User Behavior in Microblogs with a Cultural Emphasis

Ruth Olimpia García-Gavilanes

PhD Thesis UPF / 2015

Supervisor

Prof. Dr. Ricardo Baeza-Yates,
Department of Information and Communication Technologies

To my sisters and beloved parents. To science, enlightening us.

Acknowledgements

This dissertation is the result of many attempts and failures, many falls and rebirths before ultimately finding the light at the end of the tunnel. I would like to express my deepest gratitude to all those that contributed to this dissertation, both, emotionally and intellectually. First of all, I thank my advisor, Ricardo Baeza-Yates, for his amazing patience and his willingness to provide valuable feedback and for believing in me even when I didn't. If the dissertation is today readable and has a coherent structure, it is thanks to him. I am also grateful to Xavier Amatriain who introduced me to research for the first time. Without his support and guidance during my Master thesis I would have never thought about doing a Ph.D. Coming from a small country in South America where research is just starting to gain value now, little I knew about this parallel world of research. Thanks again Ricardo and Xavier for making a big difference in my life and helping me to take that further step. I am also very thankful to Fundación Carolina for the scholarship that funded my Master.

My first year was full of discoveries starting by working with Barbara Poblete and Marcelo Mendoza thanks to the support of Alejandro Jaimes who brought me to Yahoo Labs for all my Ph.D. Thanks to Barbara I found out that amazing initiatives, such as the Grace Hopper Conference, actually exist. I also want to thank Ingmar Weber for the support at my initial stage not only on research but also by motivating me to run and dance. He gave me hope that one can work hard and have an active healthy life when you "organize your time". Thank you Carlos Castillo (Chato) for leaving your code behind in the lab, I used them many times. Sandro, thanks for understanding my absence.

My second year started with an internship in Qatar, I want to thank Sihem Amer-Yahia and Ihab Ilyas for giving me the opportunity to learn from them. Not only they gave me intellectual advice but also they helped me to overcome emotional struggles. Sihem thanks for being such a great role model.

The most difficult time of my Ph.D also came in my second year. I worked on an idea of my own and I saw the dark for the first time. Without the help of Luca Aiello and Neil O'Hare, I would have never moved on with this idea, I learned an important lesson with you guys: to not give up. After many rejections, we finally got a paper award.

My last year started with the enlightenment of Daniele. Super Daniele Quercia, thanks so much for your guidance, ethical example and strength. You helped me to understand about Cultural Models and how I could apply them to data. We came up with a paper that gave me an honorable mention and the process was enjoyable. I finally learned to enjoy research with you. Thanks Yelena Mejova for supporting me in my crazy ideas that followed and for holding me when I almost gave up. Without you, I may have never submitted that paper. Thank you Yelena for the very early runs as well.

My fourth year was finally more relaxing, I moved on to pursue my own ideas, some of them failed, some of them succeeded. Thanks Andreas Kaltenbrunner, Pablo Aragón and David Laniado for the feedback, and for being an awesome team. Olivier thank you for obliging me to stop working and come back home earlier to enjoy a good cava at nights and for helping me with my coding. Thanks for introducing me to Kaggle and for helping me improve the presentation of my CV.

Thanks to all scientists, engineers, post-docs and interns that passed by Yahoo Labs Barcelona. I will never forget that I found a family friendly place in which ideas and emotions are exchanged and great friendships are forged. Thanks Ilaria, Paloma and Sara. You are the source of so many happy moments. Thanks my fellow Ph.D colleagues. Michele T. and Luca C., you taught me the value of team work. You guys are amazing! Luca you taught me that discipline, organization and hard work leads you to enjoy free times. I still miss feeling the vibrations of your mind concentration at my side. Eduardo, thank you for your friendship and for introducing me to Chilean culture. Janette, thanks for lending me your bike many times. My dear Carmen, thank you so much for your support, I have been blessed with your friendship. There are few people as compassionate, generous and giving as you Carmen. Amin and Ioannis, thank you for taking time to know me and listening to me in my best and worst moments.

I am also grateful to Diego Sáez for his support in this last stage and Lydia García for her help in all the paper work required for my legal status in Spain.

My life in the lab, the people that I met, the battles that I fought and the happiness that I lived between these walls changed my life forever. I know I am a better and stronger person thanks to this experience.

Finally, I am very thankful for the unconditional support of my family. Mom, I will always be thankful for holding me in one of the most difficult moments of my life in the last year of my undergraduate studies. Without your help nothing of this could have been possible. Dad, thanks for your moral and financial support. Only after living alone, paying my own bills and obtaining things on my own did I realize how hard life has been for you. Thanks for going to Japan to meet me, I will never forget that. Sofi, beautiful sister, thanks for taking the time to review my thesis. Thanks for the best advices during my heart broken days. Tefy, thanks for calming me down with your presence.

Last but not least, thanks to my demons and those who made this process harder. If I had not met you, I would have never dared to cross my limits and get to know what I am capable of. I know many challenges and hard moments are still to come but I'm already so far from where I used to be and I'm proud of that.

Abstract

The main objective of this thesis is to carry out a multidisciplinary study of the behavior of microblog users. To that end we first explore several user behavior patterns employing data mining techniques. Then we use social science theories of culture and socio-economic indicators to better understand differences and similarities of user behavior across countries.

We found several insights on user behavior such as *(i)* social link recommendations made by current friends have a large effect on link formation and the accepted recommendations have more longevity than other links; *(ii)* as users mature, they evolve to adopt microblogs as a news media rather than a social network; *(iii)* the collective behavior of users from some countries standout, based on certain special characteristics such as conversations, reciprocity, etc.; *(iv)* national culture determines the temporal patterns with which users post, or the extent to which they mention, follow, recommend and befriend others; and *(v)* socio-economic and cultural features improve the prediction of communication strength among users from different countries.

Resum

L'objectiu principal d'aquesta tesi és realitzar un estudi multidisciplinar de la conducta dels usuaris de microblogs . És per això que, primer explorem diversos patrons de comportament d'usuari usant tècniques de mineria de dades. Després, fem servir algunes teories de les ciències socials en cultura i indicadors socioeconòmics per tal de comprendre millor les diferències i similituds del comportament dels usuaris a diferents països.

Trobem diversos resultats interessants sobre el comportament de l' usuari, tals com *(i)* que les recomenacions d'enllaç socials fetes per amics tenen un gran efecte sobre la formació d'enllaços socials i les recomenacions acceptades tenen més longevitat que altres enllaços ; *(ii)* A mesura que els usuaris maduren i evolucionen, utilizen els microblocs com un mitjà de comunicació enlloc de com una xarxa social; *(iii)* el comportament col·lectiu dels usuaris d'alguns països es destaca en base a certes característiques especials, com per exemple converses , reciprocitat, etc.; *(iv)* la cultura nacional determina patrons temporals amb la qual els usuaris publiquen missatges, o el grau en que s'esmenten, es recomanen i es segueixen els uns als altres; i *(v)* les característiques socioeconòmiques i culturals ajuden a millorar la predicció de la intensitat de comunicació entre els usuaris de diferents països.

Resumen

El objetivo principal de esta tesis es realizar un estudio multidisciplinario sobre la conducta de los usuarios en microblogs. Para ello primero exploramos varios patrones de comportamiento de usuario usando técnicas de minería de datos. Luego usamos algunas teorías de las ciencias sociales en cultura e indicadores socioeconómicos para comprender mejor las diferencias y similitudes del comportamiento de los usuarios en diferentes países.

Encontramos varios resultados interesantes sobre el comportamiento del usuario, tales como, *(i)* las recomendaciones de enlaces sociales hechas por amigos tienen un gran efecto sobre la formación de enlaces sociales y las recomendaciones aceptadas tienen más longevidad que otros enlaces; *(ii)* a medida que los usuarios maduran, estos evolucionan a usar los microblogs como un medio de comunicación en lugar de una red social; *(iii)* el comportamiento colectivo de los usuarios de algunos países se destaca en base a ciertas características peculiares, tales como conversaciones, reciprocidad, etc.; *(iv)* la cultura nacional determina los patrones temporales con los que los usuarios publican mensajes, o el grado en que se mencionan, recomiendan y siguen los unos a los otros; y *(v)* las características socioeconómicas y culturales ayudan a mejorar la predicción de la intensidad de la comunicación entre los usuarios de diferentes países.

Resumo

O objetivo principal deste trabalho é realizar um estudo multidisciplinar do comportamento dos usuários de microblogs. Para esse fim, primeiramente, exploramos vários padrões de comportamento do usuário empregando técnicas de mineração de dados. Nós também usamos teorias das ciências sociais em cultura e indicadores sócio-econômicos para entender melhor as diferenças e semelhanças de comportamento do usuário em diferentes países.

Encontramos vários resultados interessantes sobre o comportamento do usuário , como *(i)* recomendações sociais feitas por amigos têm um grande efeito sobre a formação de novos vínculos e as recomendações aceitas têm mais longevidade do que outros vínculos; *(ii)* assim que os usuários amadurecem, eles evoluem para usar microblogs como uma mídia de notícias, em vez de uma rede social; *(iii)* o comportamento coletivo dos usuários de alguns países se destaca em base a certas características especiais, tais como conversas, reciprocidade, etc.; *(iv)* a cultura nacional determina os padrões temporais com que os usuários publicam mensagens, ou o grau em que eles mencionam, seguem e recomendam uns aos outros; e *(v)* características sócioeconômicas e culturais para melhorar a previsão da força de comunicação entre os usuários de diferentes países.

Résumé

L'objectif principal de cette thèse est de mener une étude multidisciplinaire sur le comportement des utilisateurs sur les plateformes microblogs. Pour ce faire, nous explorons d'abord différents modèles de comportements utilisateur à l'aide de techniques d'exploration de données. Ensuite, nous utilisons les modèles de culture développés dans les théories des sciences sociales, ainsi que les indicateurs socio-économiques afin de mieux comprendre les différences et les similitudes de comportement de l'utilisateur dans les différents pays.

