EXPLORING THE GOOGLE+ SOCIAL GRAPH TO UNDERSTAND USERS' COMMUNICATION

GABRIEL MAGNO DE OLIVEIRA SILVA

EXPLORING THE GOOGLE+ SOCIAL GRAPH TO UNDERSTAND USERS' COMMUNICATION

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Virgílio Augusto Fernandes Almeida

Belo Horizonte

Junho de 2014

GABRIEL MAGNO DE OLIVEIRA SILVA

EXPLORING THE GOOGLE+ SOCIAL GRAPH TO UNDERSTAND USERS' COMMUNICATION

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Virgílio Augusto Fernandes Almeida

Belo Horizonte

June 2014

de Oliveira Silva,Gabriel Magno

Exploring the Google+ Social Graph to Understand Users' Communication / Gabriel Magno de Oliveira Silva. — Belo Horizonte, 2014
xxiv, 65 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Virgílio Augusto Fernandes Almeida

1. Online Social Networks. 2. Complex Networks. 3. Google+. 4. Geo-location. 5. Internet Linguistics. 6. Microtext Analysis. I. Título.

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* Ãľ uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Aos meus pais, Márcia e Salvador, e à minha irmã, Ághata.

Acknowledgments

Agradeço a todos que direta ou indiretamente contribuíram na elaboração deste trabalho.

A Deus, pela vida.

Ao meu orientador, Virgilio, pelo conhecimento, o suporte e a compreensão, fundamentais para o desenvolvimento deste trabalho.

Aos professores colaboradores Meeyoung Cha, Marcos André Gonçalves e César Cambraia, pela oportunidade de trabalho em conjunto e suas contribuições essenciais.

Aos colegas pesquisadores Diego Saez-Trumper, Giovanni Comarela e Evandro Cunha, pelas valiosas dicas e contribuições e, sobretudo, a amizade.

Aos colegas do laboratório CAMPS, pela companhia e a amizade, fazendo com que os dias de trabalho fossem mais agradáveis.

Aos amigos, pela motivação e o suporte durante os momentos difíceis.

À minha família, pelo amor, incentivo e apoio incondicional.

"Lei de Hofstadter: Sempre vai demorar mais tempo do que você espera, mesmo quando se levar em consideração a Lei de Hofstadter."
(Douglas Hofstadter in 'Gödel, Escher, Bach: An Eternal Golden Braid')

Resumo

Este trabalho apresenta uma análise detalhada da rede social online Google+, em relação a características de rede, padrões de utilização pelo mundo e aspectos linguísticos entre membros de diferentes grupos sociais. Identificamos as principais diferenças e similaridades com outras redes populares como o Facebook e o Twitter, para determinar se o Google+ tem alguma característica inovadora ou é uma rede social como qualquer outra. O estudo é baseado em duas coletas em grande escala de milhões de usuários, que representa praticamente a rede completa, incluindo informações pessoais públicas do perfil, lista de amigos e as postagens. Comparado a outras redes, observamos que o caminho médio entre usuários é levemente maior, possivelmente devido ao fato de o Google+ ser um novo sistema onde os relacionamentos ainda estão rapidamente crescendo. O Google+ tem um nível de reciprocidade maior do que o Twitter, indicando que o Google+ é mais social. Observamos que o Google+ é popular em países com baixa penetração de Internet. Baseado na quantidade e no tipo de informação publicamente compartilhada nos perfis dos usuários, também observamos que a noção de privacidade varia significativamente em diferentes culturas. Nosso estudo revela que grupos sociais possuem peculiaridades linguísticas – tal como tendência a usar vocabulário profissional, sugerindo que o Google+ possa ser adotado, por certos usuários, para atividades profissionais, ou que os membros não se dissociam de seus trabalhos quando interagem nesse ambiente. Nossos resultados ajudam a entender não apenas peculiaridades coletivas de usuários de mídias sociais online, mas também características importantes do gênero textual "post", sendo um dos primeiros e mais compreensivos estudos nesse tópico.

Palavras-chave: Redes Sociais Online, Redes Complexas, Google+, Geo-localização, Linguística de Internet, Análise de Microtextos.

Abstract

This work presents a detailed analysis of the Google+ online social network in terms of network characteristics, patterns of utilization around the world and linguistic features among members of different social groups. We identify the key differences and similarities with other popular networks like Facebook and Twitter, in order to determine whether Google+ is a new paradigm or yet another social network. This study is based on two large-scale crawls of million users, that represent virtually the entire network, including public personal information from the profile, list of friends and posts. Compared to other networks, we observe that the average path length between users is slightly higher, possibly because Google+ is a new system where relationships are still rapidly growing. Google+ shows a higher level of reciprocity than Twitter, which also has directed social links. We find that Google+ is popular in countries with relatively low Internet penetration. Based on the amount and types of information publicly shared in user profiles, we also find that the notion of privacy varies significantly across different cultures. Our study reveals that groups hold linguistic particularities – such as a tendency to use professional vocabulary, suggesting that Google+ might be employed, by certain users, for professional activities, or that members do not dissociate from their jobs when interacting in this environment. Our findings help to understand not only collective peculiarities of online social media users, but also important characteristics of the textual genre post, being one of the first and most comprehensive studies on this topic.

Keywords: Online Social Networks, Complex Networks, Google+, Geo-location, Internet Linguistics, Microtext Analysis.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Online Social Networks are a global information infrastructure, where individuals bring their social relations online and share information, photos, songs, videos, as well as ideas. Social networking sites like Facebook now reach 82% of the world's Internet-using population or about 1.2 billion people in total according to comScore [15]. In fact social networking became the most popular online activity worldwide. Accordingly, a number of researchers have tried to understand user behaviors and characteristics of various online social networks, where Twitter and Facebook have been the two most popularly examined platforms [41, 11, 69, 3].

To compete in this field, Google has launched in June 2011 its own social networking service called Google+ (https://plus.google.com/). The platform was announced as a new generation of social network and included several new features, such as *circles* that allow users to share different content with different people and *hangouts* that let users to create video chatting session and invite up to nine people from their circles of friends to share the environment [34].

Since its launch, the Google+ social network has been adding new users at a rapid pace. In fact, it is known as the fastest growing network ever, reaching 20 million visitors in only 21 days [14]. The service has later reached 62 million registered users as of December 2011 [35] and a total of 250 million registered users of whom 150 million are active as of June 2012 [25].

Once Google+ has become a popular social media network, it is important to understand how it compares to other social network models. Typical questions follow. How are people connected on Google+? Who are the most popular users? How are users distributed worldwide? What is the impact of geography on the social relationships?

Furthermore, the rapid adoption rate of the service raises interesting questions

about online privacy. One crucial question is on what the default privacy settings should be. Along these lines, it is worth examining how "closed" social networking sites are, compared to the "open" Internet. Google has positioned itself as promoter of the Internet openness against other social networking services that are often described "walled garden" [8] due to limited access to their internal web pages. Then, is Google+ different? How open is it and how does it impact user interactions?

Increasingly, researchers have taken advantage of the vast amount of language data that online applications can provide, which gave rise to a new subfield of knowledge called *Internet linguistics* [19]. According to Crystal [18], the Internet plays an unprecedented role in the study of language, as it allows linguists to use rich documented datasets to investigate language use in various levels and the nature of the language employed by Web users. From this perspective, authors are concerned with understanding and describing computer-mediated communication, as well as developing tools to provide better online services. Opportunities arising in this area include the employment of collections from user-generated content websites as corpora of large-scale natural language data.

To better understand its typical features, the investigation of formal and functional aspects of the content shared by its members is of utmost importance. Here, we study one kind of content published in Google+: status updates, usually called *posts*. Our focus is to characterize Google+ posts and to identify differences and similarities among linguistic aspects of texts produced by users considering their distinct social characteristics. We analyze texts from male and female members from 10 countries and 15 groups of occupations, since gender, location and job are known as factors that influence language usage in a myriad of domains [42].Our main hypothesis is that the membership in certain social groups may influence aspects of the language employed by users when posting, reflecting patterns observed in other online and offline situations.

To answer these questions we have crawled millions user profiles and relationship links among users, as well as any publicly available data about the users such as gender, geo-location, and relationship status. A relatively large number of users leave personal information publicly available for anyone to see. This kind of information allows us to analyze user behavior patterns and compare them to previous research results obtained for other social networks, e.g., Facebook and Twitter.

Based on the gathered data, we characterize the novel social network model provided by Google+ in depth, its user base, its geographical distribution, and compare its main characteristics with other social network services and among different social groups like gender, country and occupation. Among various findings, some of the main results are summarized as follows:

1. Our analysis on the top users based on the circles list indicate that the majority of the top users (5 out of 30) are well-known individuals from information technology industry;

2. By looking into users who share their work or home contact information publicly (1% of all suers), we observe that a large fraction of the users who share telephone numbers are male and single;

3. We find that users share strikingly different amounts of information to public in their profiles depending on the country they are from;

4. By examining the social links between the users in relation to their countries, we observe that physical distance is crucial in the likelihood of forming a social link between two users;

5. The fraction of global and national links also vary according to the countries, indicating the different patterns of usages of the Google+ service across different cultures;

6. The fraction of misspellings in Google+ posts varies significantly among different social groups. We found a relationship between this fraction and the nature of individuals' professional activities;

7. Certain social groups organize their posts differently, so that the content and the structure of the messages may be quite distinct among users from particular countries, genders and occupations;

8. Social groups are not homogeneous with regard to the use of semantic categories of words. Particularly, we discovered that the vocabulary employed in Google+ posts is highly related to the users' occupations, which may indicate that this OSN is often used for professional activities or that members do not dissociate from their jobs when interacting in this environment;

Most of the analysis presented in this work were previously published by the author in two paper. The first one [49] contains the structural properties comparison and geographical patterns analyses presented in this work, but here we include a more recent dataset and also discuss the evolution between the datasets. The second paper [21] contains all the linguistic analyses among social groups presented here.

This work is organized as follows. In Chapter 2 we present and discuss related work. In Chapter 3 we describe the Google+ platform and how we collected the data.

In Chapter 4 we present the analysis of complex network metrics of the Google+ social graph, as well as the content of user profiles. In Chapter 5 we study the characteristics of economics, privacy and content among users of different countries. In Chapter 6 we present the analysis of the linguistic characteristics present in Google+ posts among different social groups, including gender, country and occupation. We discuss the implications in Chapter 7 and finally we conclude our work in Chapter 8.

Chapter 2

Related Work

In this chapter we present and discuss the works related to the analysis in our work. The chapter is organized in four sections related to the topics of the papers presented. In Section 1 we present works of characterization in online social networks. In Section 2 we discuss papers that analyzed the geo-location of OSN users. In Section 3 we show works related to linguistic analysis, both in the physical and the online world. Finally, in Section 4 we present papers that also studied Google+.

2.1 Characterization

Characterization of social networks and user behavior is fundamental to the understanding and engineering of these services on the Internet. Many studies focus on the characterization of the most popular social network models, such as Facebook, Twitter, Orkut, Cyworld and others. Some of the important findings of these studies include establishing power law distributions for in- and out-degree, short average distance between pairs of users, a very large connected component, and a small number of extremely popular users. Thus, in the remainder of this section, we restrict our coverage of related work to studies that concentrate on characterization of other social network models.

Mislove et al. [51] studied graph theoretic properties of social networks, based on the friend network of Orkut, Flickr, LiveJournal, and YouTube. They confirmed the power-law, small-world, and scale-free properties of these social network services. Ahn et al. [1] studied the network properties of Cyworld, a popular social networking service in South Korea. They compared the explicit friend relationship network with the implicit network created by messages exchanged on Cyworld's guestbook. They

found similarities in both networks: the in-degree and out-degree were close to each other and social interaction through the guestbook was highly reciprocal.

Recently, Ugander et al. [69, 3] used the complete Facebook dataset to study the social graph of Facebook. They show - among other things - that the degree of separation in that platform is 4.7, while we find that in Google+ it is 5.9. This difference may be explained by the fact that Google+ is a new platform at it should get denser in the future, as studied by [45] for different networks. Two recent references [41, 11] focus on the study of the Twitter graph. Other studies comparing different social network models were done by [51, 6, 7]. In general, Google+ presents a combination of the characteristics of other networks, such as Facebook and Twitter.

2.2 Geo-location

When it comes to research on geo-location of users in online social networks, Liben-Nowell et al. [46] analyzed the geographical location of LiveJournal users and found a strong correlation between friendship and geographic proximity. This work confirms that most social links in the blog network are correlated with physical distance and only 33% of the friendships are independent of geography. We find a similar pattern in the friendship structure of Google+ in this work. Recently, Scellato et al. [61] showed that there is a strong relationship between geographical distance and the probability of being friends in social networks. They discuss the implications of geo-location for social networking sites. Rodrigues et al. [59] investigate the word-of-mouth based content discovery by analyzing URLs in Twitter. They also showed that propagation and physical proximity have correlation. Finally, Poblete et al. [55] studied a large amount of data gathered from Twitter and showed the various usages of the system depending across different countries.

2.3 Linguistics

Literature on the relations between language and society is really vast. Labov's [42], Trudgill's [68] and Romaine's [60] works present the main findings of decades of research, considering also the correlations between language variation and the social factors that we contemplate here.

Bell et al. [5] used computational tools to investigate differences in language styles among men and women. Their finding that women use more social words than men could be verified by our analysis. It is also worth mentioning Lakoff's [43] seminal

work on language and gender, where the author indicates that a number of linguistic features can distinguish men's speech from women's.

The study of linguistic styles associated with particular professions was performed by Jones [38]. However, our approach that identified the use of professional vocabulary in posts published in an online social network seems to be an original contribution.

The study of topics from Facebook posts was performed by Wang et al. [71]. They demonstrated that women are more likely to write posts about personal themes, contrasting with men, who tend to share more public subjects, like politics and sports. Even though we study another OSN, this finding relates to the prevalence of usage of words from categories like *family*, *social* and *affection* by female users in our dataset.

An investigation on how men and women differ when designating hashtags on Twitter was carried out by Cunha et al. [20], who found that, in the context of political debate, Brazilian women are more prone to use approaches based on solidarity, while men tend to employ assertive strategies. Ottoni et al. [52] examined users' descriptions on Pinterest and showed differences in the linguistic style between genders, being women more likely to use words of fondness and affection. Schwartz et al. [64] investigated the relation between language and different variables on Facebook, and found associations between personality and language use of given groups.

2.4 Google+

An analysis of Google+ social graph is presented by the author [49], who studied structural properties of this network in comparison to other services and found different patterns of its usage across distinct countries. This analysis correspond to those presented in chapters 4 and 5. Also, the author studied linguistic characteristics among different social groups, including gender, country and occupation [21]. This analysis is also presented in this work, in chapter 6.

Schiöberg et al. [62] also conducted a characterization of the structure and the evolution of Google+, observing, too, that this OSN has a bias toward a highly-educated audience. A study on how members organize and select audiences for shared content in Google+ was conducted by Kairam et al. [39]. An interesting result is that users weigh limiting factors, like privacy, against the desire to reach a large audience. Gonzalez et al. [30] showed that, despite the recent growth of this OSN, the relative size of its largest connected component has decreased with time and that only a few users exhibit any type of activity.

Chapter 3

Methodology

In this chapter we describe key features of the Google+ service and the data collections process and the corresponding datasets.

3.1 Platform Description

The Google+ service was released in June of 2011 [34]. In the first 90 days, the service has been on field trial and only those users who received an invitation could create an account. During this time, the network grew *virally* through social contacts. In September 20th, 2011, the service became publicly open and no invitation was required for a sign up [33]. These two different mechanisms of spreading would have attracted different kinds of users to Google+. For instance, users who joined through invitations are likely tech-savvy users who typically adopt new services early, compared to the users who join through open sign-up.

In Google+, users can manage their contact list through *circles*. Circles are labeled groups of friends, which allows a user to share or receive information with and from a specified subset of his contacts. For example, a user may manage "family", "colleagues", and "alumni" circles. When a user adds someone in one of his circles, he starts to receive updates from that person. This manual grouping of contacts alleviates some of the privacy problems that existed in other "flat" social networks, where default privacy settings are set to maximize the visibility of users profile and only a small number of members change it [32]. There are two types of circles, namely in- and out-circles. The *in-circles* ("Have user in circles") of a user u represents the list of other users who added u to their circles, similar to the *followers* list in Twitter. The *out-circles* ("In user's circles") of a user u represents the list of users that u added to her circles, similar to the *friends* list in Twitter.

