.1. Future Work

Future work is necessary to improve the described tie strength model. A more extensive data collection, as also the refinement of more creative predictive variables would be first steps. Also, it would be crucial to design the study/survey to get a larger number of friendships rated by making it faster and prioritizing only a single tie strength question. This would undoubtedly produce tie strength models that are more accurate and reliable. Instead of working to improve the tie strength model one could choose another path, and apply the tie strength model to real facebook applications. For example a narrowcasting application that sorts and group friends according to strong or weak ties, i.e. close friends and acquaintances.

REFERENCES

Adamic, L. (2003). Friends and neighbors on the Web. *Social Networks*, *25*(3), 211-230. doi:10.1016/S0378-8733(03)00009-1

Adler, P. S., & Kwon, S. W. (2002). Social capital: Prospects for a new concept. *Academy of management review*, *27*(1), 17–40. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/4134367

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th international conference on World Wide Web - WWW '07* (p. 835). New York, New York, USA: ACM Press. doi:10.1145/1242572.1242685

Allison, P. D. (1998). *Multiple Regression : A Primer (Undergraduate Research Methods & Statistics in the Social Sciences)* (1st ed., p. 224). Sage Publications, Inc; 1 edition (December 29, 1998).

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., & Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature*, *427*(6977), 839-843. Nature Publishing Group. doi:10.1038/nature02289

Bapna, R., Gupta, A., Rice, S., & Sundararajan, A. (2011). Trust , Reciprocity and the Strength of Social Ties : An Online Social Network based Field Experiment, 1-17.

Bargh, J. A., & McKenna, K. Y. A. (2004). The internet and social life. *Annual Review of Psychology*, *55*(1), 573-590. Annual Reviews Inc. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14744227

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747-52. doi:10.1073/pnas.0400087101

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science (New York, N.Y.)*, *323*(5916), 892-5. doi:10.1126/science.1165821

Bourdieu, P. (1986). *The forms of capital*. Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470755679.ch15/summary

Boyd, Danah M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230. doi:10.1111/j.1083-6101.2007.00393.x

Boyd, D.M. (2006). Friends, friendsters, and myspace top 8: Writing community into being on social network sites. *First Monday*, *11*(12), 1–15. Citeseer. Retrieved from http://www.mendeley.com/research/gender-differences-of-isometric-hamstring-fatigue-on-muscle-activation-and-knee-extension-moments-when-landing/

Bradley, M. M., & Lang, P. J. (1999). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.

Burbary, K. (2011). No Title. *Facebook Demographics*. Retrieved August 18, 2012, from http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/

Burke, M., Kraut, R., & Marlow, C. (2011). Social capital on Facebook: Differentiating uses and users. *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 571–580). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1979023

Burt, R. (2004). Structural Holes and Good Ideas1. *American journal of sociology*, *110*(2), 349-399. Retrieved from http://www.jstor.org/stable/10.1086/421787?journalCode=ajs

Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Crawling Facebook for social network analysis purposes. *Arxiv preprint arXiv:1105.6307*, 0-7. Retrieved from http://arxiv.org/abs/1105.6307

Choi, J. H.-jeong. (2003). Living in Cyworld: Contextualising Cy-Ties in South Korea. (A. Bruns & J. Jacobs, Eds.)*Development*, 173-186. Peter Lang. Retrieved from http://www.nicemustard.com/files/jaz_c_cyworld_ch.pdf

Chun, H., Kwak, H., Eom, Y., & Ahn, Y. (2008). Comparison of online social relations in volume vs interaction: a case study of cyworld. *Proceedings of the 8th*, *v*, 57-69. Retrieved from http://portal.acm.org/citation.cfm?id=1452528

De Meo, P., Ferrara, E., & Fiumara, G. (2011). Finding similar users in Facebook. *Social Networking and Community Behavior Modeling: Qualitative and Quantitative Measurement* (pp. 1-26). Igi Publishing. Retrieved from http://cogprints.org/7634/

Donath, J., & Boyd, D. (2004). Public Displays of Connection. *BT Technology Journal*, *22*(4), 71-82. Springer. doi:10.1023/B:BTTJ.0000047585.06264.cc