Nous avons trouvé plusieurs résultats intéressants sur le comportement de l'utilisateur, tels que *(i)* les recommandations de liens sociaux faites par ses amis ont une grande influence sur la formation de liens sociaux, et les recommandations acceptées ont une plus grande longévité que d'autres liens; *(ii)* les utilisateurs à mesure qu'ils se développent, évoluent vers l'adoption de microblogs comme média plutôt que vers les réseaux sociaux; *(iii)* le comportement collectif des utilisateurs dans certains pays se distingue nettement des autres, et ce au travers des conversations, de la réciprocité, et des autres caractéristiques; *(iv)* la culture nationale détermine les tendances temporelles avec lesquelles les utilisateurs postent des messages, la façon dont ils recommandent du contenu et, comment ils se suivent les uns les autres; et *(v)* les caractéristiques socio-économiques et cultureles contribuent à améliorer la prédiction de l'intensité de la communication entre les utilisateurs de différents pays.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

For many years the Web was mostly a place where users solely read information. Nowadays, with the advent of online social networks and social media in general, users can also generate content and interact with other users. In fact, social media sites have become the dominant method of using the Internet, and it has changed the way people search for information, communicate and interact with others [50].

For this reason, Social Media has become not only a great source to study the user behavior online but also it is a repository with great cultural value thanks to the content generated by users themselves. We find it contains information in different languages regarding habits, behavioral patterns, sociocultural norms, preferences,values, etc. Likewise, these social media sites influence the way people formulate content as well as the way they request, acquire, interpret and access information. Consequently, this has increased the interest of several sectors such as the advertising industry, business applications as well as search engines who are attempting to provide more personalized search results [16; 135; 119]. Nevertheless, interpreting the large scale data generated from the social Web and understanding its benefits is a recent challenge to science that requires the use of new skills and the contribution from a wide variety of disciplines.

Data mining techniques have been used to discover data patterns in an automatic or semiautomatic way [141], but researchers have started to use social sciences theories to interpret and understand users (people). In fact this interpretation has implied a significant shift in the research done with

Online Social Networks, leading to the emergence of a "new" computational social science research area at the intersection of computer science, statistics, and sociology in which quantitative methods and computational tools are used to identify and answer social science questions [139; 16].

Similarly, social sciences also benefit from big data and automatic data processing since the increasing storage of users' footprints (millions of new users getting connected every year) can have the key in understanding the social changes and even in predicting where society and individuals are headed. So far, social sciences have mostly rely on obtrusive experiments or surveys, considering a limited number of users, making it hard to extrapolate the results to a larger scale.

This Ph.D thesis focuses on studying the behavior of microblog users considering the content they generate and the way they interact with others. In particular, we give a cultural and socio-economic emphasis in the interpretation of results. We present next the key research questions that we want to answer in this thesis.

1.2 Research Questions

Online social networks support users in a wide range of activities, such as sharing information, interacting with others and even making recommendations. The possibilities tend to increase as users become familiar with social media platforms and also due to endogenous and exogenous reasons. For example, the increase of new users from different countries leads decision makers to develop new functionalities that can better target new needs, more languages, etc. Many challenges and questions arise regarding user behavior and their needs. *What is the effect on users from human generated recommendations? How do user behavior evolve over time? Do patterns of behavior remain the same in different countries?.*

To understand the differences and similarities among users from different countries, we are also interested to explore anthropological studies of culture and socio-economic indicators that shed light on user behavior across the world. In this thesis we are particularly interested on understanding the impact of culture and socio-economic factors in the way users behave and communicate with each other. *Does culture influence the way we use social media and the frequency we communicate with others?*

Hence, the main goal of this thesis is to analyze all these aspects in order to gain insights about user behavior across different cultures.

1.3 Contributions

In order to answer the research questions proposed in Section 1.2, several experiments are done. The results contribute to the state of the art by :

- Proposing how to combine anthropological studies of culture with large scale data [Chapter 3].
- Providing a study of human generated recommendation on Twitter. This is done by building a ground-truth of acceptances and rejections based on real human generated recommendations of who to follow in microblogs [Chapter 4].
- Describing the evolution of user behavior over time regarding the content they generate. This is done by characterizing messages by language independent features [Chapter 5].
- Describing differences and similarities of users across countries regarding the way people tweet, the predominant sentiment of the words used as well as how their network of friends is structured [Chapter 6].
- Correlating how and when people tweet with dimensions of national culture and pace of life (taken from anthropological studies) [Chapter 7].
- Improving the prediction of the communication strength between users from different countries. We take into account several cultural and socio-economic indicators taken from diverse sources. Furthermore we discuss the most discriminative features in the prediction [Chapter 8].

1.4 Outline

Figure 1.1 illustrates the conceptual flow of the thesis. Each part of this flow is described in the next paragraphs.

In Part I, we give a look to the state of the art. We start in Chapter 2 by giving an overview of what has been done previously on user behavior of microblogging platforms. We emphasize the importance of Twitter and describe our datasets. In Chapter 3 we explain how anthropological studies and socio-economic indicators are used to understand the collective online behavior of users from different countries and present related work.

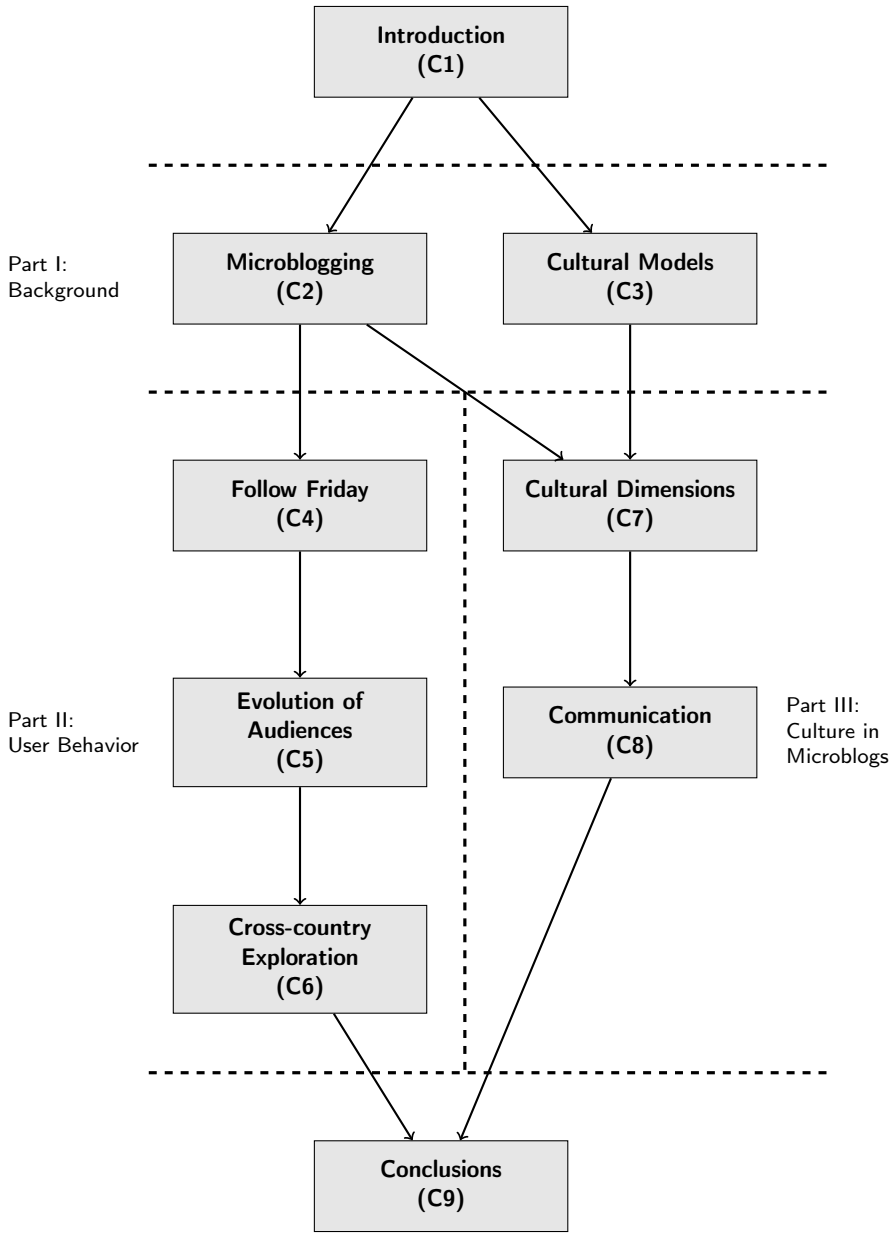


Figure 1.1: The conceptual flow of the thesis.

In Part II we study multiple aspects of user behavior. In Chapter 4, we analyze a user-generated trend in Twitter called *Follow Friday* which is used to recommend users to follow. We show that these explicit recommendations have a measurable effect on the process of link creation, increasing the chance of link creation between two and three times on average, compared with a recommendation-free scenario. Also, ties created after such recommendations have up to 6% more longevity than other Twitter ties. Finally, we build a supervised system to rank user-generated recommendations, surfacing the most valuable ones with high precision (0.52 of Mean Average Precision), and we find discriminant features to describe users and the relationships between them.

Next, in Chapter 5, we study longitudinal behavior changes in the way people tweet. User behavior in online social media is not static, it changes through the years. In this chapter, we use a taxonomy of the types of messages posted by around 4M users during 10 weeks in 2011 and 2013. We classify users according to their online posting behavior, and find 5 clusters for which we can associate a different dominant online posting type. Furthermore, we observe the evolution of users across groups between 2011 and 2013 and find interesting insights such as the decrease in conversations and increase in sharing URLs. Our findings suggest that mature users evolve to adopt Twitter as a *news media* rather than as social network.

At the end of Part II, in Chapter 6, we focus on comparing users from different countries on how and what they tweet and their network structure. To do so, we conducted a large scale analysis of the behavior of millions of users in Twitter to observe significant differences among countries and the way they use social media during one whole year. We covered the following categories: a) level of activity and languages used in the 10 countries that tweet the most, b) temporal happiness levels of tweets in two different languages for all active countries, c) the content of tweets in terms of mentions, hashtags, URLs and re-tweets, and d) the network structure: reciprocity, ties and social network. The results of this part are published in:

- Ruth García-Gavilanes, Barbara Poblete, Marcelo Mendoza, Alejandro Jaimes. Microblogging without Borders: Differences and Similarities. In *The 3rd International Conference on Information and Knowledge Management (Websci), ACM, 2011.*
- Barbara Poblete, Ruth García-Gavilanes, Marcelo Mendoza, Alejandro Jaimes. Do All Birds Tweet the Same? Characterizing Twitter

Around the World. In *The 20th International Conference on Information and Knowledge Management (CIKM), ACM, 2011.*

- Ruth García-Gavilanes, Neil O'Hare, Luca Maria Aiello, Alejandro Jaimes. Follow My Friends This Friday! An Analysis of Human-generated Friendship Recommendations. In *The 5th International Conference on Social Informatics (SocInfo), Springer 2013.* [Best paper award]

- Ruth García-Gavilanes, Andreas Kaltenbrunner, Diego Sáez-Trumper, Ricardo Baeza-Yates, Pablo Aragòn and David Laniado. Who are my Audiences? A Study of the Evolution of Target Audiences in Microblogs. In *The 6th International Conference on Social Informatics (SocInfo), Springer 2014.*

In Part III, we apply anthropological models to social media data. We carry out several experiments on the impact of culture and socio-economic factors in the way users behave and communicate with each other in microblogs.