The circles names and their user lists are private information that only the circles creator can see. A user can identify all the others who included the user in their circles (i.e., followers), because the user receives a notification when someone adds him to a circle. Similar to Twitter, people can add other users to their circles without confirmation. This is different from networks like Orkut, where all social links are reciprocal and both sides of the users should agree to own a social link. By default, both in-circles and out-circles lists are public shown, but the user has the option to set these lists as private. An example of a user profile is shown in Figure 3.1.

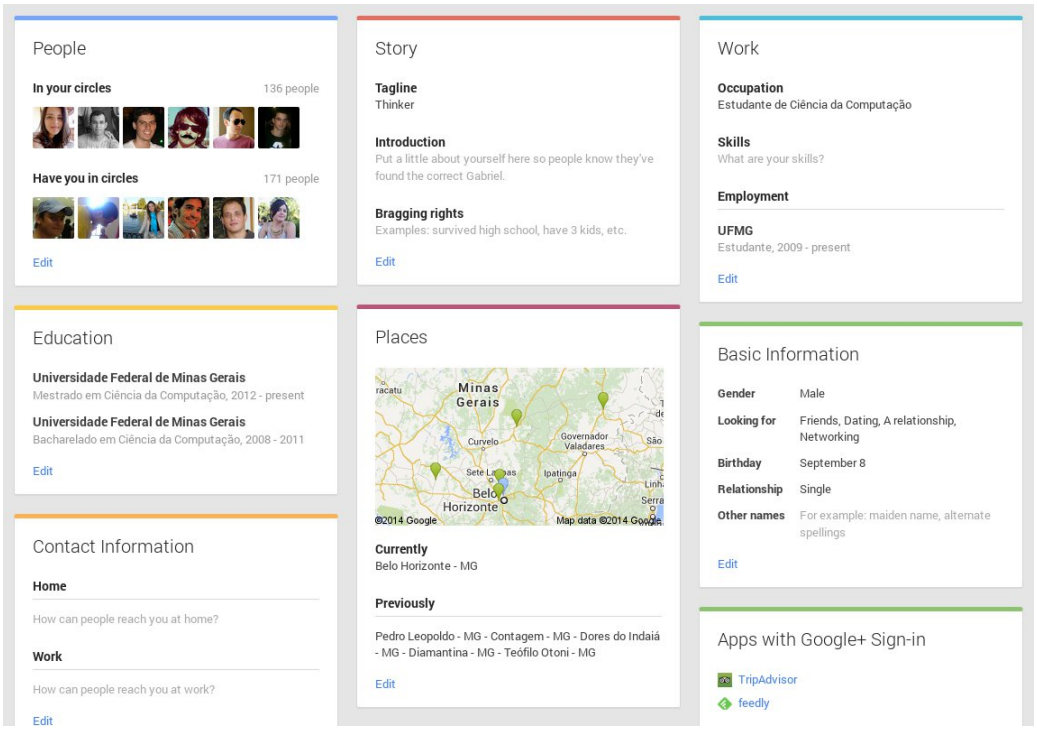

Figure 3.1. Screenshot of a Google+ user profile.

Google+ users can publish ideas (status), images, videos and any kind of URL. Whenever a user post something, she has the option to set the visibility of that content, choosing which of her circles (group of users) can see it. On the other side, a user can choose from which circles she will receive content. Therefore, circles are the way to manage information flow in Google+. The continued information flow through circles is referred to as "stream" in the system.

There are several features that allow users to interact with others. User interactions are centered around content; users can comment, share (like retweet in Twitter), and click on the "+1" button (similar to Like button in Facebook) on a given content. When a user clicks on the "+1" button, she is publicly recommending that particular content to others and it will be saved in her "+1's tab" similar to bookmark. There are

other features such as photo albums (that allow users to upload, share and organize photos), hangout (a kind of collaborative video chat with friends), and games.

3.2 Data Collection

3.2.1 System

The collection system was implemented in Python, and uses a server-client approach, where the server keep track of the user IDs to be collected, and the clients request for new IDs to be collected.

We made two crawling processes during two different periods. We used a total of 11 machines with different IP addresses to efficiently gather large amounts of data. The profile information was retrieved by making HTTP requests to publicly available user profile pages. The graph information was gathered by requesting the corresponding public circles list in the profiles of the users.

3.2.2 Approach

For the first dataset, which we call GP2011, we implemented a breadth-first search (BFS) crawler to gather new user IDs, starting the crawl with Mark Zuckerberg, the co-creator and chief executive of Facebook, because he was known to be one of the most popular users in Google+ at the time of data collection. Given that users are connected in social networks, our crawler started from a single seed node soon reached other popular users in Google+. We could not repeat the crawl with randomly chosen seed nodes, because user IDs could not be numerically generated at the time of data collection. Although the BFS technique is simple and efficient, it exhibits several well-known limitations such as the bias towards sampling high degree nodes, which may affect the degree distribution [29, 58].

For the second dataset, which we call GP2012, we inspected the *robots.txt* file and followed the sitemap to retrieve the URL's of Google+ profiles. Since we retrieved the complete list of profiles provided by Google+, we believe we have all the users with public profiles in Google+ by the time of the second data collection.

3.2.3 Datasets

The first data collection process (GP2011) started on November 11th of 2011 and ended on December 27th of 2011. In total we crawled 27,556,390 profile pages, collecting public user information and its circles lists. With the social links of the users, we

have constructed a directed graph that has 35,114,957 nodes and 575,141,097 edges. As of the data collection date, we estimated that our data set represented 56% of all registered Google+ users [2]. The lower number of profiles (27 million) in relation to the number of nodes in the graph (35 million) is due to the fact that we crawl the graph before crawling the profile information (since we discover uses through this), and only after having the list of IDs we start collection the profile information. So, between the time we discovered the user and the time we collected his profile, some users might have been deleted.

The second data collection (GP2012) ran from March 23rd of 2012 until June 1st of 2012. When inspecting the sitemap we found 193,661,503 user IDs. In total we were able to retrieve information from 160,304,954 profiles, since some IDs were deleted or we were not able to parse their information. As we did in GP2011, we created a directed graph with the circles lists of these users. The graph have 61,165,224 nodes and 1,074,088,940 edges. As mentioned before, this dataset have virtually all users in Google+. Table 3.1 presents the summary of the Google+ datasets. For this data collection we have more profiles (160 million) than nodes (61 million), because since we don't rely on the graph to discover users, a huge amount of them might not have their in and out-circles list publicly available, although having other profile information.

There is a limit on the maximum number of users that could appear in any public circle, which is 10,000 users. Since the Google+ social graph was gathered in both directions (in-circles and out-circles), we were able to recover almost all "lost edges." In order to estimate the fraction of missed links, we compared the number of users shown in their profile page with the actual number of edges we collected. Our data contained 915 users with more than 10,000 in-circles users in the first dataset, and 3,447 in the second dataset. We estimate that about 1.6% of the edges in the first dataset, and 11.4% in the second dataset, are missing because of the 10,000 limit on the circle list.

Table 3.1. Description of the two Google+ datasets.

	GP2011	GP2012
Nodes	35,114,957	61,165,224
Edges	575,141,097	1,074,088,940
Profiles	27,556,390	160,304,954
Posts	—	29,366,310

3.2.4 Posts

For the second dataset we also collet the public posts of the users. Among the 160 million profiles collected, only 8,564,462 (5%) set their posts as publicly available. We were able to retrieve up to the last ten status updates from each user's page, totaling 29,366,310 posts.

To select only messages generated in English, we used *langid.py* [47], a language identification tool that identified 20,928,557 posts probably written in this language. In order to increase the confidence that our posts are actually in near-standard English – thus avoiding the analysis of posts only partially produced in this language or written in dialects, mixed varieties or fused lects –, we additionally filtered texts with probability of at least .99 of being in English. After this restriction, we narrowed our dataset down to 7,414,679 posts. A manual evaluation of a hundred filtered posts indicated that they were indeed written in near-standard English.

Since we aimed at analyzing language characteristics of individuals, we alleviated the impact of copied posts, like chain letters and other highly replicated texts, by removing duplicated messages. We identified 265,100 types of texts that presented duplication, totaling 1,220,341 repeated posts, and removed them all from the dataset. Therefore, at this point we have 6,194,338 distinct Google+ posts.

Chapter 4

Graph Analysis

In order to characterize social relationships of Google+ users, we define a social graph. The vertices of the social graph are Google+ users present in our dataset. A user v added by user u to her circles results in $edge(u, v)$ (directed edge from u to v). Therefore the social relations among Google+ users make a directed graph $G(V, E)$, where V represents the set of users and E is the set of directed edges (u, v), $u,\ v \in V$. Given the social graph construction, we analyze two types of properties: first on the node characteristics and then on the graph structure. The former captures the characteristics of Google+ users, as defined by the fields of the user profile, while the latter represents relationships between users.

4.1 Top users

To get a sense for what users expect from the Google+ service, we first examine who the most popular and influential users are. Table 4.1 shows the top 30 users based on their in-degrees (i.e., how many users added them) and PageRank (metric of influence based on the social graph), created by joining the sets of top 20 users considering each metric. The top list of Google+ is a mix of singers, bloggers, actors, and IT professionals.

The top list is particularly different from that of Twitter in that (1) we do not see any news media outlet like the New York Times and CNN, while (2) we see founders of large Internet-based companies like Google and Facebook. In fact 5 out of the 30 users are IT related, which is uncommon in other social networks.

If we compare ranks of the in-degree and PageRank we observe that high in-degrees indeed generate high PageRanks, but not necessarily in the same order. Interestingly, the "Usher's New Look" user have relatively low in-degree (406), but still is

the third in terms of PageRank, since it is followed by the user with higher PageRank (Usher).

Table 4.1. Top 30 users in terms of in-degree and PageRank

Name	Occupation	In-degree	Rank	
			In-degree	PageRank
Britney Spears	Musician	2,046,190	1	2
Snoop Dogg	Musician	1,812,530	2	4
Larry Page	IT (Google)	1,555,474	3	6
Richard Branson	Businessman (Virgin Group)	1,433,009	4	10
Ashley Tisdale	Actress	1,409,785	5	24
Tyra Banks	Model	1,388,770	6	5
Dane Cook	Comedian	1,374,973	7	12
Tom Anderson	IT (MySpace)	1,364,270	8	21
Hugh Jackman	Actor	1,353,242	9	15
Felicia Day	Actress	1,313,685	10	27
Paris Hilton	Socialite	1,313,019	11	11
Vic Gundotra	IT (Google)	1,310,953	12	30
Trey Ratcliff	Photographer	1,280,104	13	47
Thomas Hawk	Blogger	1,277,749	14	49
Usher	Musician	1,276,174	15	1
Ron Garan	Astronaut (NASA)	1,250,329	16	40
Dolly Parton	Musician	1,242,822	17	44
Jeri Ryan	Actress	1,219,434	18	67
Muhammad Yunus	Businessman (Yunus Centre)	1,208,961	19	63
Kim Kardashian	Socialite	1,202,210	20	17
Pitbull	Musician	1,198,026	21	18
Sergey Brin	IT (Google)	1,074,711	28	16
Dalai Lama	Religious	1,072,756	29	8
Dwyane Wade	Sportsman	915,264	56	19
LeBron James	Sportsman	910,226	57	20
will.i.am	Music	876,037	71	14
Ray William Johnson	Blogger	645,352	174	9
Mark Zuckerberg	IT (Facebook)	457,682	234	7
Carmelo Anthony	Sportsman	411,051	249	13
Usher's New Look	Charity	406	164365	3

4.2 Node characteristics

We examine what kinds of interactions users perform on the network. In general, users of social networking sites reveal different types of personal information in their profile, such as basic descriptors (e.g., gender, relationship status, cities lived), contact information (e.g., e-mail, phone number, address, Web site), personal interests (e.g.,

Table 4.2. Attributes publicly available in Google+

Attribute name	GP2011 (%)	GP2012 (%)
Name	100.00	100.00
Gender	97.67	78.93
Education	27.11	11.20
Places lived	26.75	14.45
Employment	21.47	10.36
Tagline	14.79	4.33
Other profiles	13.48	7.22
Occupation	13.27	7.88
Contributor to	13.15	6.32
Introduction	7.80	5.40
Other names	4.39	3.47
Relationship	4.31	2.53
Braggin rights	3.90	2.13
Recommended links	3.63	2.02
Looking for	2.74	1.85
Work (contact)	0.22	0.08
Home (contact)	0.21	0.32

favorite TV shows, movies, books, quotes, music), education information (e.g. field of study, degree), work information (e.g., employer, position), etc.

Google+ users also publish information about themselves in their profiles. Some pieces of information are in "restricted fields", where users have to choose among some options, while in "open fields" users can write anything they want. Only the fields "relationship", "looking for", and gender are restricted fields. The rest of the fields are open fields. In the field called, places lived, a user can write the name of any place she lived and the Google+ system automatically tries to mark the place on the map.

For all the fields, except for the name that is public by default, a user can control the privacy setting and set visibility of that field. There are five options: (1) *public*, which means open to anyone in the Internet, (2) *extend circles*, which means open to people that are in circles and people that are in the circles of those, (3) *your circles*, which means open to people in one's circles, (4) *only you*, and (5) *custom*, which means a user can choose exactly which circles may view that field.

We have collected information about all the fields of users that were publicly accessible. In Table 4.2, we show fraction of users (availability) that have made each type of information available, for both datasets. We observe that the availability has decreased for all the fields but Home Contact.

4.3 Privacy concerns

Studies on human behavior [67] show that individuals with profiles on social networking sites may have greater risk taking attitudes (such as sharing private information) than those who do not use OSNs. Sharing contact information, such telephone numbers, may increases risks [37]. As far as contact details are concerned, the work in [67] shows how many Facebook users disclose identity information in the form of contact details. The majority of the users in the sample used by the study publicly showed their e-mail address (64.1%). Only a few Facebook members published their mobile phone number (10.7%). Similarly, only a minority (10.7%) of the participants revealed their home address on Facebook.

Google+ allows their users to publish contact information in their profiles. Some users publicly share their work or home contact information. In our data set, a total of 72,736 users share telephone number in Google+, which represent 0.26% of the population. We call these users *tel-users* and because they represent a class of risk taking users we look into the details of the profile of these users. We do not take into consideration the kind of profile, i.e. we do not distinguish personal, professional or business accounts. For this analysis we show results only for the dataset GP2011.

In order to examine how much information *tel-users* share publicly compared to all users, we show the CCDF (Complementary Cumulative Distribution Function) of the number of fields in the profile shared for each user in Figure 4.1, removing the fields of Home and Work information from the contabilization. (The list of the fields available are given in Table 4.2.) As we can see, *tel-users* generally share more information in their profiles than other Google+ users, which confirm their risk taking attitude. For example, 10% of all Google+ users share more than six fields, while 66% of the *tel-users* do the same.

Concerning the information sharing behavior of Google+ users, table 4.3 displays the percentage of users who give information about gender, relationship, and location for all users and *tel-users*, considering only those users that had the field public. Among all users of the dataset, 68% are male and 31% are female. However, the difference is much higher when we consider *tel-users*; 86% are male and 11% female, indicating that female Google+ users are less likely to share phone numbers than male Google+ users. Similar to the observations confirmed in [26], more risk taking behaviors can be found for men and greater concern from women with regard to information provided on the Web.

What is particular about Google+ is that it asks users to provide a very detailed level of information about their relationship status as opposed to other social networks.

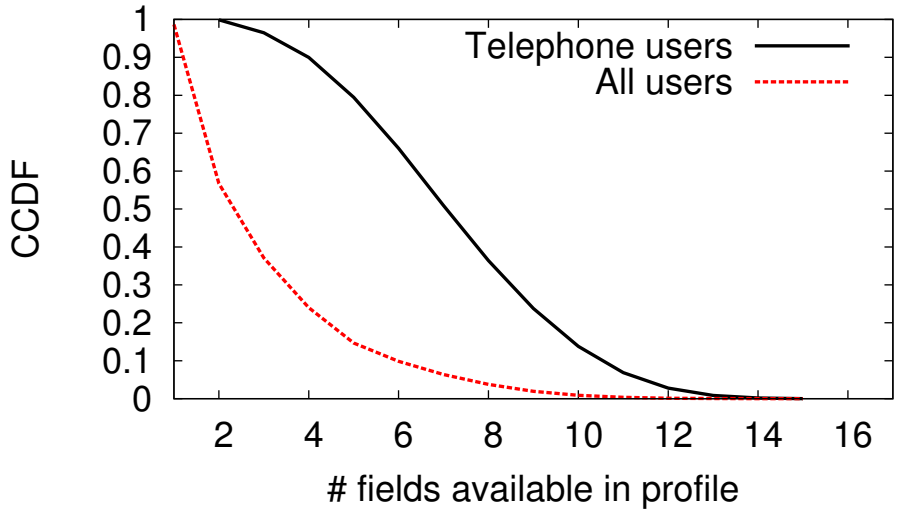

Figure 4.1. Number of fields shared by users in the profile.