Duhan, D. F., Johnson, S. D., Wilcox, J. B., & Harrell, G. D. (1997). Influences on Consumer Use of Word-of-Mouth Recommendation Sources. *Journal of the Academy of Marketing Science*, *25*(4), 283-295. doi:10.1177/0092070397254001

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook “Friends:” Social Capital and College Students’ Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, *12*(4), 1143-1168. doi:10.1111/j.1083-6101.2007.00367.x

Ennett, S. T., Bauman, K. E., Hussong, A., Faris, R., Foshee, V. A., Cai, L., & DuRant, R. H. (2006). The peer context of adolescent substance use: Findings from social network analysis. *Journal of Research on Adolescence*, *16*(2), 159–186. Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1532-7795.2006.00127.x/full

Everett, M., & Borgatti, S. (2005). Ego network betweenness. *Social Networks*, *27*(1), 31-38. doi:10.1016/j.socnet.2004.11.007

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239. Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/0378873378900217

Gilbert, E. (2010). *Computing Tie Strength*. University of Illinois.

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. *Proceedings of the 27th international conference on Human factors in computing systems* (pp. 211–220). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1518736

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of OSNs. *INFOCOM, 2010 Proceedings IEEE* (pp. 1–9). Ieee. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5462078

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, *12*(3), 211–223. Springer. Retrieved from http://www.springerlink.com/index/T24301H2268N5V46.pdf

Granovetter, M. S. (1973, May). weak ties. *The Journal of applied psychology*. doi:10.1037/a0018761

GraphML Project Group. (2002). The GraphML File Format. Retrieved May 18, 2012, from http://graphml.graphdrawing.org/index.html

Hammer, E., & Hardt, D. (2012). The OAuth 2 . 0 Authorization Framework, *5849*.

Hampton, KN, & Goulet, L. (2011). Social networking sites and our lives. *Retrieved July 12,*. Retrieved from http://www.namingandtreating.com/wp-content/uploads/2011/07/PIP-Social-networking-sites-and-our-lives.pdf

Hampton, Keith, & Wellman, B. (2003). Neighboring in Netville. *City Community*, *3*(Fall), 1-40.

Haynes, R. (2010). The Value of R-squared in Regression (it is minor). Retrieved July 25, 2012, from http://www.smartersolutions.com/blog/wordpress/2010/04/02/the-value-of-r-squared-in-regression-it-is-minor/

Helliwell, J. F., & Putnam, R. D. (2004). The social context of well-being. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *359*(1449), 1435-46. doi:10.1098/rstb.2004.1522

Joinson, A. (2008). Looking at, looking up or keeping up with people?: motives and use of facebook. *Proceeding of the twenty-sixth annual SIGCHI*, 1027-1036. Retrieved from http://dl.acm.org/citation.cfm?id=1357213

Kahanda, I. (2009). Using transactional information to predict link strength in online social networks. *International Conference on Weblogs and Social*, 74-81. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPDFInterstitial/213/411

Kavanaugh, A., Carroll, J. M., Rosson, M. B., Zin, T. T., & Reese, D. D. (2005). Community Networks: Where Offline Communities Meet Online. *Journal of Computer-Mediated*

Communication, *10*(4), 0. Blackwell Publishing Ltd. doi:10.1111/j.1083-6101.2005.tb00266.x

Krackhardt, D. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*. Retrieved from http://www.jstor.org/stable/10.2307/2786835

Krackhardt, David. (1992). The strength of strong ties: The importance of philos in organizations. In N. Nohria & R. Eccles (Eds.), *Networks and Organizations Structures Form and Action* (pp. 216-239). Harvard Business School Press.

Lampe, C., & Ellison, N. (2006). A Face (book) in the crowd: Social searching vs. social browsing. *Proceedings of the 2006 20th*, 0-3. Retrieved from http://dl.acm.org/citation.cfm?id=1180875.1180901

Lane, D. M. (1998). Introduction to Multiple Regression. Retrieved April 14, 2012, from http://davidmlane.com/hyperstat/prediction.html

Latora, V., & Marchiori, M. (2007). A measure of centrality based on network efficiency. *New Journal of Physics*, *9*(6), 188-188. doi:10.1088/1367-2630/9/6/188

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., et al. (2009). Social science. Computational social science. *Science (New York, N.Y.)*, *323*(5915), 721-3. doi:10.1126/science.1167742

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330-342. doi:10.1016/j.socnet.2008.07.002