In Chapter 7, we test three main hypotheses associated with three cultural aspects and, in doing so, we find that activity predictability in Twitter negatively correlates with Pace of Life ($r = -0.62$), tweets with mentions negatively correlates with Individualism ($r = -0.55$), and power imbalance(*e.g*, Twitter popularity) in relationships (between, for example, two users mentioning each other) is correlated with Power Distance ($r = 0.62$). We show that these three cultural dimensions matter because they are associated with a country's socio-economic aspects - with GDP per capita, income inequality, and education expenditure.

In Chapter 8, we also use cultural models and socio-economic features to predict international communication strength. We show that the *Gravity Model*, which hypothesizes that the flow between two areas is proportional to their masses and inversely proportional to the distance between them, along with other social, economic, and cultural variables, predict the communication volume at *Adjusted* R^2 of 0.80, with trade, language and racial intolerance especially impacting communication. The results of this part are published in :

- Ruth García-Gavilanes, Daniele Quercia, Alejandro Jaimes. Cultural Dimensions in Twitter: Time, Individualism and Power. In *The 7th International AAAI Conference on WebLogs and Social Media (ICWSM), 2013.* [Honorable mention]

- Ruth García-Gavilanes, Yelena Mejova, Daniele Quercia. Twitter ain't Without Frontiers: Economic, Social, and Cultural Boundaries in International Communication. In *The 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW), 2014.*

Finally, Chapter 9 summarizes the findings of this thesis and presents ideas and directions for future steps.

Part I

Background

In this part we present and overview of the state of the art on microblogging user behavior in microblogs (with emphasis on Twitter) and we propose cultural models to be applied on the data.

Chapter 2

Microblogging

2.1 Introduction

Microblogging is a form of blogging that has become increasingly popular over the last few years. It has generated large amounts of social and geographical data that triggered unprecedented opportunities to study human behavior through data analysis in a spatial-temporal-social context. As the largest microblogging platform, Twitter, in particular, is convenient for research because of the volume of publicly available information. Another relevant microblogging platform is Sina Weibo, launched in 2009 and considered the *Chinese Twitter* by westerners due to its popularity in China. Sina Weibo and Twitter currently have similar functionalities and both provide access to their micro-posts via APIs [39]. Nevertheless, at the beginning of this thesis, we found that the official documentation of Sina Weibo was unclear and inconsistent in some extent and with a lot of restrictions.[1] For this reason, we focused only on Twitter data. In this chapter, we explain Twitter conventions and the related work that we consider particularly relevant on user behavior. Finally, we describe the data sets that we have used in this thesis.

2.2 Twitter

Twitter is a microblogging service that enables users to send and read short 140-character text messages, called "tweets". Launched in 2006, Twitter has grown from thousands of users in 2007 over millions in 2009 to hundreds of

[1] Nowadays a guide is available at http://www.cs.cmu.edu/~lingwang/weiboguide/.

millions in 2013. In this platform, users (Tweeters) choose between keeping their profiles and activity (*tweets*) public or private. Users with private profiles make their information available only to a selected group of friends while users with public profiles allow their tweets to be visible to anyone in Twitter. Users can follow other users and get notified with the tweets they post. Follower links are often not reciprocated [24] and often the followee (user who is followed) can be an organization or a celebrity. So, in this sense, it is more correct to think of the followee as an information channel whose updates the user may be interested in subscribing to. Through the years, users have learned to use Twitter following certain conventions in their messages, for example a *hashtag* is any sequence of characters, without whitespace, preceded by the # symbol and is used as a way of explicitly tagging the relevant topics. A "mention" consists of any Twitter username preceded by the symbol '@' and the use of *re-tweets* is a way of supporting the content of a tweet posted by someone else. Likewise, the use of URLs (many times shortened) to share external information, etc. In this thesis, we explore all these conventions, the content of tweets, the sentiment of words, the online social network of users, etc.

2.3 Previous Research

We present next relevant topics on user behavior studies in Twitter and other popular social media sites.

Network structure. Online social networks have been studied extensively because structure is strongly related to the detection of communities and to how information propagates. Mislove *et al.* [95], for example, studied basic characteristics of the structure of Flickr, Orkut, LiveJournal, and YouTube, and found power-law, small-world, and scale free properties. The authors argue that the findings are useful in informing the design of social network-based systems. Kwak *et al.* [77], examined the Twitter network aiming to determine it's basic characteristics. One of their main findings is that Twitter does not properly exhibit a "traditional" social network structure since it lacks reciprocity (only 22% of all connections on Twitter were found to be reciprocal), so it behaves more like news media, facilitating quick propagation of news. Java *et al.* [68], on the other hand, studied the topological and geographical properties of Twitter's social network and observed that there is high reciprocity and the tendency for users to participate in communities of common interest, and to share personal information. Onnela *et al.* [106], on the other hand, present a study on a large-scale network of mobile calls

and text messages. They found no relationship between topological centrality and physical centrality of nodes within communities in the network, and examined differences amongst big and small communities.

Influence. One of the key questions relating to communities and network structure is influence. De Choudhury *et al.* [26] examine Twitter data and study how different sampling methods can influence the level of diffusion of information. They found that sampling techniques incorporating context (activity or location) and topology have better diffusion than if only context or topology are considered. They also observed the presence of homophily, showing that users get together with ''similar" users, but that the diffusion of tweets also depends on topics. Cha *et al.* [24] studied the in-degree and out-degree of the Twitter network and observed that influence is in fact not related to the number of followers, but that having active followers who retweet or mention the user is more important.

Content. Information propagation also depends on the activity of users and the content of the tweets. Naaman *et al.* [98] examine the activity of users in Twitter and classify the type of messages produced, based on whether the tweets refer to the users themselves or not. They found that the majority of users focus on the ''self" while a minority on sharing information. This clearly relates to questions regarding what sentiments, if any, are expressed in tweets. Recent studies have focused on the analysis of the "happiness" level of tweets. For example, Dodds *et al.* [32] measure temporal happiness levels in Twitter using the "Affective Norms of English Words" (ANEW) lists from Bradley and Lang [18]. They introduce the concept of "weighted average happiness level" and calculate at different temporal scales (monthly, daily, etc). Bollen *et al.* [15] study sentiment in Twitter and show that there is a happiness assortativeness beyond demographic features such as age, sex and race, and conclude that even psychological states such as "loneliness" can be assortative in a social network. Finally, Hong *et al.* [62] present a study of differences in the way people tweet depending on the language used.

Evolution of users and behavior. Liu *et al.* [86] studied the evolution of Twitter users and their behavior by using a large set of tweets between 2006 and 2013. They quantify a number of trends, including the spread of Twitter across the globe, the shift from a primarily-desktop to a primarily-mobile system, the rise of malicious behavior, and the changes in tweeting behavior. The main part of this study is based on the accumulative number of tweets. We address, instead, the evolution based on individual users' behavior.

2.4 Data

The data sets used in this thesis were crawled from *Twitter*. The data was extracted from the public APIs during 2010, 2011 and 2013. The information we were interested to collect includes the user id, the screen name, the information in the location field of the profile, the date stamp of the tweet, the number of followers and followees, the *id* and the text of the tweet. To find the geolocation of user, we use the *valid location* specified in their profile. Often these locations are either strings specified by the users themselves or GPS coordinates coming from their mobile devices. We considered as a valid location any GPS coordinate or text which could be parsed correctly into latitude and longitude (using the Yahoo! public PlaceMaker API).[2] Examples of these natural language locations are: New York, NYC, Canada, CA, etc.

For some of the subsequent chapters, we focus only on the subset of users with geo-location information. Moreover, we used the Hadoop framework and Pig Latin to analyze data and other tools or dictionaries for language detection. All processing was anonymous and aggregated. No personal user information was used. We describe the details of each data set separately in each chapter.

[2] http://developer.yahoo.com/geo/placemaker/.

Chapter 3

Cultural Models

3.1 Introduction

The concept of culture is used in many ways and it has different meanings. Some use it to refer to civilizations and others as the refinement of the mind relating it to education or art. However, the most widely accepted and used definition of culture in science comes from the studies of Geert Hofstedes. In simple words, he explains culture as a dimension that distinguishes members of one group or categories of people from others. For example, the comparison of behavior between individuals born and raised in the Unites States with individuals born and raised in Japan. The figure 3.1 is based on a pyramid discussed on [61] where it is emphasized that culture involves studying and comparing the collective behavior learned in different societies as opposed to personality, focusing on the study of individuals and their inherited and learned characteristics as opposed to human nature which focuses on the universal inherited characteristics of humans. When people demonstrate differences or similarities, it is easy to confuse these levels because their influences combine, making them difficult to distinguish.