The nine default options from which users can choose from are listed in the table. Conducting the same comparison of all users and *tel-users* over the relationship status, we find that user behaviors are similar between the two groups. However, those users who set their relationship status as "single", "it's complicated", "in an open relationship", "widowed", and "in a civil union" were more likely to share their phone numbers publicly than others. In particular, we saw a high percentage of single users (57.24%) compared to all the users (42.82%). In contrast, only half of the users "in a relationship" shared their phone numbers. This is expected if we consider that single users might be showing their numbers looking forward receiving calls from interested partners.

The fraction of *tel-users* does not follow the rank of the top 10 countries in Figure 5.1. While the US take up 31.38% of all users, it counts for only 8.92% of those users who have made their phone numbers available in Google+. In contrast, India now becomes the most populated country based on the fraction of *tel-users* count (31.90%). The fraction of Indian users in the *tel-users* group is twice as big as in all other country users group. This result is interesting, considering that there is a study [37] that specifically collected Indian users' phone numbers and actually called them, to investigate their reasons and awareness. They found out that few users did not know about the presence of their mobile number, and while some put it for promoting their business, other users put it for emergency protection.

While the different level at which users of a given country reveal their phone numbers is interesting, this may come as no surprise when we account for the fact that people's perception of what is "private" is different. According to a report in [13],

Table 4.3. Information shared by all users and *tel-users*.

	All users	Tel-users
Total	27,556,390	72,736
Gender (N)	26,914,758	71,267
Male	67.65%	85.99%
Female	31.46%	11.26%
Other	0.89%	2.75%
Relationship (N)	1,186,903	29,068
Single	42.82%	57.24%
Married	26.59%	21.03%
In a relationship	19.80%	10.23%
It's complicated	3.16%	3.98%
Engaged	4.39%	2.98%
In an open relationship	1.26%	2.77%
Widowed	0.50%	0.58%
In a domestic partnership	1.08%	0.77%
In a civil union	0.39%	0.41%
Location (N)	6,621,644	45,676
United States	31.38%	8.92%
India	16.71%	31.90%
Brazil	5.76%	4.72%
United Kingdom	3.35%	2.19%
Canada	2.30%	1.52%
Other	40.50%	50.77%

65% of people in Germany find mobile phone number as personal, whereas only 28% of people in Romania think the same.

4.4 Graph Structural Characteristics

We next present characteristics of the Google+ social graph. For each network metric, we also show the results for the Twitter graph using a dataset from other research [11] for comparison. Besides, we present a comparison of the metrics between timestamps. Comparing Google+ network metrics with Twitter makes sense, since both networks have a directed social graph, different from Orkut or Facebook (before 2012), where the connections between users are always reciprocal. Since the metrics for Google+ does not present huge difference between datasets we show in the plots only the distribution for GP2012 when comparing with Twitter.

4.4.1 Degree Distribution

One of the most common structural measures analyzed in complex networks such as the Google+ social graph is the distribution of the number of the incoming and outgoing node connections or what is so called "degree". Figure 4.2 shows the CCDF for the variables out-degree and in-degree of the Google+ social graph. We can see that these curves have approximately the shape of a Power Law distribution. The CCDF of a Power Law distribution is given by $Cx^{-\alpha}$, $x, \alpha, C > 0$. If we compare the curves with Twitter, we observe similar patterns, altough Google+ shows slightly lower degrees.

By using a simple statistical linear regression (in the log-log scale) we estimated the exponent α that best models the data. We obtained $\alpha = 1.09, C = 1.36$ (with $R^2 = 0.97$) for in-degree, and $\alpha = 1.34, C = 12.92$ (with $R^2 = 0.89$) for out-degree (considering only values lower or equal to 5,000). The out-degree curve of Google+ drops sharply around 5000. We conjecture this is because Google maintains a policy that allows only some special users to outpass a specified threshold (unknown) and add more than 5000 friends to their circles.

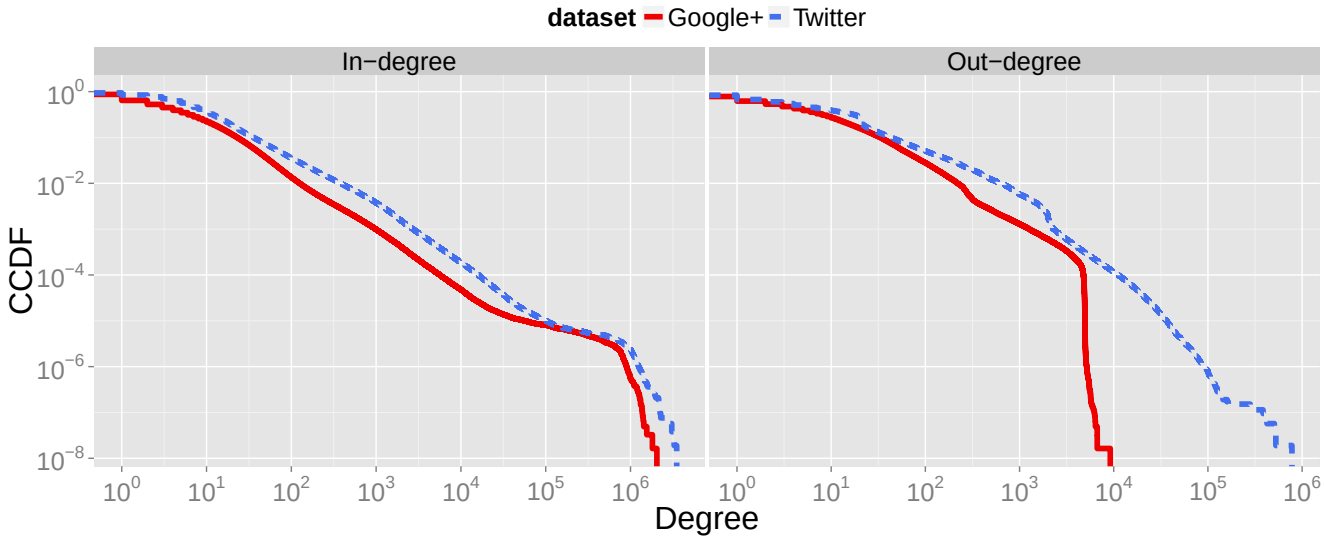

Figure 4.2. Distribution of in-degree and out-degree.

The overall power-law trend shown in the degree distribution is characteristics of the human social networks. This means that a small fraction of the individuals have disproportionately large number of neighbors, while most users have a small number of neighbors. As studied in many other research, hubs play a central role in information propagation in social networks.

4.4.2 Reciprocity

In order to evaluate discrepancies between in- and out-degrees for a given node we use a simple metric called Relation Reciprocity (RR) of a node $u \in V$ as:

$$RR(u) = \frac{|OS(u) \cap IS(u)|}{|OS(u)|} \tag{4.1}$$

where $OS(u)$ is the set of nodes (i.e, users) that have an incoming edge from u and $IS(u)$ is the set of nodes with outgoing edges pointing to u.

Figure 4.3 shows the distribution of the Relation Reciprocity. This metric is able to effectively differentiate very popular users, such as celebrities and companies, with very low reciprocity from ordinary users, that have moderate to high RR.

The analysis of the relation reciprocity in circles links suggests a strong signature of the Google+ users. Nearly 50% of the users have RR higher than 0.6, which shows some sort of structural balance between the users of this new social networking service, while in Twitter only 20% have RR higher than 0.6. This concept is related to the fraction of reciprocal relationships in a user level.

We also calculate the percentage of *global* reciprocal relations, calculated by measuring the percentage of edges in the complete graph that have the corresponding reverse edge. We find 20% in Google+, compared to 22.1% reported for Twitter [41]. This indicates that Google+ has a similar reciprocity in a global level, although having a lower reciprocity in a local per-user level (average of 0.52 compared to 0.26 in Twitter).

The high reciprocity rate in Google+ may be related to the scarcity of large media outlet profiles, since by the time of our collection the "pages" feature was new and not very popular. Media outlet attract very large numbers of followers, but do not exhibit followees. Different kinds of online sharing services exhibit higher reciprocal relations, such as 68% for Flickr [12] and 84% for Yahoo! 360 [40].

4.4.3 Clustering Coefficient

Another common characteristic of social networks is a high average clustering coefficient (CC). The CC of a node u, denoted by $C(u)$, is defined as the probability of any two of its neighbors being neighbors themselves [72], or the fraction of pairs of u's friends that are connected to each other by edges [23]. This metric is associated to the number of triangles that contain a node u. For a directed graph, we consider neighbors (or friends) as being the reciprocal friends, defined as $RS(u) = OS(u) \cap IS(u)$. the maximum number of triangles connecting the $|RS(u)|$ reciprocal neighbors of u is

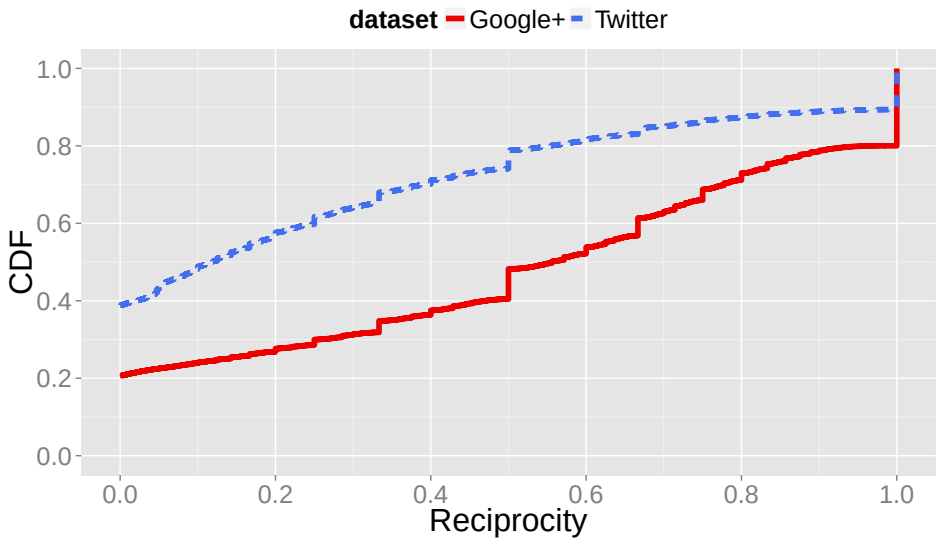

Figure 4.3. Distribution of reciprocal links.

$|RS(u)|(|RS(u)|-1)$. Thus, the CC measures the ratio between actual triangles and their maximal value. During clustering coefficient analysis we only consider the nodes with $|RS(u)| > 1$, since this is a necessary condition for this computation. We also computed the Clustering Coefficient considering the neighborhood as the outgoing neighbors, and the results are the same.

Figure 4.4 shows the distribution for the CC of all nodes in the social graph. We can see that 50% of all users have a CC greater than 0.2. An approximate calculation, based on the results presented for Facebook in [69], allows us to estimate the clustering coefficient for part of Facebook population. In [69], we have that only users with degree smaller than 50 have an average CC greater than this value. However, these users represent less than 1% of the entire network [69], suggesting that Google+ has a higher average cluster coefficient than Facebook, which represent a more tightly connected network. Comparing with Twitter (as shown in Figure 4.4), we can also see higher values of CC in Google+.

4.4.4 Strongly Connected Component

The study of the connected components of a social graph is a key factor to understand its structural properties. For example, if we know the WCCs (Weakly Connected Components) of a graph then we have information about the number of isolated nodes in the network as well if it has a giant component.

In order to investigate the connectivity of G we decided to measure the number

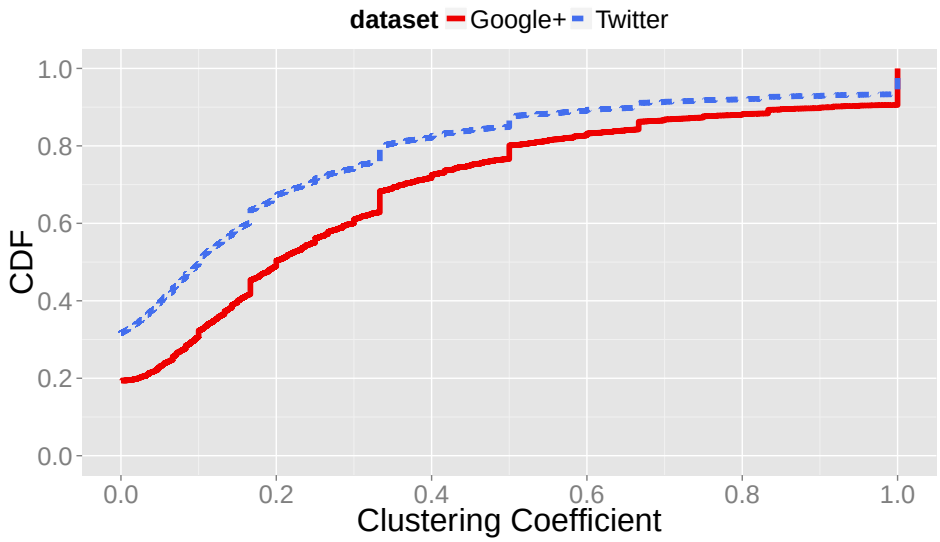

Figure 4.4. Distribution of Clustering Coefficient.

and size of all its Strongly Connected Components (SCC). A strongly connected component of a social graph (directed) is a subgraph, such that a node can be reached from any other node following edges between them. SCCs have an important role in directed social networks (like Google+) because they are central to information dissemination to the users that are part of the them. Graphs with large SCCs are amenable to quick information dissemination processes.

We identified 22,874,247 SCCs in G. To reach this number we used a procedure involving two Depth First Searches [16]. Figure 4.5 presents the CCDF of the size of all SCCs found in G. In this figure we can see that almost all of them are small. In fact, there is only one with more than 200 nodes, which is the SCC with 37,012,901, that means that G has a giant component and the graph we collected is highly connected.

4.4.5 Degrees of Separation

The degree of separation essentially describes the shortest possible routes between two nodes of our graph. Although the degree of separation has been commonly thought of in the social context, the concept has many applications in social networking such as information dissemination and friend recommendation [23]. We present an analysis of how many hops there are between two users in the Google+ social graph. In order to have the exact distribution we would need to compute the shortest path from all nodes to all nodes of the network. Due to the computational cost of this task we decided to use a random sampling procedure [1]. We sampled k different users and for each one

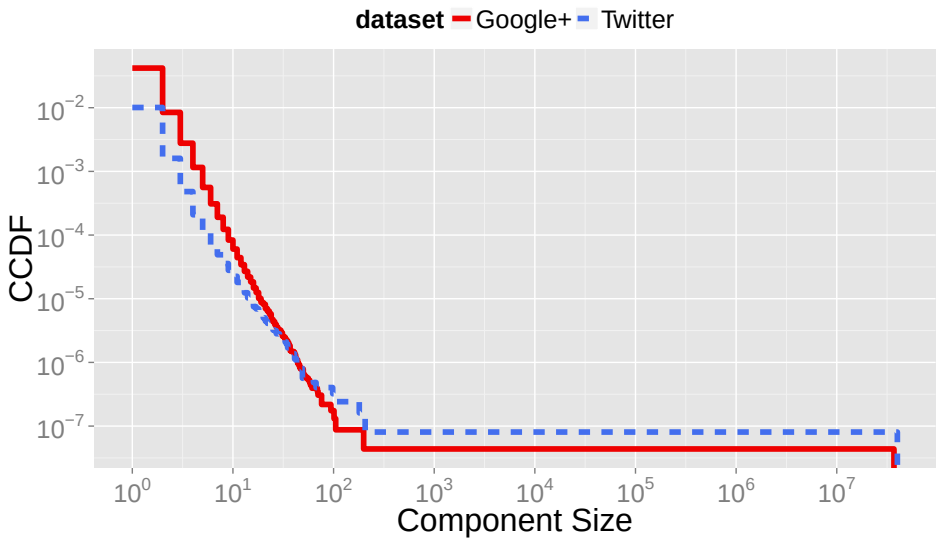

Figure 4.5. Distribution of the size of the strongly connected components.

of them we computed the shortest path to all others users in the network. We started with $k = 2000$ and increased its until 10000, stopping in this value once there were no more changes in the distribution.