Lin, K.-Y., & Lu, H.-P. (2011). Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior*, *27*(3), 1152-1161. doi:10.1016/j.chb.2010.12.009

Marsden, P. V., & Campbell, K. E. (1984). Measuring tie strength. *Social forces*, *63*(2), 482–501. Oxford University Press. Retrieved from http://sf.oxfordjournals.org/content/63/2/482.short

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 29–42). New York, New York, USA: ACM. doi:10.1145/1298306.1298311

Muncer, S., Burrows, R., Pleace, N., Loader, B., & Nettleton, S. (2000). Births , deaths , sex and marriage … but very few presents ? A case study of social support in cyberspace. *Critical Public Health*, *10*(1), 1–18. Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/713658221

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 5. APS. Retrieved from http://arxiv.org/abs/cond-mat/0205405

Nie, N. H. (2001). Sociability, Interpersonal Relations, and the Internet: Reconciling Conflicting Findings. *American Behavioral Scientist*, *45*(3), 420-435. doi:10.1177/00027640121957277

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Arxiv preprint arXiv:1103.2903*. Retrieved from http://cs229.stanford.edu/proj2009/PandeyIyer.pdf

Palmer, C. R., Gibbons, P. B., & Faloutsos, C. (2002). ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs. *Science*, *1*, 81-90. ACM Press. doi:10.1145/775047.775059

Panovich, K., Miller, R., & Karger, D. (2012). Tie strength in question & answer on social network sites. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1057–1066). ACM. Retrieved from http://groups.csail.mit.edu/uid/other-pubs/cscw2012-tiestrength.pdf

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of SocioLinguistics*, *5*(2), 180-213. doi:10.1111/1467-9481.00147

Petróczi, A., Nepusz, T., & Bazsó, F. (2006). Measuring tie-strength in virtual social networks. *Connections*, *27*(2), 39–52. International Network for Social Network Analysis. Retrieved from http://www.insna.org/PDF/Connections/v27/2006_I-2-5.pdf

Protalinski, E. (2011). 10.5 billion minutes spent on Facebook daily, excluding mobile. *ZDNET*. Retrieved August 10, 2012, from http://www.zdnet.com/blog/facebook/10-5-billion-minutes-spent-on-facebook-daily-excluding-mobile/11034

Protalinski, E. (2012). Facebook is killing top social networks worldwide. *ZDNET*. Retrieved August 15, 2012, from http://www.zdnet.com/blog/facebook/facebook-is-killing-top-social-networks-worldwide/8025

Schaefer, C., Coyne, J. C., & Lazarus, R. S. (1981). The health-related functions of social support. *Journal of Behavioral Medicine*, *4*(4), 381-406. Springer Netherlands. doi:10.1007/BF00846149

Scott, J. (1988). Social Network Analysis. *Sociology*, *22*(1), 109-127. doi:10.1177/0038038588022001007

Sledgianowski, D., & Kulviwat, S. (2009). uSIng SoCIal network SIteS : the effeCtS of playfulneSS , CrItICal maSS and truSt In a hedonIC Context. *Journal of Computer Information Systems*, *49*(4), 74-83. International Association for Computer Information Systems. Retrieved from http://www.allbusiness.com/marketing-advertising/marketing-advertising-overview/12723438-1.html

U. Brandes, M. Eiglsperger, J. L. (2000). GraphML Primer. Retrieved May 9, 2012, from http://graphml.graphdrawing.org/primer/graphml-primer.html

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. *Arxiv preprint arXiv: …*, 1-17. Retrieved from http://arxiv.org/abs/1111.4503

Wang, F.-Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligent Systems*, *22*(2), 79-83. doi:10.1109/MIS.2007.41

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. *American Ethnologist* (Vol. 24, p. 857 pages). Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(June), 440-442.