In some chapters of this thesis, we focus on the differences and similarities of the *collective behavior* manifested online by users of different countries. The collective behavior is obtained by aggregating data according to the geolocation of users. We test the associations between these values with formal anthropological studies of national culture. In order to do so, we need to count with studies that make rankings available to different countries around a particular behavior. Once counting with rankings, we can also explain findings with theoretical cultural studies such as *Huntington's*

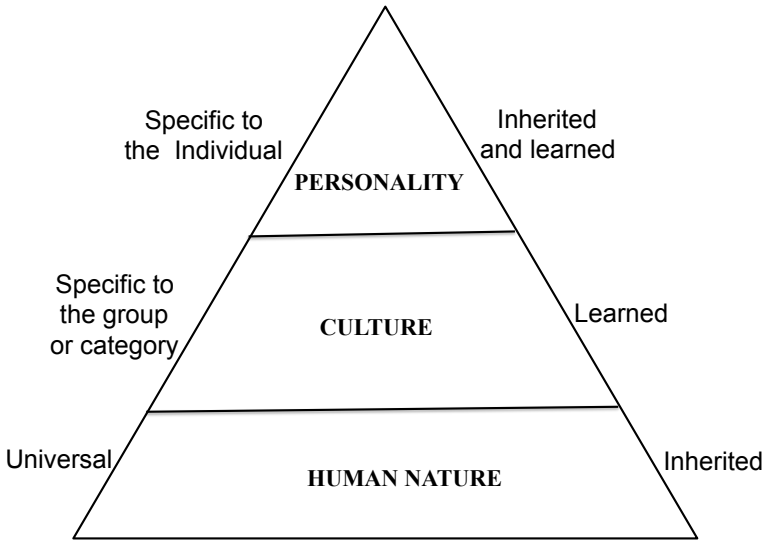


Figure 3.1: Pyramid showing that culture involves studying the collective behavior as opposed to the individual or human nature.

Clash of Civilizations [65] or Hall's Polychronic and Monochronic Tendency Model [55]. To begin with, we find that two particular studies provide these rankings: Hofstede's *Cultural Dimensions* [61] and Levine's *Pace of Life* [84] which are explained in the next sections.

3.2 National Cultural Dimensions

Hofstede is most well known for his work on dimensions of cultural variability, commonly referred to as "Hofstede's Dimensions." His first study involved four dimensions and later he expanded it to six dimensions. These dimensions were the results of detailed interviews carried out between 1978-83 with hundreds of IBM employees in different countries. He was able to determine patterns of similarities and differences among the replies to the interviews. From this data analysis, he formulated his theory that world cultures vary along consistent, fundamental dimensions. Since his subjects were constrained to one multinational corporation', and thus to one company culture. He explained the answers differences to the effects of their national cultures. He highlighted essential patterns of thinking, feeling, and acting that are well-established by late childhood. These cultural differences

manifest themselves in a culture's choices of symbols, heroes/heroines, rituals, and values. Hofstede identified six dimensions and rated the countries on indices for each dimension, normalized to values (usually) from 0 to 100. Thanks to these indices assigned to each country, we are able to associate human behavior manifested online to Hofstede's cultural dimensions. This question arises when cultural behavior is identified and available in data and there is lack of consensus to associate findings to a particular cultural dimension. In this study, we use Hofstede's six cultural dimensions as our basic framework for analysis. In the following paragraphs we introduce an explanation for each dimension and a small example of how Twitter data can be related to each one. The normalized indexes are found in [61] for around **76** countries and readers can access comparisons between two given countries in Hofstede's web page.[1]

- **Power Distance (PD)**- high vs. low: the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally. In a country classified as small-power-distance country, people would tend to easily approach and contradict their bosses. On the contrary, people from high-power-distance countries will unlikely approach and contradict their bosses directly. For example, several studies have been made on the popularity of users in social media such as Cha *et al.* [24], one could explore the level of communication between users with high, average and low popularity for certain countries and relate these findings with PD scores.

- **Individualism (IDV)**- high vs. low: the level of integration into a group. Individualist societies (high IDV) represent loose ties between individuals where everyone is expected to look after him or herself and his or her immediate family. Collective (low IDV) societies represent individuals integrated into strong, cohesive groups where they protect each other in exchange of loyalty. For example, we can explore this dimensions in Twitter by measuring the amount of conversation exchanged with others as well as comparing the unfollow dynamics of users.

- **Gender (Gn)**- masculinity vs. femininity: the description of roles between genders in a society. Femininity represents societies where emotional gender roles overlap: both men and women are supposed to

[1] http://geert-hofstede.com/countries.html.

be modest, tender, and concerned with the quality of life. Masculinity represents societies where emotional gender roles are clearly distinct: men are supposed to be assertive, tough, and focused on material success, whereas women are supposed to be more modest, tender, and concerned with the quality of life. For example, this dimension could be explored in social media by studying the use of adjectives and superlatives between women and men per country.

- **Uncertainty avoidance (UA)**- high vs. low: the extent to which members of a society feel threatened or uncomfortable in novel, surprising or unknown situations. This dimension can be measured by nervous stress, anxiety and the need for written or unwritten rules [61]. For example, measuring words expressing anxiety in social media can be an alternative to explore this dimension.
- **Time orientation (LTO)**- short vs. long: in essence, short-term oriented societies stand for the fostering of virtues related to the past and present (tradition, saving one's face, fulfilling social obligations) while long-term societies are more concerned with virtues oriented toward future rewards (perseverance and thrift). For example, this dimension could be studied in social media by identifying affiliations of users to certain groups. Many social sites are not only made of users but also groups representing ideologies, beliefs, etc.
- **Indulgence versus Restraint (IVR)**: Indulgence is related to societies allowing free gratification where there is a high correlation between happiness, life control and importance of leisure as personal value. On the other hand, restraint refer to societies that control more gratification by means of strict social norms. In social media, we can combine studies made on happiness in blogs and microblogs such as in the work of Dodds *et al.* [31; 32] with this dimension.

In Chapters 7 and 8, we show that the way we tweet can be associated with Individualism and Collectivism and how these dimensions help us predict the International communication strength across users from different nationalities in Twitter.

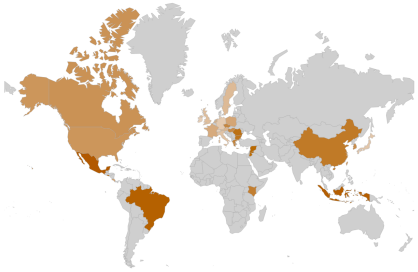
3.3 Pace of Life

An additional aspect that varies across countries, and that was not covered by Hofstede, is the perception of time across countries. Although the move-

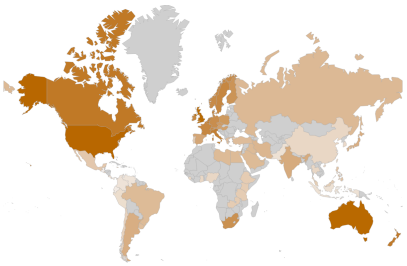
ment of time that people experience is extremely subjective, adjusting to a foreign perception of time can pose as many difficulties as learning a foreign language. Robert Levine completed a series of studies comparing what he called "pace of life" [84] in 31 different countries from throughout the world. He defines *Pace of Life* as "the flow or movement of time that people experience". As far as we know, there have been no other studies on "peace of life" on more than 31 countries. Similarly to Hofstede, the results of these experiments coupled with research findings from other social scientists. We chose Robert Levine in particular because his study counts with "pace of life" values assigned to 31 countries. In each of the 31 countries, Levine's students went into one or more of the major cities in order to measure the following three indicators of rhythm of life:

- **Walking speed**: There has been a lot of research into the speed of walking in cities and how it relates to economy [144; 94]. For this reason, Levine also measured the average walking speed of randomly selected pedestrians over a distance of 60 feet (around 18.2 meters). Several variables were controlled, such as to include pedestrians (no handicapped or window shopping) that could potentially walk at their own preferred maximum. A minimum of 35 walkers of each sex were clocked in each city. For example, they found that pedestrians in Rio de Janeiro walk only two-thirds as fast as do pedestrians in Zurich, Switzerland.
- **Work speed**: This experiment was measured by keeping track of the total time it took for postal clerks to fulfill a standard request for stamps. In each city, they presented clerks with a note in the local language requesting a common stamp and money was handed to them (a bill of $5). All measurements were taken during main business hours in main downtown areas under similar conditions. Researchers faced several difficult situations due to the way the experiment was carried out (a note and a bill) but overall they found that Western Europe was the fastest.
- **Clock Accuracy**: To measure the value of punctuality, they observed the accuracy of 15 randomly selected bank clocks in main downtown areas in each city. The times on the 15 clocks were compared to those reported by the phone company.

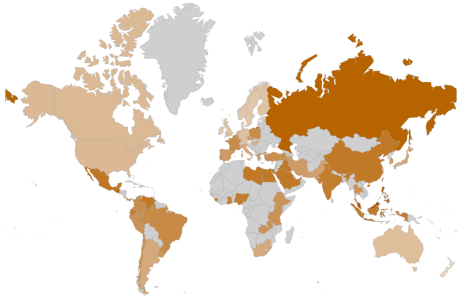
The three scores for each country were then statistically combined into an overall pace-of-life score. Levine combined the results from his experiments



(a) Pace of life ranking



(b) Collectivism vs. Individualism



(c) High vs. Low Power Distance

Figure 3.2: Maps showing (a) Levine's Pace of Life ranking, (b) Hofstede's Individualism and (c) Power Distance. Darker colors reflect lower Pace of Life, higher Individualism score and higher Power Distance. Gray areas mark countries that have not been included in Levine's study or Hofstede's.

and past research and concluded that people "are prone to move faster in places with vital economies, a high degree of industrialization, larger populations, cooler climates, and a cultural orientation toward individualism". The *Pace of Life* ranking can be found in [84] and we present these values in Table 3.1.

Figure 3.2 shows a world overview of countries that are believed to be more individualist or more collectivist (3.2b) and with higher or lower power distance (3.2c) based on Hofstede's country scores. The figure also shows which countries score higher in pace of life based on Levine (3.2a). In Chapter 7, we will associate user behavior in Twitter with these cultural dimensions.

Country	Overall Pace of Life	Walking Speeds	Postal Times	Clock Accuracy
Switzerland	1	3	2	1
Ireland	2	1	3	11
Germany	3	5	1	8
Japan	4	7	4	6
Italy	5	10	12	2
United Kingdom	6	4	9	13
Sweden	7	13	5	7
Austria	8	23	8	3
Netherlands	9	2	14	25
Hong Kong	10	14	6	14
France	11	8	18	10
Poland	12	12	15	8
Costa Rica	13	16	10	15
Taiwan	14	18	7	21
Singapore	15	25	11	4
United States	16	6	23	20
Canada	17	11	21	22
South Korea	18	20	20	16
Hungary	19	19	19	18
Czech Republic	20	21	17	23
Greece	21	14	13	29
Kenya	22	9	30	24
China	23	24	25	12
Bulgaria	24	27	22	17
Romania	25	30	29	5
Jordan	26	28	27	19
Syria	27	29	28	27
El Salvador	28	22	16	31
Brazil	29	31	24	28
Indonesia	30	26	26	30
Mexico	31	17	31	26

Table 3.1: Levine Pace-of-life scores. The 2nd column is the result of combining the scores of the 3rd to 5th columns.