Figure 4.6 presents the final estimate of the path length distribution for two cases: the directed graph G and its undirected version. In the first case we can see that the most common value is 6 with an average of 6.0. In the second we have 5 as the most common and an average of 4.8. The graph G has a diameter of 21 and for its undirected version 18. This means that most users are only a few hops away from a random user, which has the important implication that information can spread quickly and widely throughout the network.

Although we have found the same mode of the well-known study of Milgram [50], it is important to remark that we are analyzing the public graph of Google+. So, adding back the edges omitted by users due to privacy constraints may reduce the average path length further. Comparing with other networks, we refer to Twitter with a mode of 4 and average of 4.12 [41], Facebook with an average of 4.74 hops [3] and a median of 6 for the MSN messenger network [44].

4.4.6 Evolution

We now compare the distribution of the metrics between both Google+ datasets to analyze how the network evolved through time. Figure 4.7 shows the plots for all the network metrics discussed before in this section. We observe that, in general, the

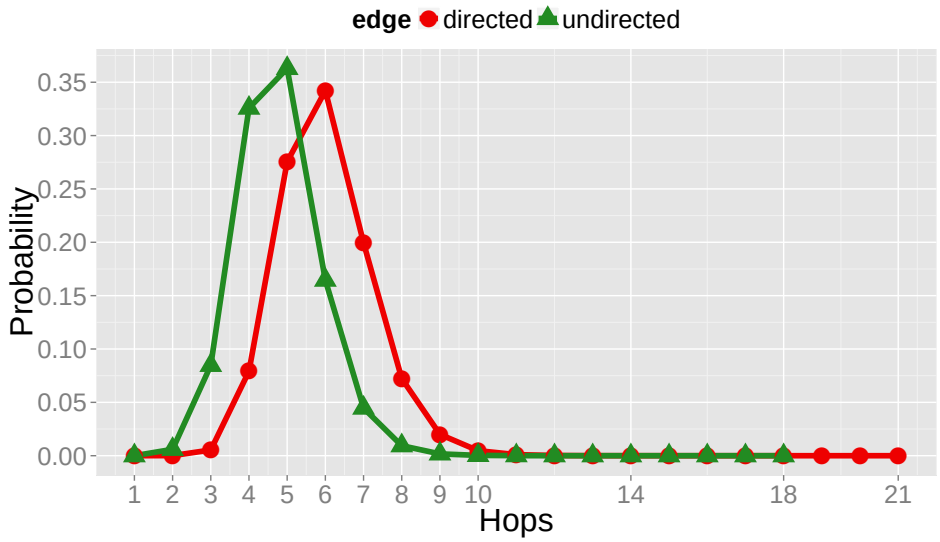

Figure 4.6. Distribution of path length (estimated).

trend (shape) of the distributions are very similar, while slightly skewed, indicating an increase or decrease of the overall value of the metric. For instance, the degrees had a minor increase, which is expected, since the network is relatively new and is still growing. Consequently, the clustering coefficient and reciprocity decreased, since having more connections increases the number of connections neededed for having the same value of these metrics when having a lower degree. When comparing the strongly connected components, we see a natural increase both in the number of components (9.8M to 22.9M) and in the size of the biggest component (25.2M to 37.0M), since the number of users in the second dataset is higher. Analyzing the path length, we observe a very similar distribution, with a slight increase of the average path length (5.9 to 6.0).

4.4.7 Summary

Table 4.4 summarizes the key structural features of Google+ and three others important OSNs in order to conclude this section. Statistics on other social networks are borrowed from [41, 3, 69, 51]. We can see some important differences, for example the Google+ social graph has a higher average path length. Its diameter is comparable to Twitter, but smaller than Facebook. Moreover, we can see that the number of friends (both in- and out-degrees) are much smaller when compared to Facebook.

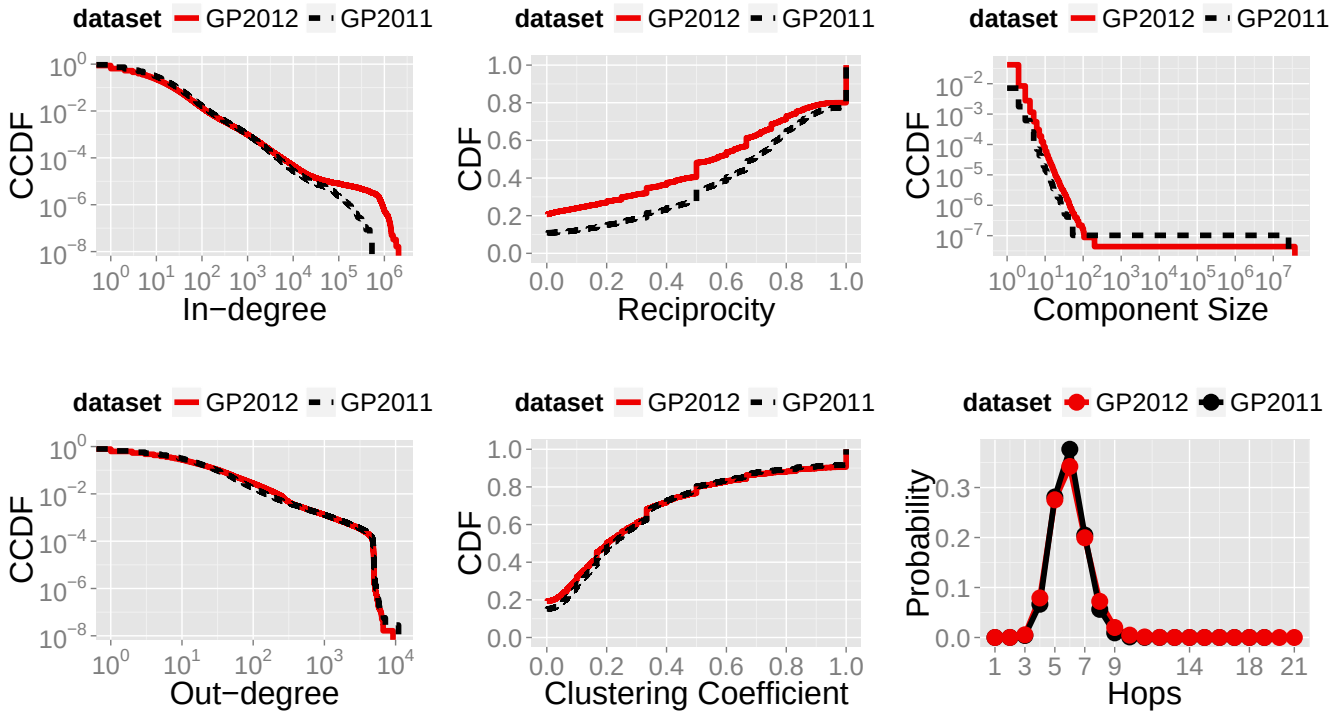

Figure 4.7. Comparison of the network metrics between the Google+ datasets.

Table 4.4. Comparison of topological characteristics of Google+ and other online social networks.

	Google+ 1	Google+ 2	Twitter	Facebook	Orkut
Nodes	35M	61M	41.7M	721M	3M
Edges	575M	1 Bi	106M	62 Bi	223M
% Crawled	56%	100%	100%	100%	11%
Path length	5.9	6.0	4.1	4.7	4.3
Reciprocity	32%	20%	22%	100%	100%
Diameter	19	21	18	41	9
In-degree	16.4	17.6	28.19	190.2	—
Out-degree	16.4	17.6	29.34	190.2	—

Chapter 5

Patterns across geo-locations

Google+ users can list all the places they have lived at a field in their profile called "Places lived" which is incorporated to the Google Map for visualization. Nearly 27% of the users in our dataset provide geo-location information. This feature is unique in Google+, for other social networks like Facebook only allows users to list their current location and, at most, their hometown information. Using the places lived field, we analyzed how Google+ users are distributed around the world. For this, we first extracted the coordinates of the last location from the places lived field for each user and translated the coordinates into a valid country identifier. In this fashion, we were able to identify the country of 6,621,644 users. In this chapter we use information from the GP2011 dataset.

5.1 Popularity

Figure 5.1 shows the top 10 countries in our dataset with their respective percentages of the registered users.[1] More than 30% of the users who share their location information are identified as living in the US. We observe Google+ is relatively popular in India and Brazil, which are also two of the countries with high presence in Orkut, the other social network from Google [31]. United Kingdom, Canada, and Germany also appear in the list, which are countries known to have high Internet Penetration Rate (IPR) or the percentage of Internet users out of the population of that country.[2] Interestingly, countries like Indonesia and Mexico appear in our top list, which are not part of the top countries based on the Internet penetration rate.

[1]County codes represent the following. US: United States; IN: India; BR: Brazil; GB: United Kingdom; CA: Canada; DE: Germany; ID: Indonesia; MX: Mexico; IT: Italy; ES: Spain.

[2]Statistics about the Internet penetration rate and population were obtained from http://www.internetworldstats.com

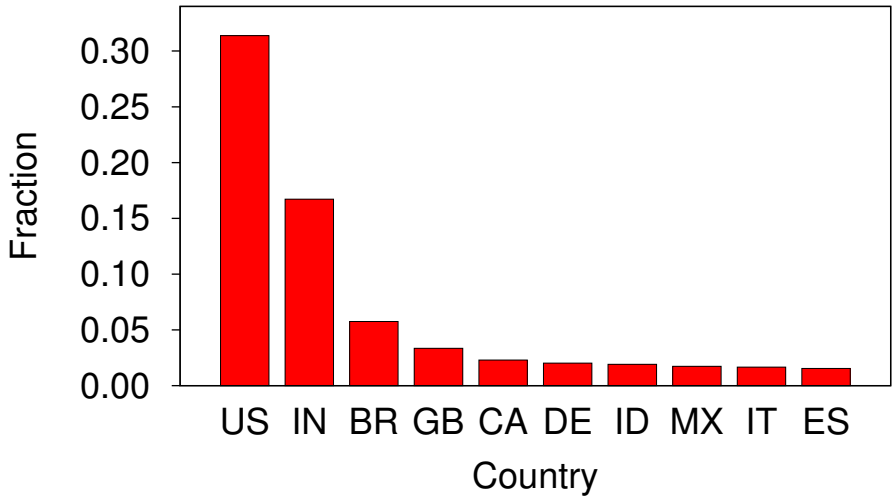

Figure 5.1. Top 10 countries with Google+ users.

5.2 Economics

Intrigued by the unique mix of countries based on Internet penetration rate, we further investigate which countries have high percentage of their Internet population on Google+. To do that we define *Google+ Penetration Rate*, that can be computed for each country C as follows:

$$GPR = \frac{\text{number of users in our dataset living in } C}{\text{Internet population of } C}. \tag{5.1}$$

Note that our measure is meaningful only for the relative ranking of different countries, because our data is a sample taken from Google+ and in the sample only 27% of the users provide geo-location information.

Figure 5.2 shows the Google+ penetration rate for the top 20 countries. The top country in Google+ adoption now becomes India. We also see that countries like Taiwan and Thailand appear in the top ten list. For comparison, we show the Internet penetration rate of the same top 20 countries in Figure 5.3. The top five countries of Internet penetration are United Kingdom, Germany, Canada, Japan, and Australia. Both of the figures have Gross Domestic Product (GDP) per capita in the X-axis.

We make several observations. First, while there is a linear relationship between the GDP per capita of a country and its Internet penetration rate , we do not see the same trend in Google+ penetration rate. Countries with lower GDP per capita like Brazil, Mexico, and Thailand have equal footing in the penetration rate as with much wealthier countries such as United Kingdom, Australia, and Canada.

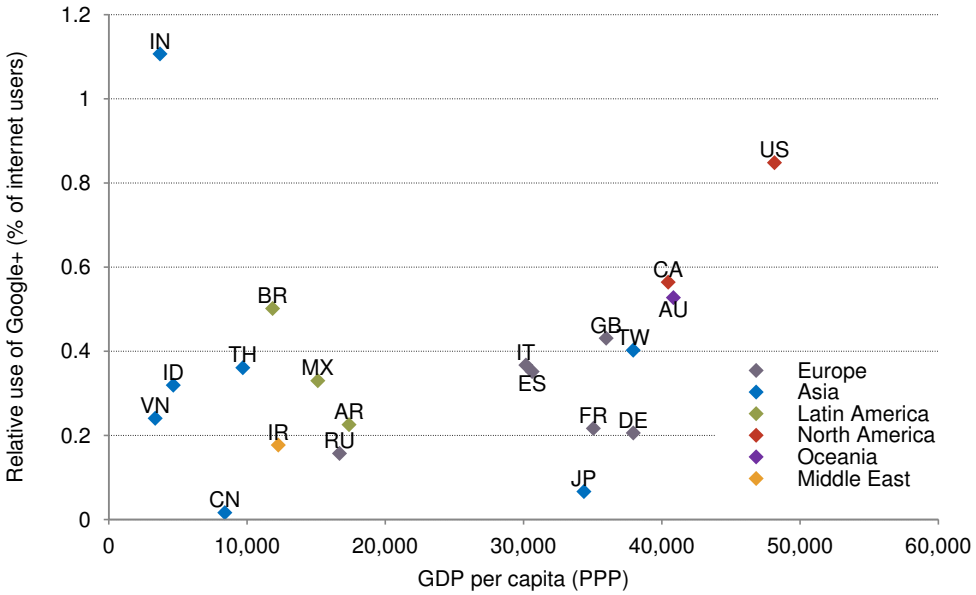

Figure 5.2. GDP Per Capita and Use of Google+.

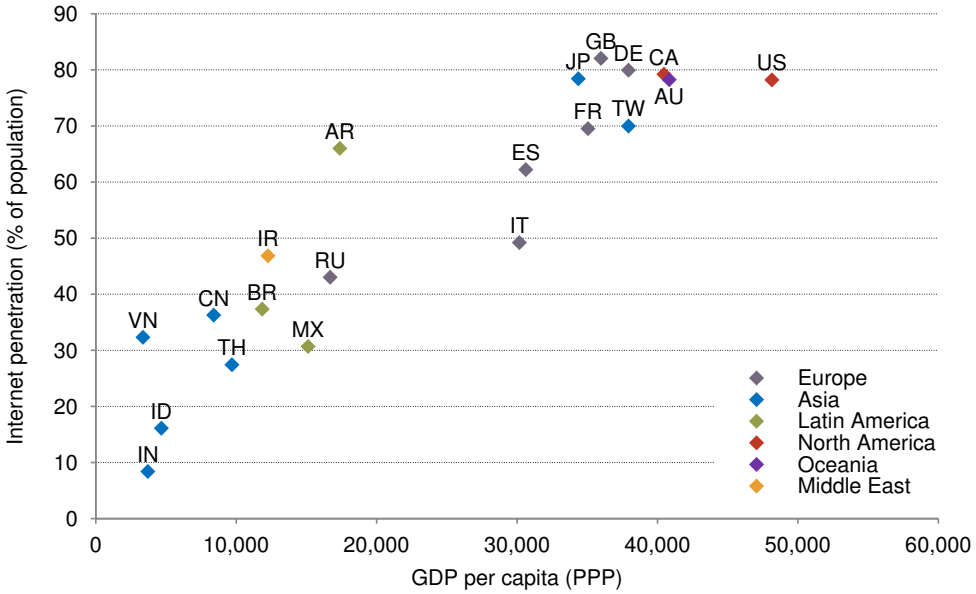

Figure 5.3. GDP Per Capita and Internet Penetration.

Second, certain countries showed a large gap between the Internet and Google+ penetration rate such as Japan, Russia, and China. In both of these countries, domestic social networks like Mixi in Japan, Odnoklassniki in Russia, and QQ in China are widely used. International social networking services like Facebook and Twitter are known to have little presence in these countries. Also in case of China, international social networking sites have been blocked [27].

Third, countries with lower Internet penetration like India and Brazil had a very high Google+ penetration rate. As we mentioned before, these two countries are known to favor Google's other social network, Orkut. It is possible that users in India and Brazil are familiar with the Google product, hence are more likely to adopt the service quickly than other countries.