Whittaker, S., Terveen, L., Hill, W., & Cherny, L. (1998). The dynamics of mass interaction. *… of the 1998 ACM conference …*. Retrieved from http://dl.acm.org/citation.cfm?id=289500

Wuchty, S., & Uzzi, B. (2011). Human Communication Dynamics in Digital Footsteps: A Study of the Agreement between Self-Reported Ties and Email Networks. *PloS one*, *6*(11), e26972. doi:10.1371/journal.pone.0026972

Xiang, R., Neville, J., & Rogati, M. (2010). Modeling relationship strength in online social networks. *Proceedings of the 19th international conference on World wide web - WWW '10*, 981. New York, New York, USA: ACM Press. doi:10.1145/1772690.1772790

Ye, S., Lang, J., & Wu, F. (2010). Crawling online social graphs. *Web Conference (APWEB), 2010 12th International Asia-Pacific* (pp. 236–242). IEEE. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:University+of+California#5

Zhao, J., Wu, J., & Feng, X. (2011). Information propagation in online social networks : a tie-strength perspective. *Knowledge and Information Systems*. doi:10.1007/s10115-011-0445-x

Zhao, J., Wu, J., & Xu, K. (2010, July). Weak ties: subtle role of information diffusion in online social networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20866687

APPENDIX 1. SNA Metrics

One important concept of SNA is their metrics, which allow measuring the topology of a social network, for posterior analysis and conclusions. In this section the most relevant SNA metrics to the current work are presented. Centrality metrics are a group of metrics particularly important to the current work because it permits to identify key nodes, i.e. identify the most popular/important nodes within a network. Centrality measures include: degree, betweenness eigenvector, and closeness centrality. In this study the degree and betweenness will be covered with some detail. Other metrics of interest, such as: clustering coefficient, assortativity, disassortivity, diameter, and disparity are also presented. All of these metrics are essentially important if one wish to analyze the GraphML files provided in this thesis.

Before continuing it is important to distinguish two different types of networks: ego/personal networks and whole/complete networks. Ego networks may be defined as networks consisting of a single actor (ego) together with the actors (alters) they are connected to, and all the links among those alters. On the other hand, whole networks as the name indicates, are complete networks. Most studies have adopted ego-centric network because these networks provide a simple structure which brings the benefit of simplicity in data collection (Everett & Borgatti, 2005).

Degree Centrality

The most common measure of Centrality is the degree. A degree of a determined node is the count of the number of ties to other actors in the network. This metric usually corresponds to the number of friends of a SNS, but it can be extended to other social activities and therefore obtain multiple graphs and their respective degree (De Meo, Ferrara, & Fiumara, 2011). The standard representation of a social network assumes that the graph is undirected, for instance the notion of friendship in Facebook, which does not differentiate a receptor from a receiver, yet some studies apply the use of directed graphs for example to map an activity network (Chun et al., 2008; Xiang et al., 2010). In these cases the degree is divided and represented by the in-degree (edges directed to the node) and out-degree (edges which directs out from a node).

Betweenness Centrality

Defined by Freeman (1979) as an index measuring one's potential to control communication in a given network. Put in another words, it can be simply defined as the number of shortest paths from all vertices to all others that pass through that node. This measure considers the connectivity of the node's neighbors and assigns a higher value for nodes that serve as a bridge between clusters, as it is visible in figure 2. Although there are several metrics to compute the centrality of a node, such as degree, closeness, eigenvector and betweenneess centrality, the latter one has been shown as the more accurate metric to calculate the actual importance of a node in a network (Latora & Marchiori, 2007;

Wasserman & Faust, 1994). So, a person can have a low degree (few connections) but if one has an high betweenneess centrality, then it plays an important role in a social network, because it contributes to a more connected network (Everett & Borgatti, 2005). For instance, Ennett et al., (2006) argue that students who demonstrate high betweenneess centrality play a central role in the transmission of behaviors, norms, and cultural knowledge. In the work of Catanese, De Meo, Ferrara, Fiumara, and Provetti (2011) it was concluded that the study of betweenneess centrality in Facebook is fundamental for all those aspects related to discovering central nodes of the network, and it is a numerical property for applications, e.g., for marketing purposes, broadcasting news, etc. Also, Lewis, Kaufman, Gonzalez, Wimmer, and Christakis (2008) showed that betweenneess centrality was used to understand the role of social categories, such as race/ethnicity, socioeconomic status, and gender in network behavior more generally and in online social network behavior in particular. In two similar experiments realized on Facebook and performed by Zhao, Wu, & Feng, (2011) and Zhao, Wu, & Xu, (2010) the nodes with higher betweenneess centrality were used preferentially to diffuse information, and it was found out that it achieved better efficiency of information propagation than random, strong and even weak nodes.