3.4 Monochronic versus Polychronic Cultures

How can we associate the perception of time of people with the way users tweet?. To do so, we depart by assuming that there is a link between pace of life and temporal predictability from the finding that countries with higher pace of life tend to schedule their time in more predictable ways [142]. From this assumption, we can measure the predictability of when people will likely tweet and associate our findings with Levine's "Pace of Life" indexes. This assumption is possible thanks to the ethnographic studies of Hall [55] regarding cultures and their focus on monochronic or polychronic time.

Monochronic time refers to paying attention to only one thing at a time. In monochronic cultures, people tend to schedule their activities in a linear way, tend to be less flexible, and perceive time as a measurable, quantifiable entity, something with real weight and value. Moreover, monochromic time stresses adherence to present schedules and completion of tasks over social relationships. For monochromic people, time is a resource that is bought, spent, wasted, and inevitably deleted as we move through life [131]. For these reasons, monochronic countries are also considered more predictable. Comparing these characteristics to Levine's work, we can hypothesize that people from countries with high pace of life tend to be more monochronic or predictable. In Twitter, monochromic users will likely tweet in similar hours everyday.

In contrast, in polychronic countries, people are more flexible with time, adapt their schedules to others' needs, and see time as a general guideline, something without substance or structure. In polychromatic cultures, appointments are not taken seriously and, as a consequence, are frequently broken. For polychromic people, "time is seldom experienced as wasted and is apt to be considered a point rather than a ribbon or road" [55]. Consequently, polychromic countries are less (temporarily) predictable. We compare these characteristics to Levine's work and hypothesize that people from countries with low pace of life tend to be more polychronic or predictable. In Twitter, polychronic users will likely tweet in random hours everyday.

Similar assumptions have already been considered before in Social Computing. For example, Reinecke *et al.* [112] showed that the number of polls in *Doddle* per internet user correlates with Levine's pace of life country ranking associating their results with Hall's polychromic versus monochromic time orientation. They assume that Doodle users are primarily from those cul-

tures that are highly concerned with time. In Chapter 7, we show that this also holds true for predictability.

3.5 Culture and Socio-Economic Indicators

So far we have presented different cultural models to study how culture is related to the users' behavior in Twitter. Similarly, the relationship between national culture, economic and social country indicators has been studied extensively in the literature [29; 124]. These studies explain the causes and nature of within and between country variations in cultures, showing that culture is shaped by a variety of individual and country level factors. We build on the research by exploring the relationship between national culture, social and economic indicators *and* the aggregated results of Twitter's user behavior. We show that similarly to culture, the effect of economic and social indicators also influence how people use Twitter by using this data in correlations and predictions. To obtain economic and social indicators we have visited the sites of following entities: the US's Central Intelligence Agency, the WorldBank API for R, the World Values Survey and OpenFlights tool. We explore these relationships in Chapter 7 and 8.

Part II

User Behavior

In this part we present three studies on user behavior. First we explore human generated recommendations in Twitter. Second, we analyze how users evolve over time. Finally, we present a cross-country study of the similarities and differences of microblog usage.

Chapter 4

Human-generated Friendship Recommendations

4.1 Introduction

As seen in Chapter 2, in some networks, such as Twitter, connections need not be reciprocal, and any user is free to *follow* any other user with a public profile, to be able to see their posts or status updates. Since users are allowed to follow people they do not know, an important question is who *else* they should follow, in particular people who might be sources for the type of information they are interested in. In response to this need, *Follow Friday* emerged in 2009 as a spontaneous behavior from the Twitter user base, inspired by a blog post of an influential blogger:[1] users post tweets with the *#followfriday* or *#ff* hashtag, and include the usernames of the users they wish to recommend. As the name suggests, by convention these recommendations are made on Fridays. The key idea behind *Follow Friday* is that people you already follow should be able to suggest new contacts that you will be interested in following.

In 2009 and 2010, in particular, the popularity of these hashtags on Twitter rose considerably, up to the point that the Twitter hashtags *#followfriday* and *#ff* were among the most popular hashtags observed in several large-scale Twitter corpora [126; 109].

Although Twitter now has an automatic recommender system for contacts, the analysis of the dynamics of the *Follow Friday* phenomenon is interest-

[1] http://mashable.com/2009/03/06/twitter-followfriday/.

ing from multiple perspectives. From the angle of complex systems analysis, measuring the effect that collective recommendation processes have in driving the connectivity choices of individuals is very valuable to quantify the ability of a system to self-organize. Additionally, our analysis identifies features that are most predictive of tie formation in a peer-to-peer link recommendation process. This is useful on the one hand to alleviate the information overload of users receiving recommendations from their peers, by identifying the 'strongest' recommendations among hundreds or even thousands, and on the other hand to improve the design of automatic contact recommendation algorithms.

We focus on the dynamics of *Follow Friday* as a form of broadcast recommendations, making the following main contributions:

- We analyse for the first time the dynamics of a large-scale human-driven recommendation system and, by comparing it with two baseline conditions, we measure its impact on the process of follower-link creation. We find that recommended users have a chance of being followed that is roughly two or three times higher than a recommendation-free scenario. We also measure how long the recommendation effect lasts, as well as the effect of repeated recommendations and the longevity of the accepted recommendations (*i.e.*, how long these follower links persist).
- We develop a recommender system for ranking the human-generated recommendations received by a user. We evaluate this system against a corpus of known 'accepted' recommendations, identifying the features that are more predictive of link creation. Our recommender achieves a MAP of around 0.52, which is extremely high given the sparsity of the link recommendation problem. To the best of our knowledge, this is the first friend recommender system built and evaluated on human created recommendations.

The rest of this Chapter is organized as follows. In the next section we summarize related work, followed in Section 4.3 by a description of the data set and a summary of key terminology. In Section 4.4 we analyze the *Follow Friday* phenomenon along a number of dimensions, and quantify the extent to which it has a real effect on users' *following* behavior. We then, in Section 4.5, propose and evaluate a recommender system for ranking a user's received *Follow Friday* recommendations. Finally, we discuss our findings in Section 4.6. This Chapter is based on [42].

4.2 Related Work

The study of user-generated recommendations based on *Follow Friday* tags lies between two streams of research on recommender systems: recommendation based on user-generated content and social link recommendation.

A number of studies have been done on friendship recommendations in the context of Twitter, for instance [57] compared collaborative filtering and content-based recommendation for the purpose of link recommendation on Twitter. In [40], authors presented a movie recommendation system that extracts information from a Twitter-like microblog platform for movie reviews. They profile 537 users and 1080 movies according to words and tags, and offer content-based and collaborative-filtering recommendations. Several aspects of user profiles have been studied for recommendations, for example [1] propose a methodology for modeling Twitter user profiles to support personalized news recommendation. They compare profiles constructed from the complete long-term user history with profiles based only on users' most recent tweets.

The task of predicting link formation (or deletion) in social graphs is one of the major challenges in the area of link mining, and has been well studied in the last decade [85; 87]. Approaches have been proposed based on attributes of the nodes [78], structural graph features [88; 147], or both [2]. Unlike most of the work on link prediction that tries to predict future links in balanced sets of positive and negative samples, we are interested in a variant of link prediction, namely *link recommendation*, that is strictly user-centered and aims to provide a list of contacts to a user with the objective of maximizing the acceptance rate. Due to its inherent sparsity, this problem is more difficult than general prediction, and it has received little attention so far [6]. Previous studies also investigated what are the most predictive network and profile features for link formation in Twitter [67].

Despite the previous work in the area, we are not aware of any other attempt at characterizing human-generated recommendations and to leverage them to provide automatic contact suggestions. We also quantify the power of different features in predicting the formation of new links, not just considering structural or profile features of the user accounts, but focusing also on features that are descriptive of the human-driven recommendation process, such as the characterization of the relationship between the different human parties involved: the user who produces the recommendation, the one who receives it, and the one who is recommended.

4.3 A Data Set of Broadcast Friend Recommendations

As seen in Chapter 2, in Twitter, users can follow other users and have access to the tweets they post in their own accounts. In this chapter, for convenience, we will refer to the follower-followee relationship as a friendship relationship, although strictly speaking this relation only occasionally represents a true friendship: follower links are often not reciprocated [24]. We define *Follow Friday* recommendations as broadcast mentions of usernames in tweets containing the hashtag *#followfriday* or *#ff* (case-insensitive).[2] So, for example, the tweet **''#followfriday @Lula and @Obama for being such great leaders''** recommends people to follow the Twitter users `Lula` and `Obama`.

In March 2011, using the Twitter stream API, we randomly selected a *seed set* of 55K users. To remove profiles that are unlikely to be legitimate or active, we follow the approach of Lee *et al.* [81] and exclude users who have more than 1000, or less than 100, followers or followees. This filter also excludes celebrities, who usually do not interact with other users [77]. This choice was made to not exceed the limit of the API calls at that time. It also has the added benefit of filtering out less legitimate (*e.g.*, spam) users, since, according to Lee *et al.* [81], the majority of spam users tend to have out-degree and in-degree outside the range of [100; 1000]. Also Kurt *et al.* [129] showed that 89% of users following spam accounts have fewer than 10 followers. So, while we cannot guarantee that our data set does not contain spammers, previous studies indicate that our sample will indeed have a higher probability of containing mostly legitimate users.

Next, we monitored the evolution of the seed users' followees over time by collecting snapshots of the seed users' contact networks during a 24 week period from March 24^{th}, 2011 to September 5^{th}, 2011. The snapshots were taken twice a week, every Thursday and Monday yielding a total of 48 network snapshots. This choice is motivated by the fact that, although the recommendations are mostly broadcast on Fridays (76%), there is still a non-negligible amount of recommendations broadcast on Saturday (14%) and Sunday (3%), therefore Thursday and Monday snapshots can describe the status of the network right before and right after the recommendation takes place.

[2] We use the term *Follow Friday* to refer to the use of either of these *Follow Friday* hashtags.

Set name	Total
Initial seed	55,000
Receivers	21,270
Recommenders	589,844
Recommended Users	3,261,133
Recommendation Instances	59,055,205
Accepted Recommendation Instances	354,687
Rejected Recommendation Instances	58,700,518

Table 4.1: Unique # of Receivers, Recommenders, Recommended Users, Recommendation Instances, and Accepted and Rejected Recommendations.