5.3 User Occupation

Among various statistics we examined, the occupation-job title of the top users clearly distinguished Google+ from other well-known social networks, as we examined in Table 2. Interestingly, the top occupations also varied slightly across different countries. Table 5.1 shows the occupation-job title of the 10 most connected users in each of the top 10 countries, based on their in-degree (i.e., how many circles these users are added to by others). The number of users analyzed is limited because a manual inspection and classification of the occupation is needed.

At a glance, the top list of Google+ is a mix of singers, bloggers, actors, and IT (i.e., Information Technology) professionals. When we compare the list to that of Twitter [41], the top list is particularly different in that we do not see any news media outlet like the New York Times and CNN, while we see founders of large Internet-based companies like Google and Facebook. In fact, five out of the top 30 global users were IT related in Google+, which is uncommon in other social networks.

In the table, we also show the Jaccard index, used to compare the similarity and diversity of occupations in these country when compared to occupation-job titles in US. The top users in Canada have a very similar profile to that of the United States. Furthermore, the US, Canada, UK, and India share several top professions, which we may be due to the common British colonization. In contrast, Brazil, Italy, and Spain show a different set of celebrities and professions, and is worth noting that these three countries are Latin cultures, different from anglo-saxon cultures (US, CA, GB).

The top countries have very different kinds of popular users. IT professionals are popular in Google+. In Brazil, there are no famous IT related public figures, hence the list is dominated by comedians and bloggers. In Mexico, half of the top users are related to music. Italy is the country with more journalists among top users, 4 in total. Spain is the only country having Politicians in the top 10 user list. These lists suggest that each country has a different pattern of utilization of the information network provided by Google+, because the occupations of the top individuals represent what a typical user expect from Google+.

Table 5.1. Occupation-Job Title of the top users.

Country	Profession codes* of the top-10 users										Jaccard
United States	Co	Mu	IT	Mu	IT	Mu	Bu	IT	Mo	Ac	1.00
India	Mu	So	IT	Mu	Mo	Mo	IT	Bu	IT	Mu	0.57
Brazil	Co	TV	Jo	Wr	Ar	Bl	Bl	Co	Mu	Co	0.18
United Kingdom	Bu	Mu	IT	IT	Mu	Mu	IT	Mo	So	IT	0.57
Canada	IT	IT	Mu	Co	Bu	Ac	IT	Mu	Co	Ac	0.83
Germany	Bl	IT	IT	Jo	Bl	IT	Jo	Ec	Mu	Bl	0.22
Indonesia	Mu	IT	So	Mo	Mo	IT	Mu	Ec	Ph	Jo	0.30
Mexico	Mu	Mu	Mu	IT	Mu	Bl	Bl	Mu	Ac	Jo	0.33
Italy	Jo	Jo	IT	IT	Jo	IT	Jo	Mu	Mu	IT	0.29
Spain	Jo	Po	Po	IT	Mu	Mu	IT	Mu	Po	IT	0.25

* Co: Comedian; Mu: Musician; IT: Information Technology Person; Bu: Businessman; Mo: Model; Ac: Actor; So: Socialite; TV: Television Host; Jo: Journalist; Bl: Blogger; Ec: Economist; Ar: Artist; Po: Politician; Ph: Photographer; Wr: Writer

5.4 Openness

The notion of privacy is an individual characteristic. What is considered private information for a person might not be for the other. This notion can be influenced by different factors, such as age, gender and culture. In this section we want to analyze if such a difference exist among countries. We examine how the 10 countries differ in the notion of privacy, by looking at the number of different types of information publicly shared by users in their profiles (e.g., name, gender, education, occupation). As mentioned earlier, the name field is mandatory. Also, because of our methodology to utilize geo-location, all of the sample users studied in this section have shared "places lived" field. Therefore, the minimum number of shared fields is 2.

Figure 5.4 shows the CCDF of the number of fields users of the top 10 countries shares in their profiles. We present the X-axis in the range 2-14 for better visualization. First, compared to the "all users" distribution in Figure 4.1 we observe higher amount of shared information for all the countries. Since the users considered in Figure 5.4 are, by definition, users that share the places lived field, it is expected that they also share other fields more than the average user.

We observe that, although the difference between the countries is not very pronounced, the ranking is slightly different. Indonesia and Mexico share more information than other more popular countries like United States and United Kingdom. Germany is the most conservative when it comes to sharing personal information; it was the only country having less than 10% of the users sharing more than 12 fields and also the only

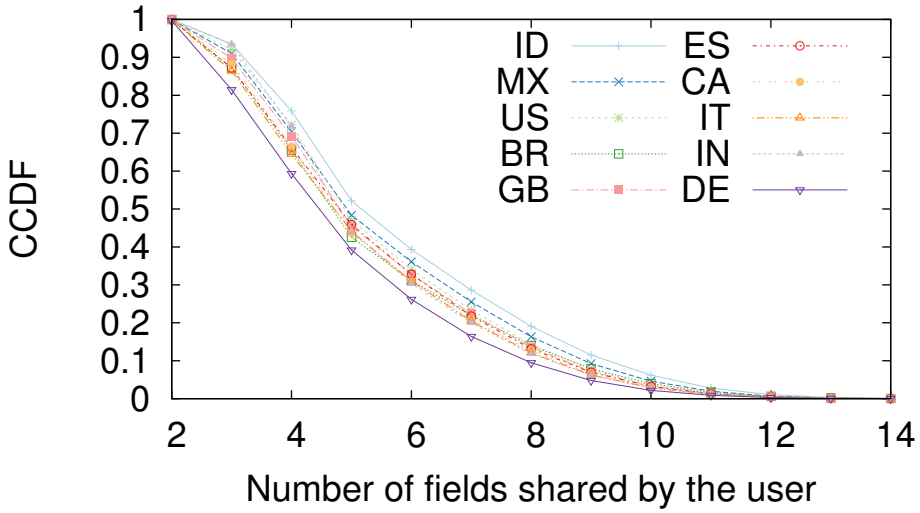

Figure 5.4. Number of fields in profiles in each country.

country having less than 30% of the users sharing more than 10 fields.

5.5 Average Path Miles

We now investigate the relationship between social network structure and geographical properties. We start by answering the following question: *is the geographical location of users an important factor in the formation of social links?* To understand if the distance has some influence on the formation of social links as described in circles, we estimated the physical distance of pairs of users in three cases: (1) every pair of socially connected users (approximately 60 million pairs), (2) pairs of reciprocally connected users (approximately 13 million pairs) and (3) randomly chosen pairs of users (20 million, not linked by a social relation). We then computed the physical distance between them. It is important to remark that we conducted this analysis only for users that share geo-location information, which represents 26.75% of the crawled Google+ network.

Figure 5.5 shows the cumulative distribution on the expected physical distance—which we call the *path mile*, similar to the notion of the path length—between pairs of circle friends and random user pairs in Google+. The friendship links in Google+ have higher geographical proximity than a random pairs of users. Nearly 58% of the users (friends) were separated by less than a thousand miles and 15% of them were separated by in fact 10 miles. This observation reinforces the high chance that the Google+ network largely capture the offline social relationships among users. As expected,

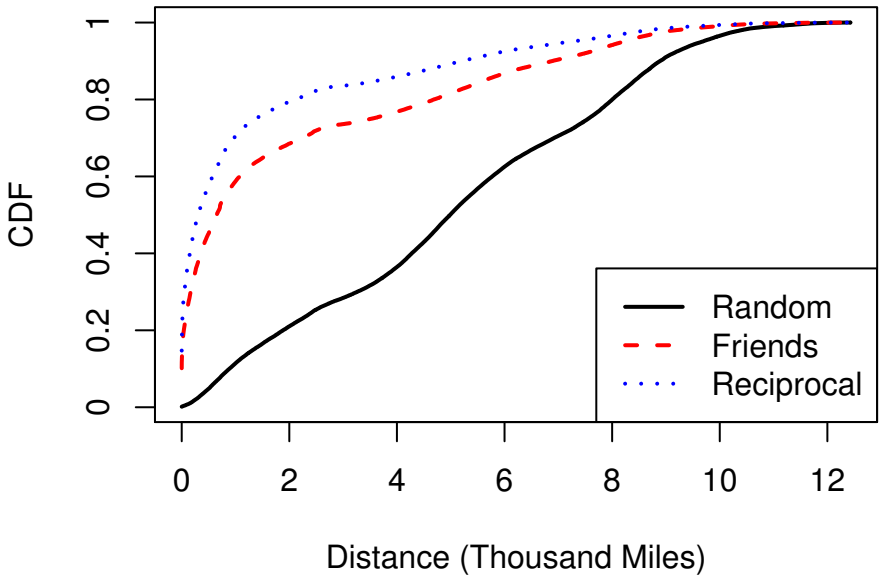

Figure 5.5. Path Mile distribution of Google+ users.

users with symmetric links (reciprocal) live closer than those with asymmetrical links, indicating the influence of physical distance on the intensity of the relationship.

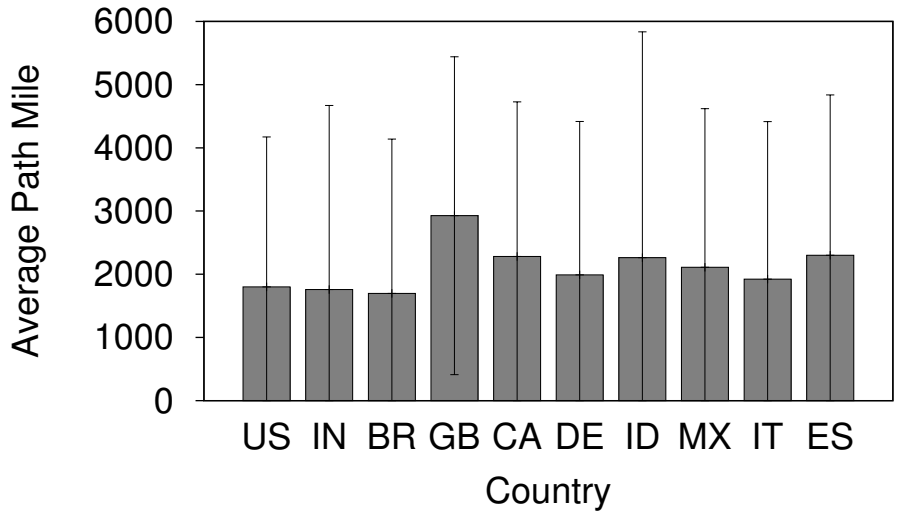

Figure 5.6. Average path mile with standard deviation.

One natural question that arises from this result is whether it depends, or not, on the country size. For example, do geographically large countries like US have a

higher average path mile than small countries like Italy? If this is the case, large countries should have better investment in content distribution in order to minimize jitter and delay especially for the delivery of user generated videos. Figure 5.6 shows the average path miles along with the standard deviation error bar for the top 10 countries. Contrary to our expectation, there is no specific pattern relating the size of the country and its average path mile. One possible explanation could be that small countries have a considerable fraction of edges going outside the country. In fact, this result is discussed in next section.

5.6 Social links across geography

The final question we ask is about the impact of country on friendship link formation. In particular, we ask: are users in the same country more likely to be friends in Google+ than users in different countries? To answer this question, we constructed a graph of countries, where each node is represented by one of the top 10 countries and the weight of each directed edge is given by the proportion of outgoing links from one country to another. Self-loop edges hence would represent the fraction of friendship links that bind two users in the same country.

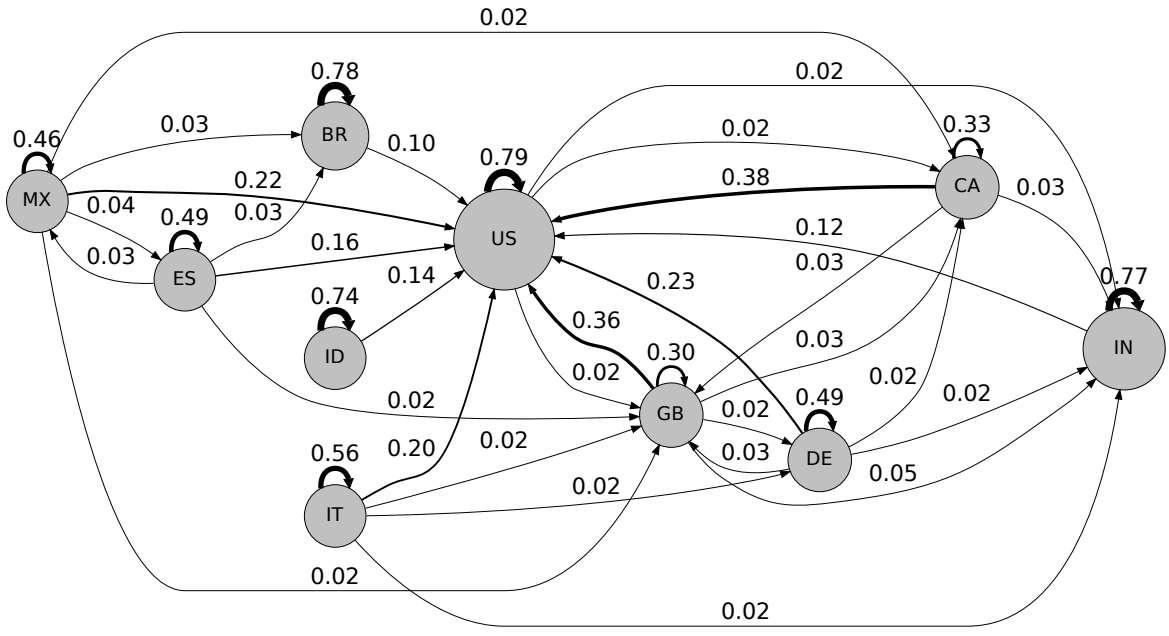

Figure 5.7. Link distribution across the top countries.

Figure 5.7 shows the visualization of the links across the top 10 countries of

Google+. The size of each node is normalized to represent the proportion of Google+ users in the associated country and the thickness of each edge is proportional to its weight. Edges with weight smaller than 0.01 were omitted to improve visibility. With this result we find that US has an important role in the overall landscape of Google+, as seen from the dominant influx of edges from most countries to the US. Moreover, the US is a node with low reciprocity, which means that there are a significant number of people of other places adding people in the US to their circles while those in the US, in general, prefer to form friendships among themselves.

Highly populated countries like Brazil, India, and Indonesia (and the US as already mentioned) tend to have a high weight in the self-loop edges. For the remaining six countries, the proportion of self-loops is much smaller. In particular, only 30% of the links are self-loops in United Kingdom and 33% in Canada. These two countries, as a result, have a large number of out-going edges to the US, which might be explained by geographical proximity and cultural similarity (e.g., sharing the same spoken language).

It is also worth noticing that in Figure 5.7, the countries that exhibit self-loop edges grater than 0.50 are those that do not have English as their first languages, which are Indonesia, India, Brazil, Italy. Perhaps because of its economical and technological leadership, the US also exhibits a high degree of self-loop edges. This indicates the language barrier in the set up of cross-national social relationships. Furthermore, this also means that the nature of language and geography will introduce interesting opportunity for growth strategies (e.g., advertisement of Google+ in a non-English speaking country will likely show a similar organic growth pattern with many national links).

The average path mile discussed earlier could mean that content distribution in Google+ faces similar challenges for both small and large countries. In fact, smaller countries like United Kingdom may require more sophisticated measures to reduce delay in content delivery, as seen from its high average path mile. Furthermore, we see varying patterns of link formation across different countries. When it comes to building recommender systems, it may make sense to recommend domestic users and their content for those countries that have high degree of self-loop such as Brazil and India. However, it may be of more interest to the users to recommend foreign users and content to those in Germany and United Kingdom due to their low fraction of self-loops.

Chapter 6

Linguistics

In this chapter, we present the linguistic analyses performed on Google+ posts. They are all independent investigations, not necessarily examining the same text attributes, which makes it possible to test distinct aspects of language behavior. It is important to note that the results presented here apply only to language behavior in the specific context of Google+ and may not be valid for offline environments or even other online social networking systems. In this chapter we use data from the GP2012 dataset.