Clustering Coefficient

In simple terms this metric computes the tendency of a network to cluster together and is computed as the ratio of number of connections in the neighborhood of a node and the number of connections if the neighborhood was fully connected. In figure 2 it is visible the calculation of the clustering coefficient in three different simple networks. The clustering coefficient of a network is defined by the average of the individual clustering coefficients (Wasserman & Faust, 1994). Barrat, Barthélemy, Pastor-Satorras, and Vespignani (2004) proposed a slightly modified model for calculate the clustering coefficient of a weighted network. Based on Barrat's model, Chun et al., (2008) calculated both the unweighted and weighted clustering coefficient of Cyworld which corresponds the clustering coefficient of the network and the activity network respectively, with the purpose to check if the activity network of a network flows between their close neighbors. Networks with a low value of clustering coefficient reveal a poor rate of information propagation because there is a negative correlation between the number of positive weak ties and the clustering coefficient of the network, weak ties are fundamental to disseminate information more efficiently (Zhao et al., 2011). In order to solve the apparent conflict between clustering coefficient and short paths Watts and Strogatz (1998) defined a simple model of social networks to show that as long as there is a small fraction of 'random' connections between cliques, social networks could display both high clustering and small average shortest paths.

Other metrics of interest

Here other interest metric are presented but with less detail. Other important centrality measure is the metric of closeness which is based on the notion closeness or distance. It focuses on how close a node is to all other nodes in the network. An actor which has high closeness score can communicate information more efficiently to other actors (Wasserman & Faust, 1994). Still in the centrality metrics there is eigenvector centrality which is a more sophisticated view of centrality, it makes an approximation of the importance of each node in a graph. Ye, Lang, and Wu (2010) states that a person with few connections could have a very high eigenvector centrality if those few connections were themselves very well connected.

The assortativity metric is directly related with homophily (tendency of individuals to associate with similar others), if the nodes in a social network are connected with similar nodes it is said to have an assortative mixing pattern, otherwise they have a diassortative mixing pattern. Assortative mixing patterns are an unique characteristic of social networks (Newman, 2002) and basically most known studies follows this pattern (Ahn, Han, Kwak, Moon, & Jeong, 2007). We can say that a network has diameter D if every pair of nodes can be connected by a path of length of at most D edges (Wasserman & Faust, 1994). Today's SNSs (e.g., Facebook, Twitter, LinkedIn, etc.) are huge, and their diameter computation is infeasible. Palmer, Gibbons and Faloutsos (2002) proposed an approximation for the effective diameter of massive networks but it requires the entire knowledge of the network's topology. Not very know, yet important, it is the disparity (Almaas, Kovács, Vicsek, Oltvai, & Barabási, 2004) between two nodes, this metric allows us to identify if the nodes interact evenly, e.g., two friends in Facebook have a reciprocal relationship. In Chun et al. (2008) experiment, users with a smaller number of correspondents tend to interact more with a subset of correspondents, while users with a very large number of correspondents actually spread their activity evenly across all of the correspondents.

APPENDIX 2. GraphML Files

The Tie Strength Survey application generated a GraphML file for each participant, containing the relationships between all the participant's friends. A GraphML file format is a XML based format for graphs (GraphML Project Group, 2002; U. Brandes, M. Eiglsperger, 2000). This thesis did not follow this line of research since it was out of its scope, therefore the study of GraphML files have been discontinued. However this work produced a total of 85 GraphML files, which is available for anyone who wants to study and analyze.