In the remainder of this chapter, we use the following terminology:

- **Receivers (Rcv)**. Users from the initial seed set who accepted at least one *Follow Friday* recommendation at any time during the 24 week period.
- **Recommenders (Rdr)**. The followees of the *receivers* (*Rcv*) who made at least one *Follow Friday* recommendation during the 24 week period.
- **Recommended users (Rdd)**. The users mentioned after the *Follow Friday* hashtag in the messages of the *recommenders* (Rdr).
- **Recommendation Instance (Rec)**. The tuple $\langle rdd, rdr, rcv, w\rangle$ identifying an instance of a recommended user, made by a recommender, and exposed to a specific receiver in a given week (w). We use lowercase letters to identify elements in the actors sets (*e.g.*, $rdr \in Rdr$)
- **Acceptance**. We consider a *recommendation instance* made at time t to be *accepted* if its receiver becomes a follower of the recommended user between time t and time $t + \Delta$. Unless stated otherwise, the Δ considered is one week. Although we use the term *acceptance*, we cannot be sure about the causal relation between recommendation and acceptance (see discussion in Section 4.6).
- **Rejection**. We consider a recommendation instance made at time t to be *rejected* if the receiver does not follow the recommended user between time t and $t+\Delta$. Recommended users who are already followees of the rcv are not considered in the analysis.

Table 4.1 summarizes the quantities of followers, receivers, recommenders, recommended users and recommendation instances in our data set.

4.4 Analysis of Broadcast Recommendations

During the 24 weeks captured by our data set, we have a total of 144,180 *unique* new followees, from 354,687 *Follow Friday* accepted *recommendation instances*: this means that, on average, for accepted *recommendation instances*, the *receiver* got recommendations to follow the same *recommended user* from 2 distinct *recommenders*. Table 4.2 shows the acceptance rate (the number of accepted recommendations divided by the number recommendation instances) for recommendations under various conditions where one of the actors involved in the *recommendation instance* mentioned one of the others in the previous week (using the '@username' convention). The first column indicates the direction of the mention and the users involved. The case of $rdr \rightarrow rdd$ (recommender mentions recommended) involves all recommendations since this is a necessary condition of a *Follow Friday* recommendation.

We can see that, overall, the acceptance rate is very low at 0.006 (*i.e.*, 0.6% of recommendations are accepted), which is to be expected, since the recommendations are broadcast, as opposed to being personalized, and may not even have been seen by the receiver. When one of the actors mentions another, the acceptance rate tends to increase, which is expected, since these mentions are indicators of an active relationship. When either the *recommended user* (rdd) or the *receiver* (rcv) mention each other, the acceptance rates are roughly 10 times higher than the average (10% to 14% of recommendations accepted), which is not surprising since it shows that there is already a connection between these two users who form the new link. Note that while the acceptance rate for these particular cases is relatively high, the volume is low, indicating that these cases of pre-existing relationships are not typical of *Follow Friday* recommendation acceptances.

4.4.1 Effect of #FF recommendation

Since *Follow Friday* is a spontaneous recommendation phenomenon, the first question that arises is whether it has an actual impact on the creation of new follower links, and to what extent. In complex social systems, determining the causes of observed evolutionary phenomena is a very challenging task, due to the intrinsic difficulty in disentangling all the factors that produce

Mentions	Volume	Proportion	Acceptance Rate
rdr→rdd	59,055,205	1.000	0.006
rdd↔rdr	4,667,056	0.079	0.009
rcv→rdr	9,071,311	0.154	0.010
rdr→rcv	9,199,224	0.156	0.011
rdr↔rcv	6,242,059	0.106	0.012
rcv→rdd	205,447	0.003	0.095
rdd→rcv	238,822	0.004	0.097
rcv↔rdd	76,482	0.001	0.145

Table 4.2: Acceptance Rates for *Follow Friday* Recommendations, under various conditions where the users mention each other in the preceding week. For example 'rdd→rcv' indicates that the recommended mentioned the receiver, and 'rdd↔rcv' indicates that the recommended and receiver both mentioned each other (rdd→rdr is omitted because it is identical to rdd↔rdr in this data set: by definition, the recommender mentions the recommended for all recommendations).

the events observed in a-posteriori data-driven studies [117]. Even when controlled experiments are performed [3; 9], it is very difficult to know with absolute certainty which factors trigger the observed dynamics.

In our case, the inclusion of a new *recommended user* in the followee list cannot be interpreted directly as a cause-effect sequence, since the adoption may be driven by factors that are not related with the recommendation itself, such as unobserved online interactions or even exogenous events. Nevertheless, when sufficiently extensive temporal data is available, it is possible to compare the evolution of the system under different conditions, or null models [11; 116], to understand if the the target factor has an effect, distinguishable from the other conditions, on the evolution of the system.

Specifically, we measure the added value of *Follow Friday* by comparing the acceptance rate of *#ff* or *#followfriday* recommendations with two alternative conditions:

- **Implicit recommendation model**: all usernames mentioned in any tweet received by users in the receiver set (*Rcv*) are considered implicit recommendations, based on the assumption that being exposed to the names of some users may increase the probability of adopting them as new followees. The implicit recommendations we consider are all mentions appearing in non-*#ff* tweets during the week before

the target week, and that never previously appeared as an explicit *#ff* recommendation (for the same *receiver*) in the 24 week sample.

- **Unobserved recommendation model**: for this model, we assume that, due to unobserved factors, the contacts recommended through *#ff* hashtags would have been added by the *Rcv* set even in absence of any explicit *#ff* recommendation. These unobserved factors could include, for example, the rising popularity of the recommended user or relevance of the topics discussed by the recommended user to external breaking events. To model this condition, we apply a temporal shift: for the set of *#ff* recommendations made at time t, we measure their acceptance rate at time $t-1$, before the actual recommendation is made, that is, we measure the acceptance rate in a situation where the external conditions are similar (one week previously), but where no *Follow Friday* recommendation has been made. To keep this model separate from the implicit one, we exclude cases where the receiver received implicit recommendations, up to time $t-1$, for the same recommended user.

The difference in the acceptance rate between the three models, depicted in Figure 4.1, shows that *#ff* recommendations lead users to follow a higher proportion of contacts compared to models in which *#ff* is not considered. Apart from an outlier at week 1, the margin between the *#ff* model and the two alternative conditions is large, with *#ff* having an acceptance rate always between two and three times that of the others. By disentangling the role of the presence of the *#ff* tag from other factors that play an important role in the creation of social links, mainly homophily, the comparison with these alternative conditions provides strong evidence that the recommendation has an effect on the probability of link creation.

Whereas homophily may have a role in the selection of a recommended profile among other recommended ones, it seems not to be the main reason for the recommendation acceptance itself. Since the unobserved condition is simulated by performing a one week temporal shift, if we assume that the homophily effect between two users is not likely to change drastically in this one week time frame, then if the probability of acceptance is mainly determined by homophily, the *#ff* and *unobserved* conditions would have similar acceptance rates. The fact that this is not the case is, we believe, strong evidence that the *#ff* recommendation, and not purely the similarity between the profiles, drives the creation of the new link. Of course, there

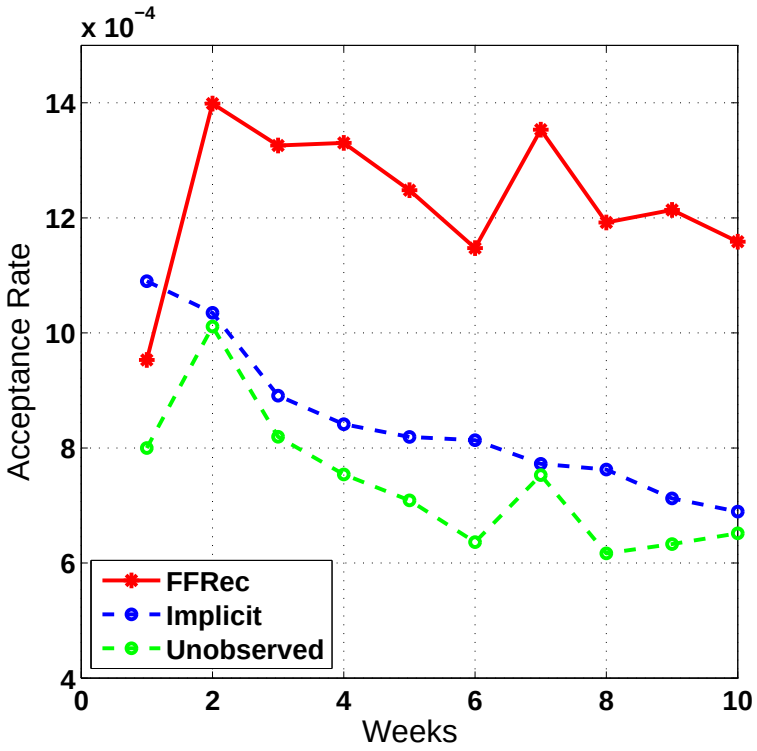


Figure 4.1: The acceptance rate for *Follow Friday* recommendations in different weeks, compared with implicit and unobserved recommendation models.

may be cases where the homophily effect changes drastically over the one week time shift, but it seems unlikely that this would explain all of the recommendation acceptances in this very large corpus.

A slight decreasing trend in acceptance rates over time is observed for all conditions, most likely due to the effect of the residual signal of explicit and implicit recommendations from the previous weeks (*i.e.*, due to recommendations made before week 1 of our study, which we have no information about). To further verify this hypothesis, we measure how much the effect of an implicit or explicit recommendation lasts in time by computing the acceptance rate n weeks after the recommendation is made. To do so, we split our 24-week sample in half and observe the percentage of recommendations (from the first 12 weeks) that receivers followed up to 12 weeks after the recommendation was made. We do not consider cases where the rec-

ommendation was repeated after the week of the initial recommendation. Figure 4.2 reveals that the likelihood of subscribing to a recommended profile extends over several weeks and, after an initial substantial drop, fades slowly. We observe that the probability does not seem to stabilize even after 12 weeks. Even though the scenario in which a user remembers a *Follow Friday* recommendation after several weeks is unlikely (especially if the recommendation has not been repeated), the probability decay is evident. The reasons behind such a long-lasting decay are difficult to find, since over such a large time scale many other interconnected events co-occur in the network's evolution. We argue that the effect of the *#ff* recommendation may introduce a perturbation in the network structure that may lead to delayed adoptions.