6.1 Basic characterization

6.1.1 Activity

Among the more than 160 million users with profiles collected, only 8,564,462 set their status updates as publicly available. We were able to retrieve up to the last ten shared contents from each user's page, totaling 29,366,310 posts. The fact that our collection has only up to ten posts from each user enables the results not to be influenced by idiosyncratic behaviors of very active members. Figure 6.1 depicts the distribution of the number of posts per user in our dataset and shows that most users have published only a few posts. Nonetheless, we have no information about the real amount of content published by 1,258,684 members who posted at least ten messages.

6.1.2 Language

Our work is focused on the analysis of posts written in English. To identify the language of a post we used `langid.py`[1], a language identification solution that provides the

[1] `https://github.com/saffsd/langid.py`

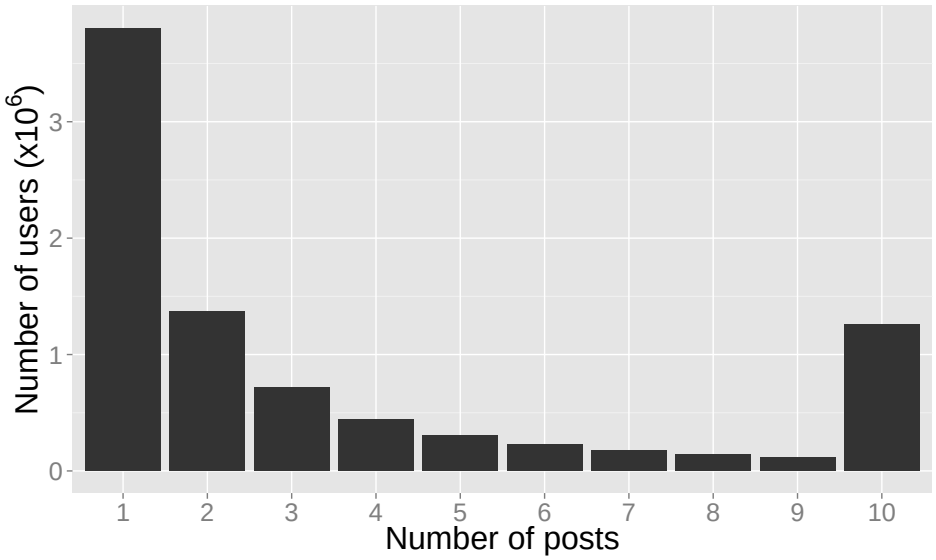

Figure 6.1. Number of posts per user in our dataset.

probability of an evaluated text being in a particular language, working well for both long and short documents, including microblogs [47]. Figure 6.2 shows the number of posts per language according to `langid.py` and illustrates that the vast majority of Google+ content is published in English.

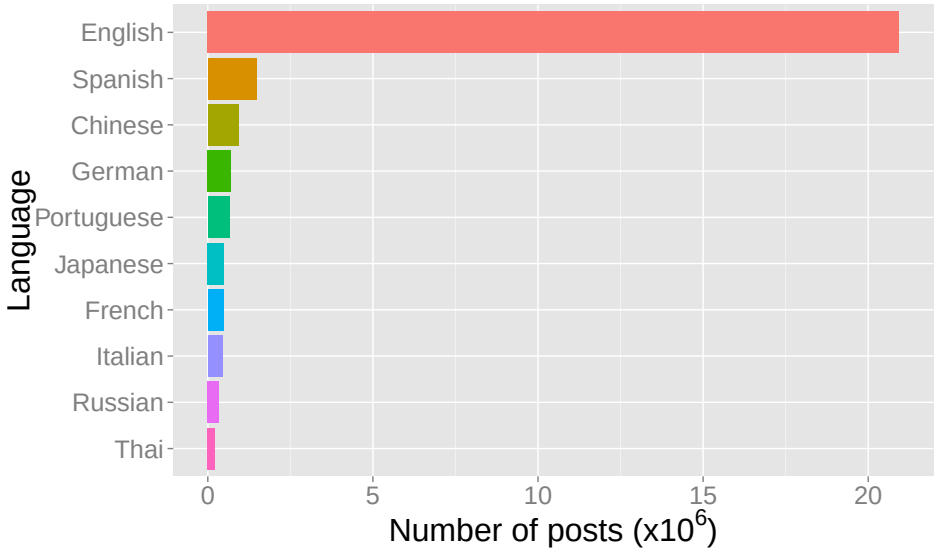

Figure 6.2. Number of posts per language. The graphic exhibits information about the ten most popular languages in our dataset.

6.1.3 Length

Figure 6.3 displays a general characterization of distinct Google+ posts written in English. The first two graphics show, respectively, cumulative distribution functions of numbers of characters and words per post. On average, posts have 111.2 characters and 25.6 words. The third graphic indicates that the majority of posts have only a few sentences: 53% of them have one sentence, while 26% have two and 10% have three sentences. This shows that, even though Google+ posts are not compulsorily limited to a small number of characters like Twitter updates and Foursquare tips, they can still be considered microtexts.

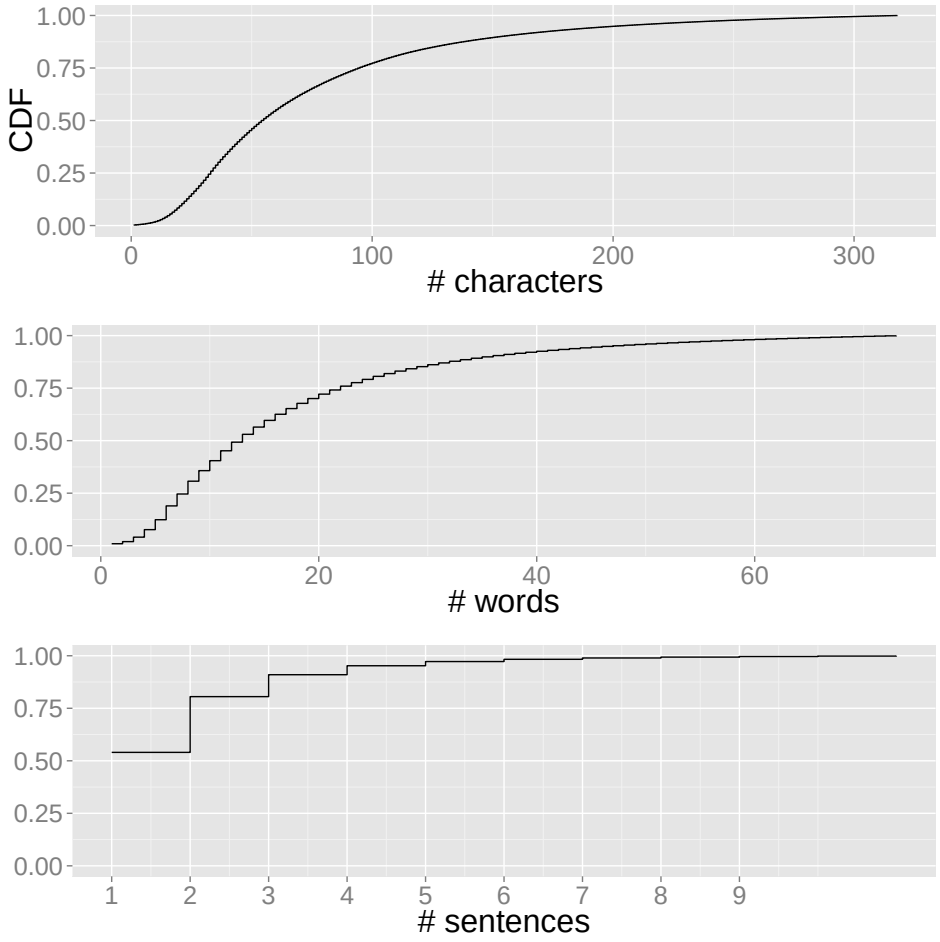

Figure 6.3. Cumulative distribution functions of numbers of characters, words and sentences per post.

6.2 Social Groups

6.2.1 Gender

Gender information is shared by 126,531,842 users (78.93% of the complete dataset) and by 770,997 users with posts collected in the ten countries studied. Considering members who set this information publicly available, 63.77% chose *male*, 34.38% chose *female* and 1.85% chose *other*. Here, we do not consider users who set their own gender as *other*.

6.2.2 Country

We inferred users' location using information available in the field *Places lived*, in which members can create a list of places where they have lived. This is an open field, meaning that users can type any text they want to. Therefore, the same place can be written in different ways (e.g. *New York*, *NYC*, *New York City*) or using distinct geographic levels (e.g. *Los Angeles*, *California*, *USA*).

To identify an user's country, we extracted the geographic coordinates of the last location cited and translated them into a valid country identifier. In this fashion, we were able to identify the country of 22,578,898 members (14.08% of the full dataset). Remaining users set this information as private or simply did not fill this field.

Here, we consider only members located in the ten countries with most posts in English: United States (US), Great Britain (GB), India (IN), Canada (CA), Australia (AU), Indonesia (ID), Germany (DE), Philippines (PH), Malaysia (MY) and France (FR).

6.2.3 Occupation

The field *Occupation* is an open field, so users can type any text they want to in order to describe their activity. As a result, we gathered a very large number of different occupations and had to summarize the information introduced by users: first, we manually aggregated the most common strings present in the dataset, since the same occupation can be written in different ways (e.g. *student*, *study*, *graduate student*, *go to school*); second, we selected the top 30 occupations; third, we used the Standard Occupational Classification (SOC) by the U.S. Bureau of Labor Statistics [70] to divide these occupations into the major groups of professional activities used here. The occupations *student* and *retired*, although not shown in the SOC, are also considered

in our analyses. Table 6.1 shows the number of posts and users per social group in our dataset.

Social group	# posts	# users	Social group	# posts	# users
Country			Occupation		
United States (US)	1,460k	494k	Student	85k	36k
Great Britain (GB)	182k	62k	Computer and math.	61k	19k
India (IN)	177k	96k	Arts and design	25k	7,9k
Canada (CA)	101k	34k	Archit. and engin.	15k	6,0k
Australia (AU)	60k	21k	Business and financ.	11k	3,9k
Indonesia (ID)	40k	24k	Media	8,3k	2,1k
Germany (DE)	35k	15k	Educ. and library	6,7k	2,2k
Philippines (PH)	32k	14k	Management	5,9k	1,9k
Malaysia (MY)	22k	10k	Sales	4,6k	1,6k
France (FR)	21k	10k	Legal	2,6k	0.8k
			Retired	2,2k	0.9k
Gender			Healthcare	1,9k	0.8k
Male	1,549k	557k	Religious	1,5k	0.4k
Female	526k	203k	Science	1,2k	0.4k
Other/NA	55k	18k	Food preparation	0.7k	0.3k
			Other/NA	1,897k	695k

Table 6.1. Number of posts and users per social group (round).

We observe that United States is the country with the highest number of posts written in English, since it is also the most popular in Google, followed by Great Britain and India. Regarding the gender, we observe a disproportional number of posts between females and males, as was expected due to the unbalanced gender distribution in Google+. The occupation distribution is also unbalanced, where the 2 most popular occupations ("Student" and "Computer and Mathematics") have more posts than all the other 13 occupations together. This is related to the phenomena of Google+ being very popular among IT students and professionals, as stated earlier in this work. The high number of unknown occupations is due to the necessity of manual inspection and classification of the professions, since the Occupation is a free-text field in Google+.

6.3 Misspellings

The occurrence of misspelled words in texts may signify unawareness of standard orthographic rules or carelessness during typing, due to negligence or lack of revision. Thus, calculating the extent to which misspellings emerge in our dataset might indicate how high literacy levels in English of the communities are or how concerned individuals are about the quality of their posts, since, for most users, it may not matter whether they make misspellings in OSN posts. In a few cases, also, misspellings may be on purpose, in order to create specific effects on readers.

By using a list of 4,238 common misspellings in English[2], that encompasses 31.3% of the whole vocabulary employed in the dataset, we investigated the occurrence of these non-standard linguistic elements in Google+ posts produced by different social groups. This list, that considers spelling differences in distinct varieties of the language, comprises misspelled items and their corresponding standard spellings, which are, therefore, the only words susceptible to misspelling in our analysis.

We calculated the fraction of misspellings per post by dividing the number of misspelled words by the number of words susceptible to misspelling. To avoid biases due to the small number of words susceptible to misspelling in some posts (e.g. if a post has only one word susceptible to misspelling, its fraction of misspellings is either 0 or 1), we did not consider posts with less than five words that appear in our list, thus evaluating 758,233 posts.

Figure 6.4 exhibits the average fractions of misspellings per post. It expectedly shows that non native English speakers, with exception of French users, are more prone to make misspellings in English written posts. We also found that, in general, women's fraction of misspellings is higher than men's: we believe that the difference between the topics of posts written by men and women – a fact that will be considered in section 6.6 – does not force women to be so demanding on the formal linguistic attributes of the content published.

Figure 6.4 also states that workers who deal more with written texts make fewer misspellings in Google+ posts: while media, legal and education professionals have the smallest fractions of misspellings, food and health professionals have the highest ones. It is worth remembering that, by the nature of these occupations, review of written material is sometimes part of the activities performed daily by media, legal and education professionals.

6.4 Readability and structural complexity

The readability of a text can be described as the ease in which readers can properly comprehend it. A series of formulas that return numerical scores estimating the level of difficulty of texts have already been proposed [28] and should not be seen as metrics of quality of documents, since *easier* or *more difficult* texts are not necessarily *worse* or *better* texts. In this study, we employ a readability index to diagnose differences in the organization of speech by distinct groups in Google+.

[2] `http://bit.ly/1ieaEOa`

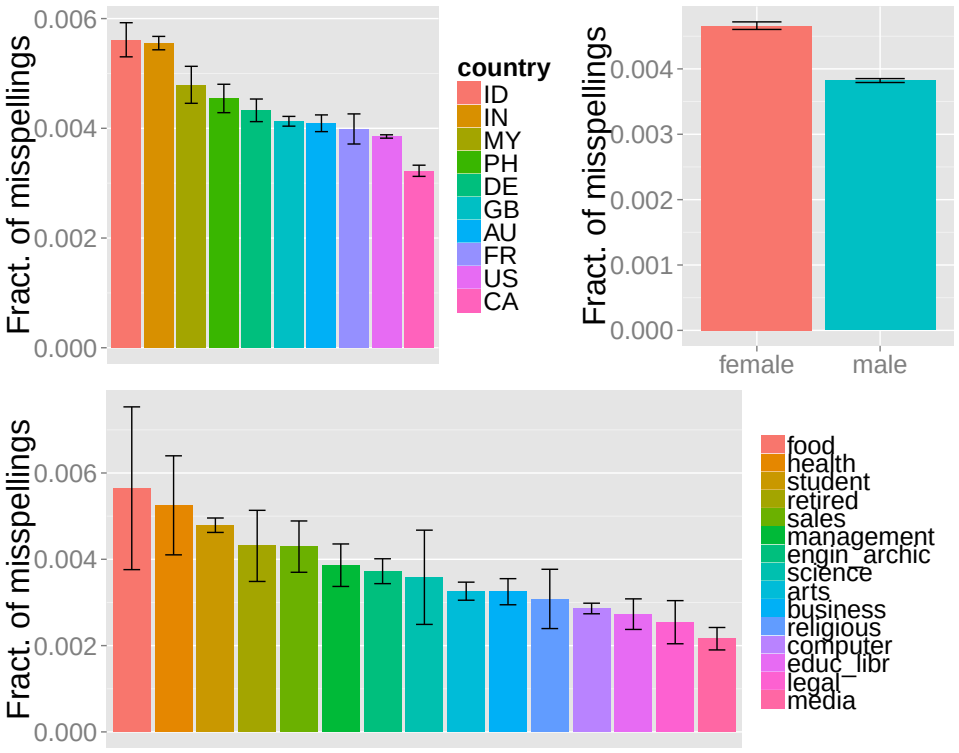

Figure 6.4. Average fractions of misspellings per post for different countries, genders and occupations ± standard errors.

We used the Unix command `style` to calculate readability values of the posts. It returns results for the Automated Readability Index (ARI), which calculates the readability of a text using the formula $ARI = 4.71 \cdot \frac{\#ofcharacters}{\#ofwords} + 0.5 \cdot \frac{\#ofwords}{\#ofsentences} - 21.43$. As one can see, the ARI relies mostly on a factor of characters per word and, on a lesser extent, on a factor of words per sentence. Thus, ARI's assumption is that the adoption of big words and the construction of large sentences are features that enhance the complexity of a text. Naturally, it considers only the structural complexity of the passages, not their conceptual complexity. For a detailed explanation of the derivation of the formula and the precise meaning of the constants, we refer readers to the original source [65].