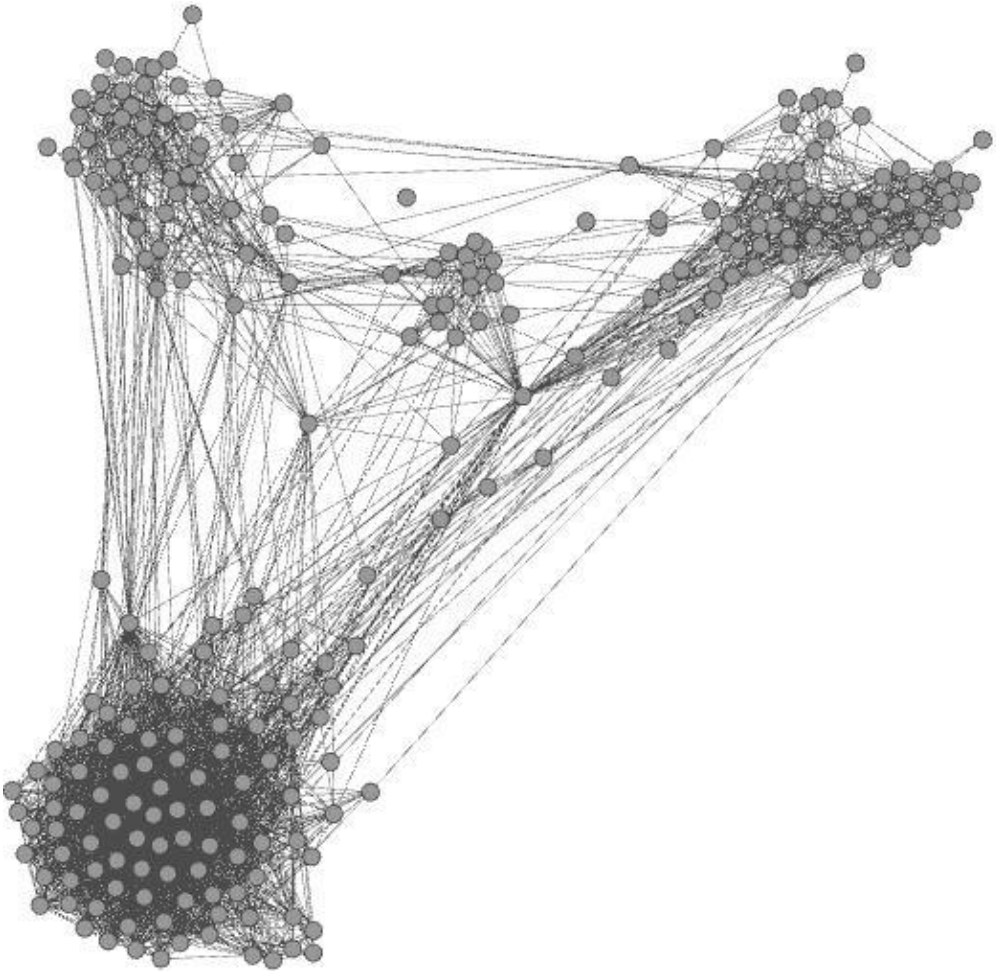

Figure 15 - A participant's network displayed in the program Gephi and obtained from the GraphML file