For instance, a user who received a recommendation before, but did not accept it, may create the link later because other users in his neighboring network accepted it, leading to new opportunities for social triangle closure [83].

To go beyond the acceptance rate of recommendation, we now look at the longevity of the new social ties created as a consequence recommendations. Figure 4.3 shows the percentage of acceptances that were still in the receivers network after n weeks. The curve labeled as *Others* represents all the users that were followed for reasons not related to the conditions considered in this study. After 12 weeks, we can see that 83% of *#ff* links are still in the receiver's network, versus 80% of links that follow implicit recommendations, and 76% of other follower links. This is an important finding in an environment such as Twitter where social ties have been observed to be very volatile [78].

4.4.2 Repeated Recommendations

In social sites such as Twitter, it is likely that a single broadcast Tweet may not be seen by many of a user's followers. Repeated recommendations, therefore, are likely to increase the likelihood of a recommendation being accepted, because the follower is more likely to see the recommendation at least once, and because repeated viewings of the recommendation may reinforce it. Figure 4.4 plots the acceptance rate against recommendation repetitions, where repetitions are counted as recommendations received previously by a user within the time frame covered by the corpus. We consider two cases: when the recommendation is in the form of a *Follow Friday* recommendation only, and when there are only implicit recommendations. The

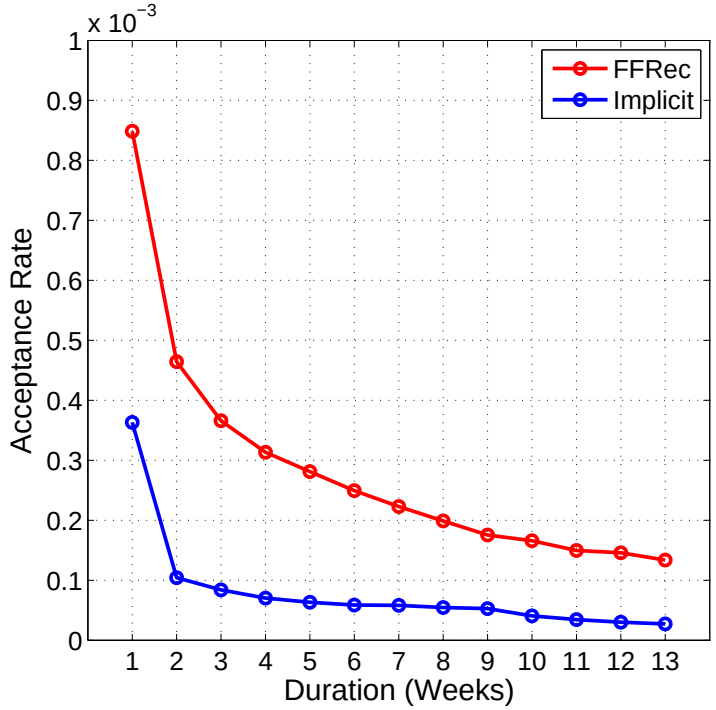


Figure 4.2: Acceptance rates n weeks after a recommendation is made.

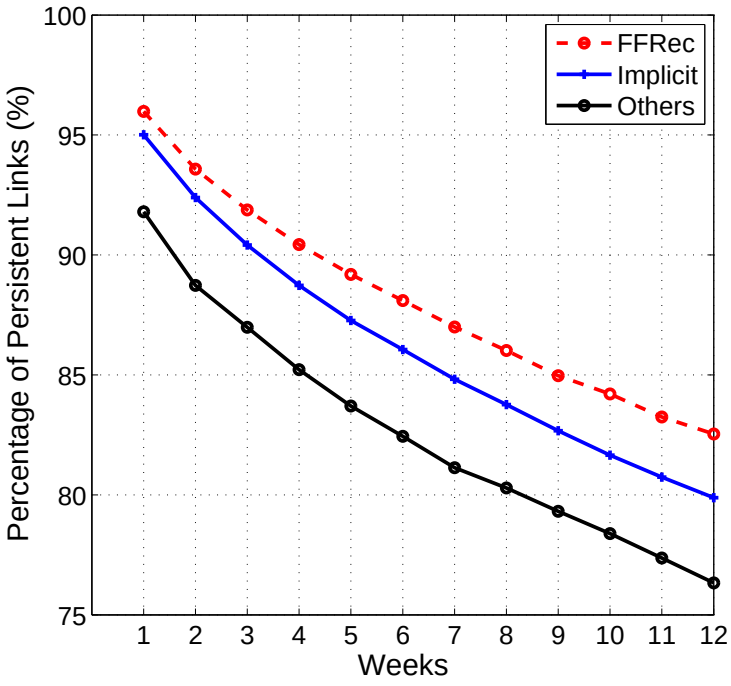


Figure 4.3: Longevity of accepted recommendations.

results show that repeated recommendations make a significant difference. We can also see that it takes many implicit recommendations to have a similar effect as even a single *Follow Friday* recommendation, with 15 implicit recommendation having a similar acceptance rate as 1 *Follow Friday* recommendation.

Figure 4.5 plots the acceptance rate versus the number of distinct recommenders who recommend the same recommended user a receiver, and it shows a similar increase in the acceptance rate as the number of distinct recommenders increases, but with a bigger gap between the *Follow Friday* recommendations and the implicit model.

4.5 Recommender System

In the broadcast recommendation setting given by *Follow Friday*, users are exposed to a large number of friend recommendations every week. In a situation of information overload, the 'good' recommendations are likely to get lost among noisy ones, therefore automated methods are needed to detect the most valuable recommendations. We envision a scenario where all recommendations received by a user in a given week are ranked such that the good recommendations are at the top of the ranking. This essentially corresponds to providing recommender service built on top of the human-generated recommendation system.

In the following, we verify that it is possible to rank Twitter friendship recommendations and surface the most valuable ones, and we evaluate the utility of various features for this task. Secondly, by analyzing the predictive value of different features for ranking recommendations, we supplement the analysis of the previous section, giving further insight into features that can predict the creation of a link after a recommendation is made.

4.5.1 Features for Ranking Recommendations

For each recommendation instance $\langle rdd, rdr, rcv, w \rangle$ we calculate a number of features, and group them into 3 main types: *user-*, *relation-*, and *format-based*.

User-Based features. These features describe an individual Twitter user, whether it be a *receiver*, a *recommender* or a *recommended user*. We identify two types of user-based features, attention-based and activity-based:

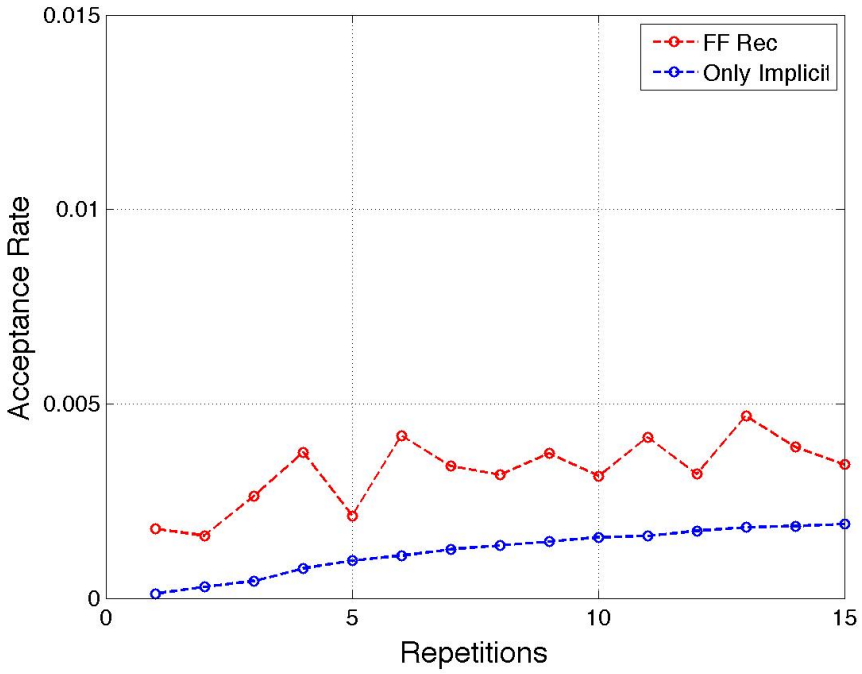


Figure 4.4: The number of repeated recommendations vs acceptance rate.

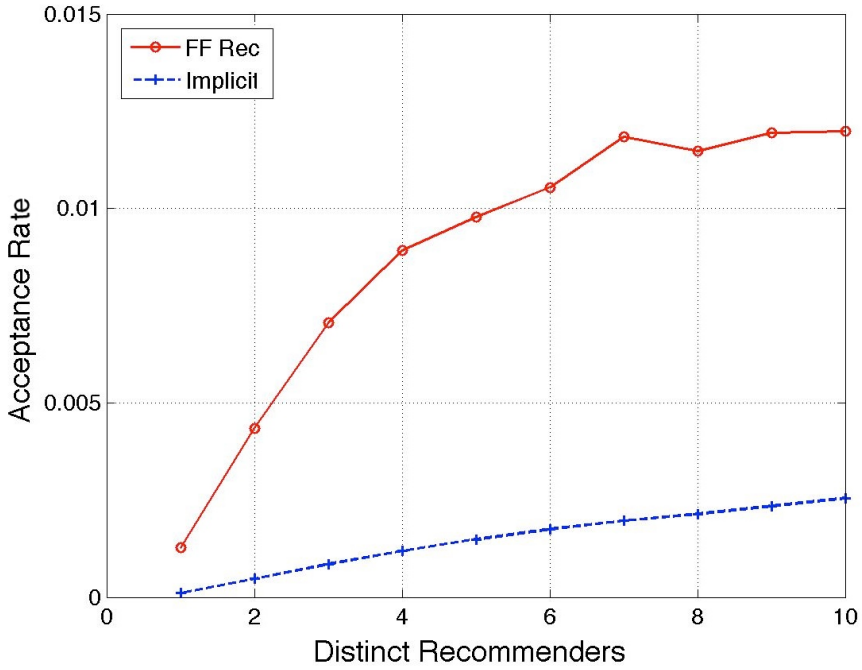


Figure 4.5: The number of distinct recommenders vs acceptance rate.