Even though no single set of criteria comprises an universal concept of readability [4], the assumption that complex words and sentences hamper the understandability of texts is shared by the majority of readability indices: other aspects being equal, easier words and shorter sentences should result in increases of comprehension [66]. Defining which words can be considered complex, however, is not a trivial task: some indices use lists of words predetermined as difficult, while others – taking into account the *principle of quantity* [22], that correlates the quantity of information to the length

of linguistic forms – approximate the complexity of a term calculating its number of syllables. Nevertheless, the factor of syllables per word is not always easily and accurately obtained by computer programs. Since the number of characters is reasonably proportional to the number of syllables of a word in English [65], ARI's strategy of relying on a factor of characters per word is quite plausible.

Figure 6.5 depicts average values of ARI for distinct groups. Higher scores indicate higher structural complexity, as they correspond to bigger words and sentences. According to our results, texts of German, French and Indian users on Google+ are the most complex ones; on the other side, posts of Malaysians, Filipinos and Indonesians are the least complex. Interestingly, native speakers of English – from Australia, Great Britain, Canada and USA – present the central values, which seems to indicate that non native English speakers must have transferred linguistic patterns of their mother tongues to the second language [10]. This hypothesis is strengthened when we observe that users from countries with prevalence of speakers of Indo-European languages have the highest values of ARI and those from countries with prevalence of speakers of Austronesian languages have the lowest indices. We also observed that the average number of characters per word is very similar across countries, showing that, in this case, the discriminant factor of the readability index is the number of words per sentence, which may be highly influenced by the linguistic structures of mother tongues.

ARI scores for female and male users show that posts written by men are, on average, structurally more complex than those written by women. This fact is observed for most countries and professions. The examination of the structural complexity of posts of users with different occupations can be related to the previous analysis on misspellings: in the same way that workers from fields more associated with written communication and traditionally elaborated texts, like legal and media professionals, publish texts with fewer misspellings, they also produce more structurally complex posts than those from fields that do not necessarily deal with written texts, like food preparation and sales professionals. Ahead, in section 6.6, we will advocate that: (a) men and women make distinct use of this OSN, which could explain the differences in the complexity of the posts between genders; and (b) Google+ users are often talking about their own professional activities and, therefore, talking about topics that ask for either more or less elaborated linguistic constructions, according to their respective occupations.

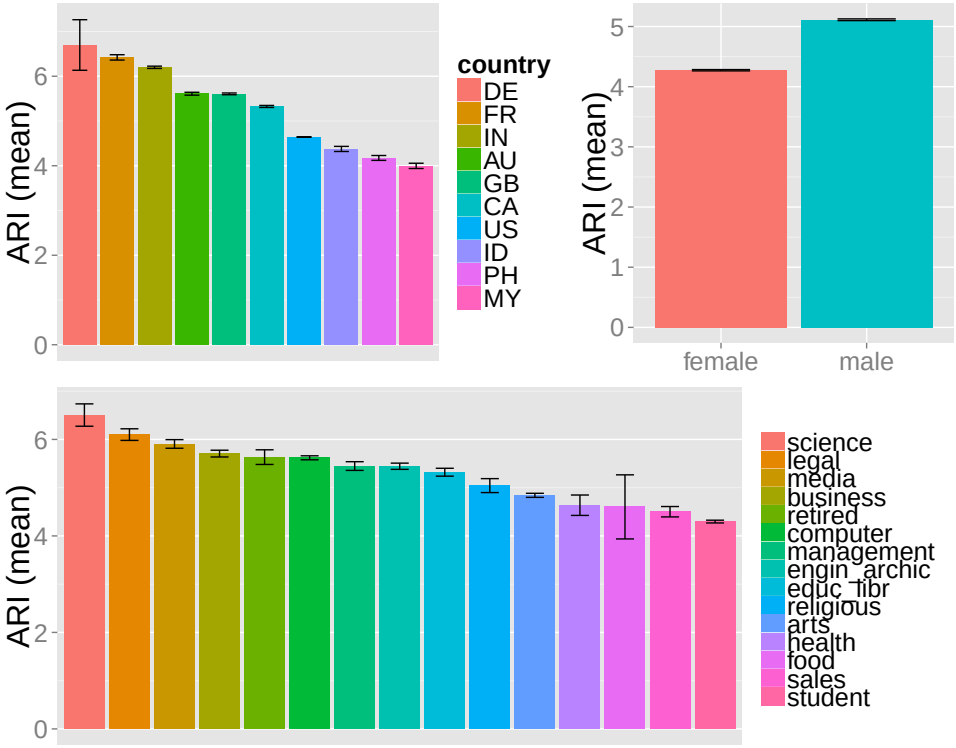

Figure 6.5. Average values of ARI for posts of users from different countries, genders and occupations ± standard errors.

6.5 Entropy

We also considered vocabulary variability across different groups, since this could add relevant insights into statistical regularities of the language employed by users. Differences of entropy values are related to the specific style of each community: lower values mean more predictable word usage, while higher ones mean more vocabulary variability. After removing stopwords and applying stemming based on Porter's algorithm [56], we calculated Shannon's entropy of the concatenation of all posts from a given group g as

$$E(g) = \sum_{\forall w_i \in g} p(w_i, g) log[p(w_i, g)]$$

where $p(w_i, g)$ is the probability of a word w_i in group g, calculated as $p(w_i, g) = freq(w_i, g) / \sum_{\forall w \in g} freq(w, g)$, in which $freq(w_i, g)$ is the frequency of word w_i in group g.

Since the number of users in each group differs and the number of unique words is directly affected by the total number of words, we applied an undersampling methodology across our three categories of social groups, randomly selecting, for each group,

the number of users of the group with the lowest number of members in each category: for countries, Malaysia (9,761 users); for genders, female (203,294 users); and for occupations, food preparation professionals (259 users). We repeated this process 25 times and calculated the mean.

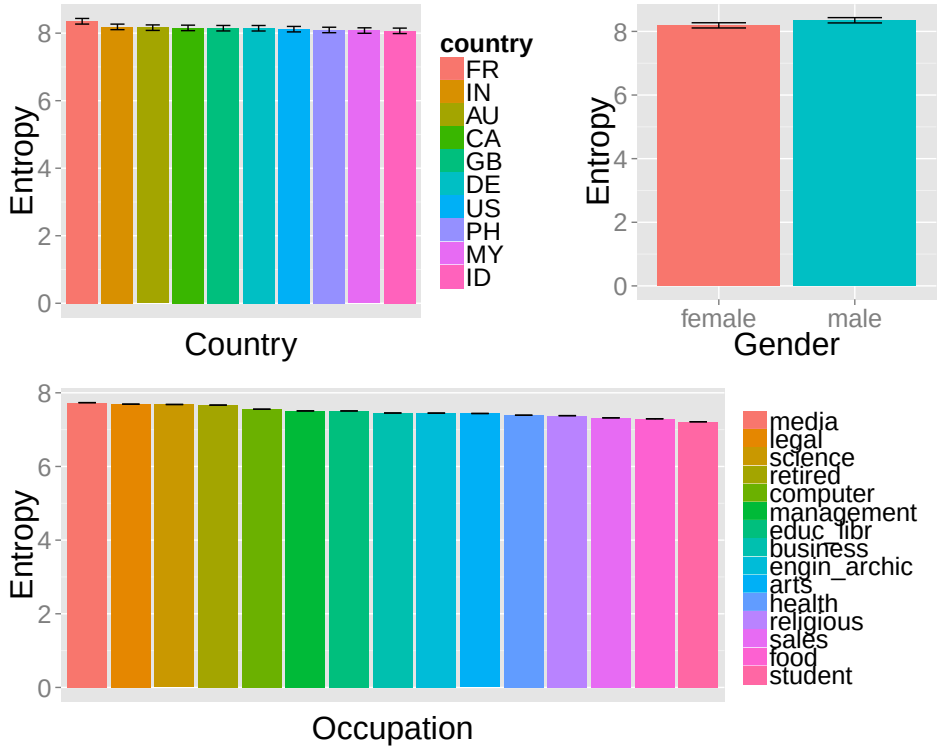

Figure 6.6. Entropy for different countries, genders and occupations ± standard errors.

In Figure 6.6, we present the results for each social group. We did not find significant differences among entropy values of different social groups, indicating that they are not discriminant on the variability of vocabulary in the context of Google+ posts. We must also consider that, since posts are small texts, they could not be relevant to this kind of analysis, since they are not long enough to allow real vocabulary diversity.

6.6 Semantic categories of words

An interesting way of investigating language differences across distinct groups is through the analysis of the vocabulary used by their members. Since vocabulary is a system of mapping the world, this kind of investigation reveals how groups perceive reality, showing what the main concerns of certain communities are.

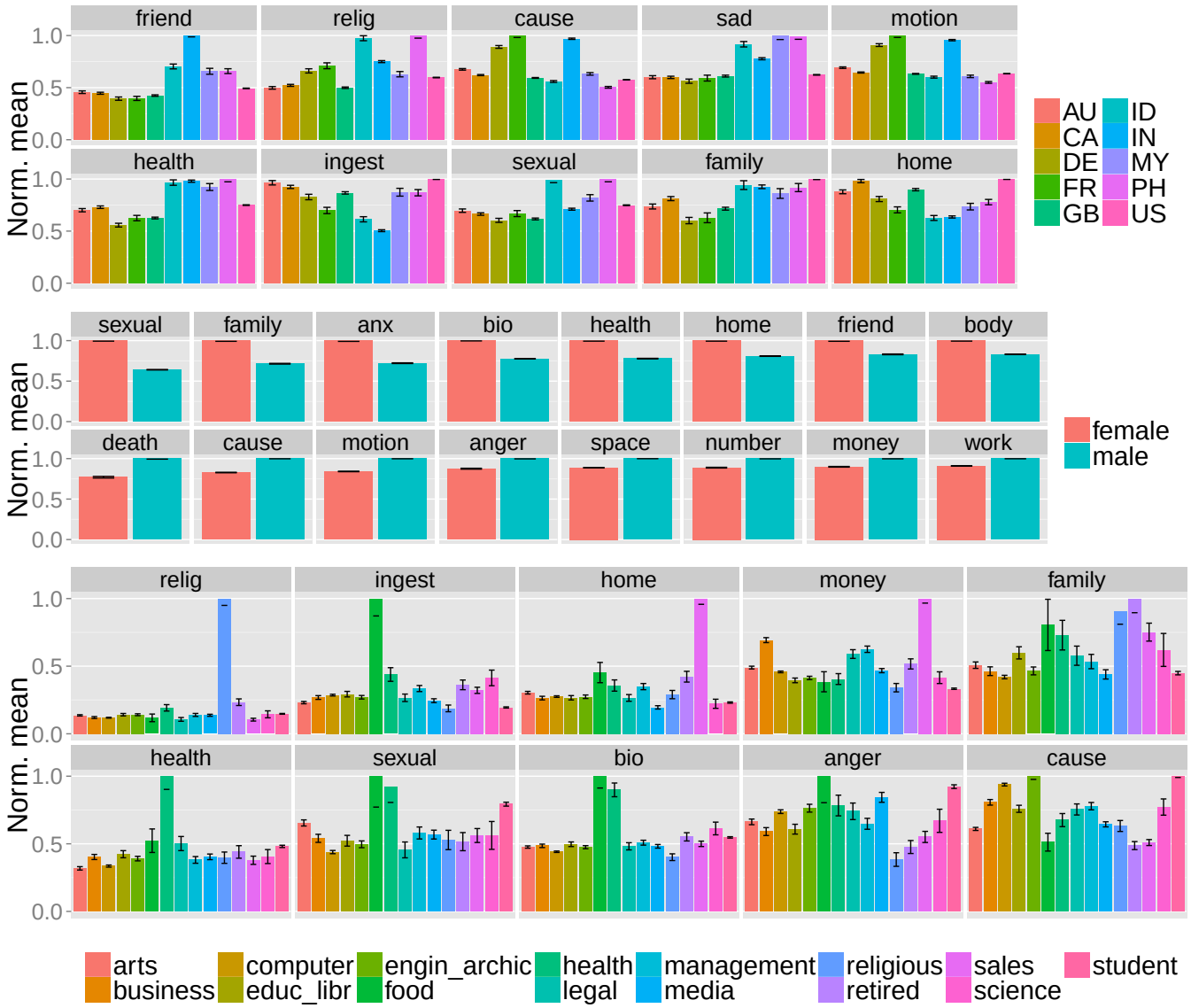

Figure 6.7. Semantic categories of words with most significant differences across distinct groups of users (countries, genders and occupations, respectively) ± standard errors.

We aim to identify if some given semantic categories of words are more common in texts produced by members of particular countries, genders and occupations. To accomplish this task, we used the Language Inquiry and Word Count (LIWC) [54], a tool that examines texts and verifies the occurrence of words previously classified as members of grammatical (e.g. pronouns, articles, prepositions etc.) or semantic (e.g. social, money, religion etc.) categories. A comprehensive list of all LIWC categories, including examples of words that are part of each category, is available at `http://www.liwc.net/descriptiontable1.php`.

We calculated LIWC scores for a given category of words as the fraction of words of this category in the total amount of categorized words of a particular post. After having calculated LIWC scores for 41 categories of semantic words, we compared them across the social groups. Figure 6.7 shows the categories of words with most significant differences across the groups considered in this study. The magnitudes of the differences among groups were obtained after the calculation of the Gini coefficient [17] for each category and those with highest coefficient are displayed (i.e. highest inequality, or dispersion). The Gini coefficient is a measure of statistical dispersion, calculated by measuring the inequality among values of a frequency distribution.

We observed that users from different countries hold distinct patterns in the usage of certain semantic categories of words in their posts. For example, Indians have the highest scores in the use of words from categories such as *friend*, *humans* and *social*, while they have low scores in categories like *negative emotions*, *anger* and *time*. Also, users from most of the Western countries considered here tend to be the main users of words related to *home*, *money* and *work* and the least users of words from the categories *health*, *affection*, *positive emotions* and *family*. These categories might be revealing the topics more covered in the posts and are a sign of cultural differences among users from different countries.

Considering gender, we found that women are more prone to use words from categories such as *family*, *home*, *friend*, *social*, *humans*, *affection* and *emotions*, while men are the main adopters of words from categories like *cause*, *motion*, *space*, *numbers*, *money* and *work*. We interpret these results suggesting that men have a tendency to use Google+ to talk about technical topics, their achievements and professional activities, while women are more likely to use this OSN to talk about their social and familial relations. These distinct approaches toward this specific online social networking service may also be the reason why men's posts are more structurally complex and more formally accurate, having fewer misspellings, as described in the sections above.

We also found a clear correlation between word usage and users' occupations. For instance, words related to religion are extremely more frequent in posts from religious professionals; the same for money vocabulary in posts from salespeople, body-related words in posts from health workers, among many others (interestingly, the category *family* is adopted mainly by retired users). This fact suggests that vocabulary employed in Google+ posts is highly related to users' working activities, indicating that this OSN may be often used for professional activities or that members' professional vocabulary is maintained even in this environment.

As far as we are concerned, these significant differences among the vocabulary of users with different occupations have been found for the first time in online social

media.

6.7 Inference of social groups

To illustrate a possible application of these results, we propose the task of inferring social characteristics of users based on linguistic analysis of their posts. This type of application is useful to assist in the development of tools aiming authorship attribution for purposes like personalization of services and identification of fake profiles.

We conducted a preliminary classification experiment using textual metrics related to the ones contemplated above. For each user, we created a vector containing 76 features: 4 size metrics (numbers of characters, words, sentences and paragraphs per post), 7 readability indices (ARI and other indices provided by the Unix command `style`), 64 LIWC categories (including categories of semantic words considered in section 6.6 and categories of grammatical words) and fraction of misspellings. By using support vector machine classifier (SVM), we tried to infer users' gender (2 classes), country (10 classes) and occupation (15 classes). We used the `scikit-learn` library [53] to conduct the SVM classification and parametrization. For the experiments, we employed a 5-fold cross-validation technique randomly selecting a fixed number of users per class: 1,000 for countries and genders; 259 – the number of members in the smallest occupation group – for occupations. The results reported in Table 6.2 are the averages of the 25 runs and their respective confidence intervals at 95%.

	Accuracy random	Accuracy SVM	F1 weighted
Gender	0.5000	0.5985±0.0093	0.5768±0.0079
Country	0.1000	0.1830±0.0032	0.1788±0.0027
Occupation	0.0666	0.1563±0.0054	0.1515±0.0044

Table 6.2. Results of the inference experiments.