APPENDIX 3. Tie Strength Survey Screenshots

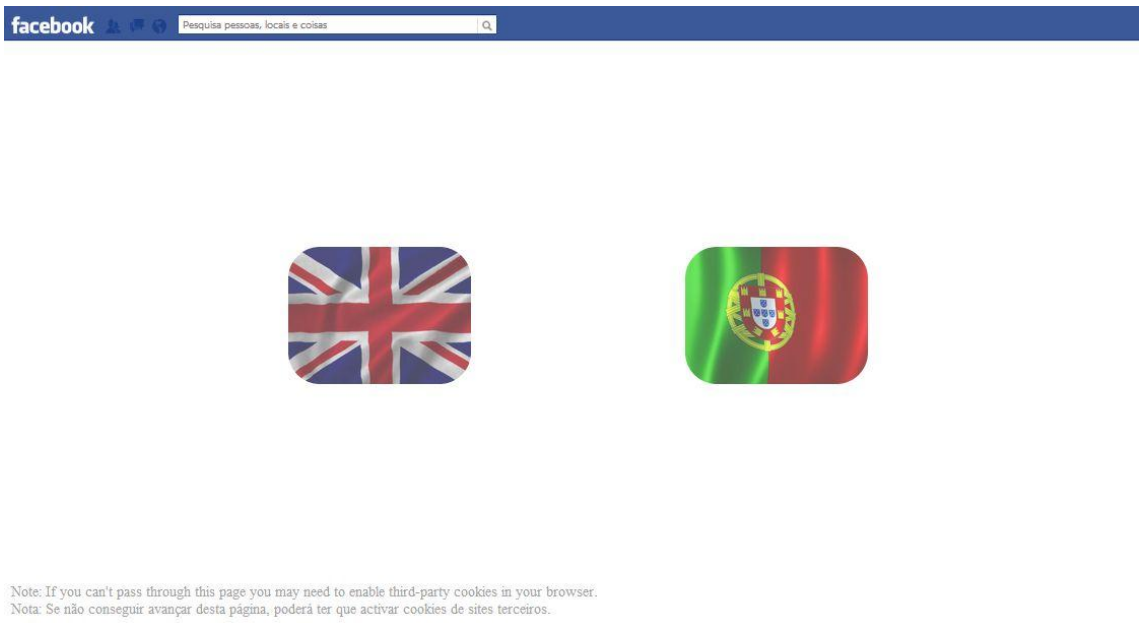

Figure 16 – Choosing language page, the first page to be shown

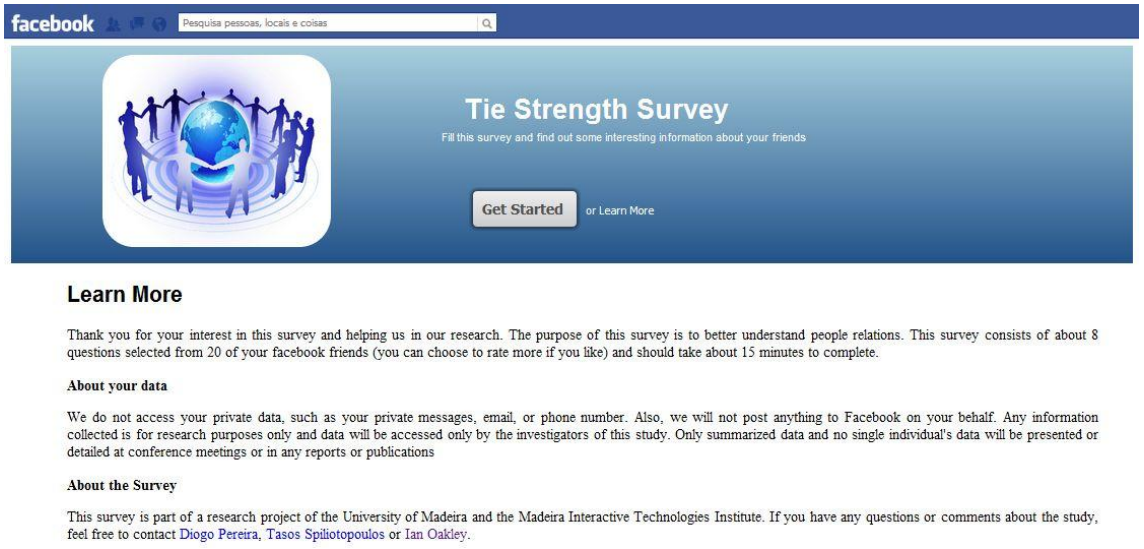

Figure 17 - Learn more section, it is available from within the application, right after choosing the language