Table 6.2 shows that, when using our vector of linguistic features, the SVM classifier increased in 19.7% (for genders), 83.0% (for countries) and 134.6% (for occupations) the accuracies of the inferences if compared to a random classifier. We advocate that this vector can be used in conjunction with other metrics, such as profile information and network topology, with the goal of increasing the quality of predictors of social characteristics of members in information networks. It is important to note that this is a preliminary experiment, with the aim of evaluating the potencial of our linguis-

tic features for authorship attribution, and that many improvements shall be made in future.

Table 6.3 depicts values of F1 per class, indicating that some groups – like Indians ans religious professionals – are much more easily identified by our classifier than others – like Australians and architects/engineers.

Social group	F1	Social group	F1
Country		Occupation	
India (IN)	0.2593	Religious	0.4191
Philippines (PH)	0.2365	Sales	0.2277
Indonesia (ID)	0.2030	Retired	0.1879
United States (US)	0.1910	Media	0.1761
Canada (CA)	0.1851	Business and financial	0.1465
Great Britain (GB)	0.1845	Healthcare	0.1393
France (FR)	0.1605	Legal	0.1364
Germany (DE)	0.1553	Student	0.1354
Malaysia (MY)	0.1148	Computer and mathematical	0.1227
Australia (AU)	0.0990	Arts and design	0.1177
		Education and library	0.1075
Gender		Management	0.0994
Male	0.6179	Science	0.0931
Female	0.5768	Food preparation	0.0672
		Architecture and engineering	0.0463

Table 6.3. Score of the classes of social groups.

Other studies already proposed solutions for gender classification in different online social systems. Schler et al. [63], who investigated language use in blogs, achieved up to 80.1% of accuracy in this task; Burger et al. [9], in their Twitter classifier relying only on text attributes, achieved 75.5% of accuracy; and Rao et al. [57], who also studied Twitter, achieved up to 72.33% of accuracy. Although the accuracy of our preliminary gender classifier is not high if compared to these previous ones, we believe that they and other classifiers can benefit from the use of some of the features proposed here.

Eisenstein et al. [24] addressed the issue of inferring users' geographic location from Twitter texts. Differently from us, they only considered users from different states in the United States, which makes comparison between our and their studies quite difficult. The task of predicting the professional activity of OSN users, however, seems to be an unexplored subject, since we did not find studies regarding the inference of occupations in online systems.

We advocate, then, that our vector of linguistic features can be used in conjunction with other metrics, such as profile information, network topology and other linguistic metrics, with the goal of increasing the quality of predictors of social characteristics of members in information networks.

Chapter 7

Implications

So far, we have made a series of observations about the service, network topology, and users of the Google+ social network based on large-scale data. In this chapter, we discuss the implications of these findings.

First, given that Google+ is a new social network, our first interest is to compare the topological structure of Google+ against other social networks. Compared to other social networks, Google+ cherishes openness in content sharing and is not a "walled garden" service like Facebook, where only the members can access content [48]. On the other hand, Google+ enforces a strong notion of friendship links by allowing users to manage different circles of friends. Our data analyses indicate that Google+ is in fact truly a social network, where the social links are correlated in geography reflecting offline friendship (i.e., friends are more likely to be located close), is far more reciprocal (i.e., bidirectional links), and have higher clustering coefficient (i.e., have triangle structures) compared to Twitter. The average path length is shown slightly longer than the other networks. As shown earlier, it is 6.0 in Google+ compared to 4.1–4.7 in other networks. A possible reason stem from the Google+ network is new and is still in the growing phase.

Second, observing the patterns of Google+ penetration worldwide can give insight into other new social networking service providers who would like to enter the market. While most new social network services typically starts their operation as a third party application or an adds-on service to the existing OSN services, Google+ is leading a full-fledged competition in the field. Therefore the pattern of how this new service is being adopted is important. While popular social networks like Facebook are known to have extremely high penetration rate of 50% or above [73], there is still room for a new social network service to become a hit in some countries [36]. In particular, Google+ have been successfully adopted by countries with lower GDP per capita and

this trend is important because the Internet penetration rate of these countries are growing fast–meaning the user base could potentially grow more rapidly for Google+.

Third, our findings about the privacy concern of users indicate that users exhibit different privacy notions and expectations in Google+, based on geography. Such differences could be taken into account when trying to build a recommender system or run an advertisement campaign on top of Google+, for instance, the system could feature newly emerging musicians to users in Mexico, while recommend journalists to newly joining users in Italy. Also, running a political campaign on Google+ may turn out more successful for countries with high participation of politicians in the network, like Spain. Another example would be that marketers could build appealing profiles for companies by following the right level of privacy concerns in each country.

Fourth, based on the information about the circle list and the geography of users, we have examined how the social links are distributed across different countries. The resulting map in Figure 5.7 shows an interpreting landscape of user interactions. We find very different user behaviors in this case. Certain countries like Brazil, India, and Indonesia appear far more inward looking when forming social links, than those outward looking countries like United Kingdom and Canada. This means that based on the geographical location of where a user lives, her expectation towards finding a stronger local community in the network is different. We believe this kind of social network analysis allows us to study the collective and deviant behavior of particular demographics, which are increasingly considered important and useful both in research and practice. This analysis also has impact on epidemics, since the global spread of a system depends on its structure.

Chapter 8

Conclusion

In this work we study characteristics of the Google+ social graph. We present a comprehensive description of the platform, highlighting the main differences from other popular social network models. Our study is based on a large amount of data gathered encompassing 160 million user profiles and their connections to other users. With this dataset we analyze unique features of the Google+ demographics, especially on the gender, occupation, relationship status, and geo-location of users. We construct a graph representing the social relations of Google+ and analyze its structural properties, such as reciprocity, clustering coefficient, node degree distribution and connected components. The Google+ social graph has a giant connected component that included 61% of the crawled users, which means that information can flow freely among all such users.

We also compute the average physical distance between two connected users. Exploiting the geo-location of users, we could see how aggressively Google+ has been adopted in different countries. We investigate relationships between economic indexes of countries and the adoption rate of Google+. We find that Google+ is popular in countries with relatively low Internet penetration rate. By examining the top users based on the circle link information, five out of the top 30 users turned out to be in the information technology industry, a trend that is rather uncommon in other online social networks, where popular figures are media outlets, celebrities, and public figures. By looking into users who share their contact information publicly, we observe that a large fraction of the users are male and single.

We evaluate linguistic elements among members of particular social groups. These analyses not only describe the posts, but especially identify how distinct groups differ when posting content on the Web. Contributions of our study go beyond the mere characterization of posts – which per se is an important supplement to the literature on

language use in social media –, since implications on authorship attribution may follow. For this reason, we implemented a preliminary classifier to infer social characteristics of Google+ users, which may be an useful tool to improve the task of automatically detecting fake profiles through the analysis of their linguistic behaviors and to improve language modeling focused on personalization of services.

There are several interesting directions for future research. First, we are interested in measuring the speed at which a new social network service grows and whether we can predict the phase transitions in the growth sparks (e.g., tipping point when a network suddenly shows a rapid growth or the point where the growth stabilizes and turns into a dormant phase). By collecting multiple snapshots of the Google+ topology, we hope to gain insight in the dynamic changes in the internal structure of the social network over various adoption phases. Second, having seen the key differences of Google+ from other online social networks, we would like to understand how different privacy settings and openness impact the types of conversations and the patterns of content sharing in Google+.

Regarding the posts, future work should include the analysis of other relevant linguistic and social factors, such as the topic of posts and the educational level of users. Also, it would be interesting to compare the outcomes reported here for Google+ with other popular OSNs, such as Facebook and Twitter. Another related issue to be analyzed in future studies is the question of how these different social groups express their feelings on the Web and which linguistic elements are used to indicate tones of happiness, angriness, hope and hatred, among others: are these elements also distinctive across different social groups in the context of online social networking services?

Bibliography

[1] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *Proc. of ACM International World Wide Web Conference*, pages 835--844.

[2] Allen, P. (2011). Google+ growth accelerating. passes 62 million users. adding 625,000 new users per day. prediction: 400 million users by end of 2012. `http://bit.ly/gp_nusers`.

[3] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2011). Four degrees of separation. *CoRR*, abs/1111.4570.

[4] Bailin, A. and Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: a critique. *Language and Communication*, 21:285--301.

[5] Bell, C. M., McCarthy, P. M., and McNamara, D. S. (2012). Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles. In McCarthy, P. M. and Boonthum-Denecke, C., editors, *Applied Natural Language Processing: Identification, Investigation, and Resolution*. Information Science Reference, Hershey, PA.

[6] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proc. of ACM SIGCOMM Internet Measurement Conference*, pages 49--62.

[7] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2012). Characterizing user navigation and interactions in online social networks. *Inf. Sci.*, 195:1--24. ISSN 0020-0255.

[8] Brin, S. (2012). Web freedom faces greatest threat ever, warns google's sergey brin. `http://www.guardian.co.uk/technology/2012/apr/15/web-freedom-threat-google-brin`.

[9] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301--1309. Association for Computational Linguistics.

[10] Cadierno, T. and Ruiz, L. (2006). Motion events in Spanish L2 acquisition. *Annual Review of Cognitive Linguistics*, 4:183--216.

[11] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. of AAAI International Conference on Weblogs and Social Media*.

[12] Cha, M., Mislove, A., and Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proc. of ACM International World Wide Web Conference*, pages 721--730.

[13] Comission, E. (2011). Attitudes on data protection and electronic identity in the european union. `http://bit.ly/wkZhGR`.

[14] comScore (2011a). Google+ Reaches 20 Million Visitors in 21 Days. `http://tinyurl.com/3ox3lp9`.

[15] comScore (2011b). It's a Social World: Top 10 Need-to-Knows About Social Networking and Where It's Headed. `http://tinyurl.com/cfhqxke`.

[16] Cormen, T. H., Stein, C., Rivest, R. L., and Leiserson, C. E. (2001). *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition. ISBN 0070131511.

[17] Cowell, F. (2011). *Measuring Inequality*. Oxford University Press, USA, 3rd edition.

[18] Crystal, D. (2004). *The Language Revolution*. Polity Press, Cambridge, UK.

[19] Crystal, D. (2005). The scope of Internet Linguistics. *American Association for the Advancement of Science*.

[20] Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014a). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLOSONE*, 9(1):e87041.

[21] Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014b). How you post is who you are: Characterizing google+ status updates across social groups. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14. ACM.

[22] Dirven, R. and Verspoor, M. (2004). *Cognitive Exploration of Language and Linguistics*. John Benjamins Publishing, Philadelphia, PA.

[23] Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

[24] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277--1287. Association for Computational Linguistics.

[25] engadget (2012). Google+ has 250 million users, more mobile than desktop. `http://tinyurl.com/d3ausse`.

[26] Fogel, J. and Nehmad, E. (2009). Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153--160.

[27] Freshtrax (2010). Top 10 most popular websites in china. `http://bit.ly/xgRIAq`.

[28] Fry, E. (2006). Readability. In *Reading Hall of Fame Book*.

[29] Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2009). A walk in facebook: Uniform sampling of users in online social networks. *CoRR*, abs/0906.0060.

[30] Gonzales, R., Cuevas, R., Motamedi, R., Rejaie, R., and Cuevas, A. (2013). Google+ or Google-? Dissecting the evolution of the new OSN in its first year. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW 2013)*.

[31] Google (2012). Orkut demographic data. `http://www.orkut.com/MembersAll`.

[32] Gross, R. and Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proc. of ACM CCS Workshop on Privacy in the Electronic Society*, pages 71--80.

[33] Gundotra, V. (2011a). Google+: 92, 93, 94, 95, 96, 97, 98, 99... 100. `http://bit.ly/gp_noinvite`.

[34] Gundotra, V. (2011b). Introducing the google+ project: Real-life sharing, rethought for the web. `http://bit.ly/gp_released`.

[35] Hardy, Q. (2011). Google+ gains traction, researcher says. `http://nyti.ms/wrEsOx`.

[36] Internet World Stats (2014). Facebook growth and penetration in the world. `http://www.internetworldstats.com/facebook.htm`.

[37] Jain, P., Jain, P., and Kumaraguru, P. (2013). Call me maybe: Understanding nature and risks of sharing mobile numbers on online social networks. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 101--106, New York, NY, USA. ACM.

[38] Jones, N. L. (2011). Talking the talk: the confusing, conflicting and contradictory communicative role of workplace jargon in modern organizations. Master's thesis, University of Rhode Island.

[39] Kairam, S., Brzozowski, M. J., Huffaker, D., and Chi, E. H. (2012). Talking in circles: selective sharing in Google+. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'12)*.

[40] Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 611--617.

[41] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proc. of ACM International World Wide Web Conference*, pages 591--600.

[42] Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Blackwell, Malden, MA.

[43] Lakoff, R. (1975). *Language and Woman's Place*. Harper and Row, New York, NY.

[44] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proc. of ACM International World Wide Web Conference*, pages 915--924.

[45] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 177--187.

[46] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623--11628.

[47] Lui, M. and Baldwin, T. (2012). langid.py: an off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 25--30.

[48] MacManus, R. (2010). Tim berners-lee calls facebook a walled garden - is that fair? `http://www.readwriteweb.com/archives/tim_berners-lee_says_facebook_is_a_walled_garden.php`.

[49] Magno, G., Comarela, G., Saez-Trumper, D., Cha, M., and Almeida, V. (2012). New kid on the block: Exploring the Google+ social graph. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, IMC '12, pages 159--170. ACM.

[50] Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60--67.

[51] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In *Proc. of ACM SIGCOMM Internet Measurement Conference*.

[52] Ottoni, R., ao Paulo Pesce, J., Las Casas, D., Franciscani Jr., G., Meira Jr., W., Kumaraguru, P., and Almeida, V. (2013). Ladies first: Analyzing gender roles and behaviors in Pinterest. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*.

[53] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825--2830.

[54] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). *The Development and Psychometric Properties of LIWC2007*. The University of Texas at Austin and The University of Auckland, New Zealand.

[55] Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do All Birds Tweet the Same? Characterizing Twitter Around the World. In *Proc. of ACM Conference on Information and Knowledge Management*, Glasgow,UK.

[56] Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, 14:130--137.

[57] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 37--44. ACM.

[58] Ribeiro, B. F. and Towsley, D. F. (2010). Estimating and sampling graphs with multidimensional random walks. *CoRR*, abs/1002.1751.

[59] Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K. P., and Almeida, V. (2011). On Word-of-Mouth Web Based Discovery of the Web. In *Proc. of ACM SIGCOMM Internet Measurement Conference*.

[60] Romaine, S. (1994). *Language in Society: An Introduction to Sociolinguistics*. Oxford University Press.

[61] Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. (2010). Distance matters: geo-social metrics for online social networks. In *Proc. of ACM SIGCOMM Workshop on Social Networks*, Berkeley, CA, USA. USENIX Association.

[62] Schiöberg, D., Schneider, F., Schiöberg, H., Schmid, S., Uhlig, S., and Feldmann, A. (2012). Tracing the birth of an OSN: social graph and profile analysis in Google+. In *Proceedings of the 4th International Conference on Web Science (WebSci'12)*.

[63] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199--205.

[64] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOSONE*, 8(9):e73791.

[65] Smith, E. A. and Senter, R. J. (1967). *Automated Readability Index*. University of Cincinnati, Ohio.

[66] Swanson, C. and Fox, H. (1953). Validity of readability formulas. *The Journal of Applied Psychology*, 37(2).

[67] Taraszow, T., Aristodemou, E., Shitta, G., Laouris, Y., and Arsoy, A. (2010). Disclosure of personal and contact information by young people in social networking sites: An analysis using Facebook profiles as an example. *International Journal of Media and Cultural Politics*, pages 81--101.

[68] Trudgill, P. (1983). *Sociolinguistics: an introduction to language and society.* Penguin, London, UK.

[69] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *CoRR*, abs/1111.4503.

[70] U.S. Bureau of Labor Statistics (2010). Standard occupational classification and coding structure. `http://1.usa.gov/14INxmQ`.

[71] Wang, Y. C., Burke, M., and Kraut, R. (2013). Gender, topic, and audience response: An analysis of user-generated content on Facebook. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'13).*

[72] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440--442.

[73] ZDNet (2012). Top 10 countries in Facebook adoption. `http://zd.net/xmA4rt`.