Figure 18 - Demographic questions about the participants

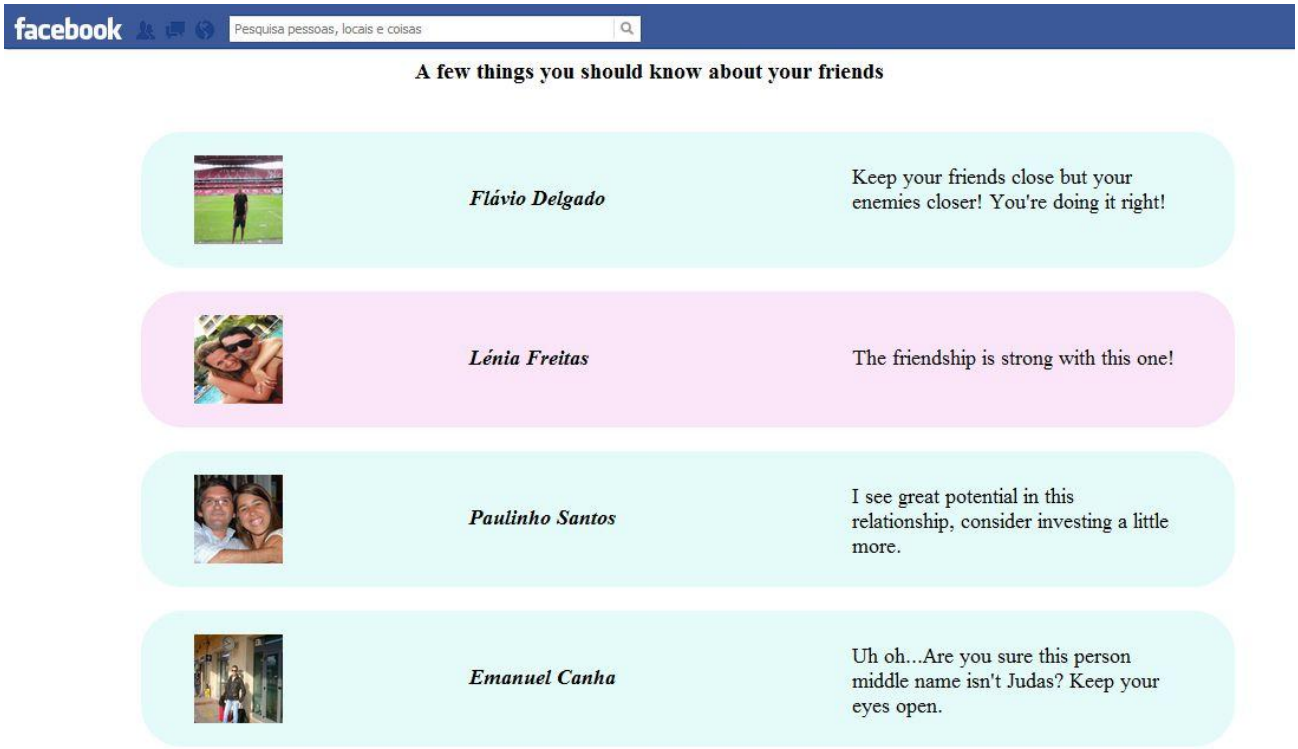

Figure 19 - Some funny phrases for some friends that the participant rated. These phrases were only available for the friends that the participant rated

Top 10 friends

Rank	Picture	Name
1º		Ilidio Freitas
2º		Zé Rodrigues
3º		Fabio Santos
4º		Cátia Neves
5º		Jaime Andrade
6º		Fábio Góis
7º		Lita Freitas
8º		Gonçalo Gouveia
9º		Joana Pereira
10º		Filipe Graca Santos

Figure 20 - Top 10 friends, provided after the participant rated at least 20 of their friends

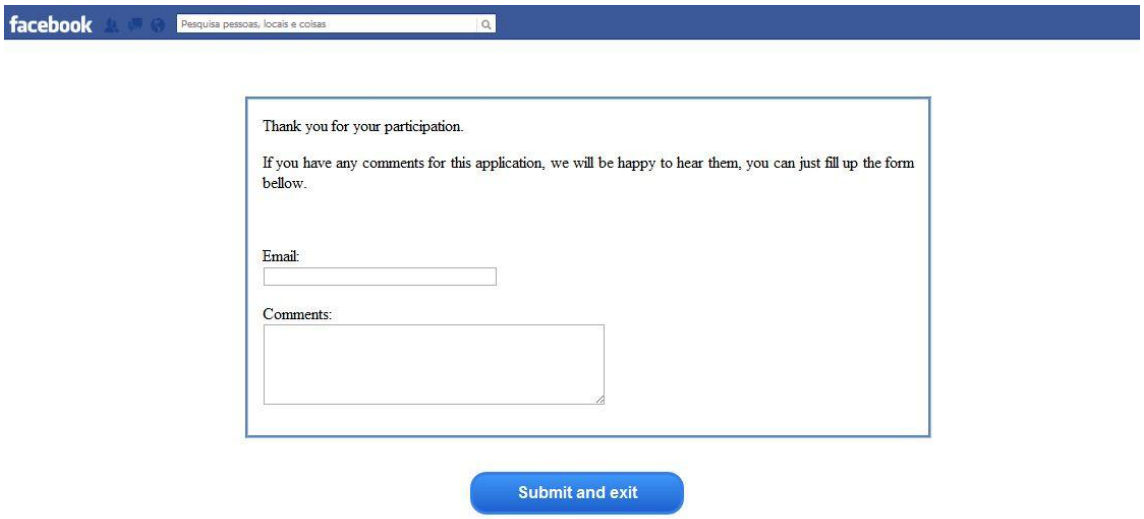

Figure 21 - Last page shown in the application