1. *Attention-Based features* are related to the level of attention given to the user by other users. We measure *popularity* ($followers/(followers+followees)$), the number of times the user has been *mentioned*, the number of people mentioning the user, the number of times the user has been recommended with a *Follow Friday* hashtag, and the number of distinct recommenders.

2. *Activity Based features* describe the level of activity of a user. We count the number of *new followees* of a given user for a given week, the *average tweets per day* of the user (over the entire history of the account), the number of *recommendations accepted* by receivers, and the number of distinct recommenders the a receiver has accepted recommendations from. Finally, we also count *mentions*, the number of distinct Twitter accounts mentioned by the user.

Relation-Based features. These features describe the relation between pairs of users in the $\langle rdd, rdr, rcv \rangle$ triple, based on either profile similarity or communication patterns.

(a) *Communication-based features* describe the level of communication between two users. *Conversational mentions* count the number of times a user mentions another user, and is calculated separately for each pair of users involved in a recommendation. The *number of Follow Friday recommendations*, (the number of times a user recommended another user) is calculated for each pair of users. We also count the number of previous acceptances between the *receiver* and the *recommender*, based on *Follow Friday* recommendations and on *implicit* recommendations. Last, we measure the friendship duration between the receiver and recommender by number of weeks.

(b) *Similarity-based features* describe the similarity of users. Separate *content-based similarity* features calculate similarity between all the tweets of two users, hashtags only, mentions only, and URLs only. All these features use the weighted Jaccard similarity coefficient, as in Sudhof *et al.* [125]. *Geograpical similarity* is a binary feature, set to 1 if actors are in the same country or 0 if not. The location is parsed from the users's declared location using the Yahoo! PlaceMaker API.[3]

[3]http://developer.yahoo.com/geo/placemaker/

Ranking	MAP
Rotation Forest	0.4959
Linear Combination	0.0565
Random	0.0368

Table 4.3: Recommendation Mean Average Precision using all features.

Format-based features. These features describe a recommendation with information of the profile of the users, based either on the context or the format of the recommendation.

(a) The *repetitions* counts the number times the recommendation has been repeated in a *receiver's* timeline, the number of distinct *recommenders* who made the recommendation, and the number of previous weeks in which the recommendation was received.

(b) *Context* features describe the format or the context of the tweets containing the recommendation. We consider the *day of week* on which the recommendation was made, we record whether the recommendation was made in a retweet or not, the number of other users appearing *together* with the recommended user in the *Follow Friday* tweet(s), and the *length of recommendation tweet* (the number of tokens in the tweet, excluding *#ff* hashtags and *@mentions*). Finally, we count the number of URLs in the recommendation tweets. Since many of these measures can have more than one value for a given user (*i.e.*, they receive the same recommendation from different people) we calculate both the maximum and minimum for all of them.

Most of the features are calculated over a temporal window prior to the recommendation. For all such features, we calculate two versions of the feature: (1) based on the one week period prior to the recommendation (to capture recent activity, similarity, etc), and (2) based on all previous weeks in the corpus (to capture longer-term activity).

4.5.2 Evaluation Methodology

We consider all the unique *Follow Friday recommendation instances* that a given receiver is exposed to at week t and rank them with the aim of putting the 'best' recommendations at the top of the ranking. For the recommender, we set $\Delta = 2$, meaning that we consider the acceptances within two weeks

of the recommendation, based on the ground truth of known acceptances. Recommendations accepted after two weeks are considered as not accepted.

Using the *acceptance* information as a ground truth, we evaluate our method by its ability to place the accepted recommendations towards the top of the ranking.

We split the data into training and test sets based on time, with data from weeks 1 to 16 used for training, and weeks 17 to 23 used for testing. We do not test against data from week 24, as we do not have details of the evolution of the followee network one week later.

We use two distinct methods to rank the *Follow Friday* recommendations received by a user in a given week: (1) based on a linear combination of the normalized scores from each feature, and (2) using the confidence score from a supervised classifier trained to classify recommendations as acceptances or rejections.To build the supervised classifier, we take a random subset of recommendations from the training set, ensuring that this subset contains a balanced set of acceptances and rejections. We train a binary classifier on this data using the Rotation Forest algorithm [113] as implemented in the WEKA library [47]. The Rotation Forest method constructs an ensemble of decision trees using random subspaces and principal components transformation applied to the input data [113]. For the linear combination of features, we normalize each list of recommendations by dividing by the feature's maximum value.

We do not normalize the similarity features based on the weighted Jaccard index, since those features are already normalized.

Since a receiver can receive many recommendations in a given week, and can accept one or more of them, we evaluate our various ranking approaches using the standard Information Retrieval measure *Mean Average Precision* (MAP). MAP evaluates a ranking by averaging the precision at the rank position where each relevant item is retrieved [30]. In the evaluation of friendship recommendation, an accepted recommendation is analogous to a relevant item, and a recommendation that is not accepted is non-relevant.

4.5.3 Results

Table 4.3 shows performance of the Rotation Forest classifier, compared against the linear combination and a random baseline. The linear combination performs very poorly, while the Rotation Forest gives encouraging

Features	MAP
All	0.4959
User-based	0.0741
Relation-based	0.3976
Format-based	0.0615
User + Relation	0.5176
User + Format	0.0790
Relation + Format	0.3787

Table 4.4: Recommendation performance for subsets of features (Rotation Forest).

performance, with a MAP of almost 0.5, showing that machine learning approaches can give good results for this task.

In Table 4.4, we show the results when using various subsets of features, according to the grouping of features described in Subsection 4.5.1. The *relation-based* features are the most discriminative for friend recommendation, while the *format-based* features are not useful at all, and always harm performance. Finally, the *user-based* features, while they do not perform particularly well on their own, improve performance when combined them with the *relation-based* features. Overall, the best performing set of features is *user-based* + *relation-based* (*i.e.*, ignoring the format-based features), with a MAP of almost 0.52.

Due to space we do not show detailed results for individual features, but the single best performing individual feature is the previous behaviour of the *receiver* in accepting recommendations from the *recommender*. Other relation-based features based on similarity (not communication) are also important, however, and the results in Table 4.4 show that optimal performance is achieved when we also consider user-based features.

4.6 Discussion

We describe the first study of the *Follow Friday* phenomenon, which aims to better understand the dynamics of a large scale collective process of human-generated link recommendations, and to understand the features and conditions that may predict the creation of new social links.

Furthermore, in contrast with other studies of link prediction in social media, we use a direct and reliable ground-truth of acceptances and rejections,

based on real user behavior. We compare acceptance rates of *Follow Friday* recommendations with baseline conditions where (a) another user is mentioned, without being explicitly recommended, and (b) we simulate a condition where there is no observed (explicit or implicit) recommendation made via Twitter. Through this comparison, we show that explicit *Follow Friday* recommendations have a large, measurable, effect on who users choose to follow on Twitter. We also show that the effect of a recommendation (explicit or implicit) lingers for a number of weeks, that repeating recommendations has a strong effect, and that ties formed after *Follow Friday* recommendations tend to have more longevity than other ties, an important finding in Twitter, where social ties are quite volatile.

To surface more valuable recommendations above others, we propose an automated recommender system based on a number of features, which we group into three distinct categories: *user-based*, *relation-based* and *format-based*. We show that the most discriminative features for friendship recommendation are those features based on communication and similarity between users. In particular, past behavior in following recommendations coming from a given recommender is the most predictive feature of future recommendation acceptance.

Chapter 5

Evolution of Microblogging Behavior

5.1 Introduction

As seen in Section 2.2 users in Twitter are used to follow certain conventions in their messages, like the use of the symbol @ (at) before a user, the use of *re-tweets*, shortened URLs (often shortened), etc.

As a consequence, Twitter is used in several contexts, for different audiences and with different purposes. In fact, scholars have argued that Twitter is used as an hybrid between a communication media and an online social network [77; 143]. Additionally, user behavior is not static, it changes through the years, the way the first Twitter users interacted with the platform when it started may differ from how they interact now. While the set of research using Twitter data has expanded rapidly, little work has studied the change/evolution in Twitter ecosystem itself.

In this chapter, we propose a step towards understanding the evolution of user behavior focusing on *how* people tweet and their audiences. To this end, we carry out a longitudinal study of tweets posted during 10 weeks in 2011 and 10 weeks in 2013 by more than 4M users who have been active in Twitter in both of these periods.

First, we propose a taxonomy of messages based on Twitter conventions (mentions, links, re-tweets). In doing so we obtained 6 tweet formats. To identify models of behavior, we cluster users based on these types of tweets and study how users change their behavior in time. To present our results,

we organize the chapterr as follows. Section 5.2 provides related work. Section 5.3 describes the data. In Section 5.4 we explain our methodology and the taxonomy given to the types of tweets. In Section 5.5 we report how user behavior changes in 2013 with respect to 2011. Section 5.7 looks at relationships between popularity and activity with the clusters, providing insights for future work. We finish with conclusions and next steps. This chapter is based on [44].

5.2 Related Work

The goal of this work is to study the variation of tweeting behavior across time based on a taxonomy of tweet types and audiences. In a similar way, researchers have already analyzed how a variety of aspects change across time in Twitter and other online platforms. They have studied the following aspects:

Audiences. Marwick and boyd [92; 107] claim that users in Twitter *imagine* their target audiences since they do not know "which few" will read their tweets. They find that users do not have a fixed target audience and that having one would be a synonym of "inauthenticity".

Behavior and clusters. Naaman *et al.* [98] find 4 relevant categories of tweets based on the content of the messages. For each one of these categories, they cluster users and find two types of users: Meformers (talking about one self) and Informers (sharing news). Luo *et al.* [89] classify tweets based on language and syntactic structure and Huang *et al.* [64] show that tagging behavior (hashtags) has a conversational, rather than organizational nature.

Many attempts have been done to classify users according to their audiences and tweet content. However, most of these studies are language-dependent and need manual labeling. In this work, we categorize audiences and tweet types using a language-independent approach.

5.3 Data Set

For the results we present here, we used part of the dataset described in Chapter 4 and in [42]. This is a dataset of 55K randomly selected users with number of followers and followees in the range of [100, 1000] and their corresponding followee network (for a user u, it contains all users who u is following) during 10 weeks.

	Full Data Set 2011	Active Users in 2011	Active Users in 2013
Users	8,092,891	4,350,583	4,350,583
Tweets	2,280,707,094	1,527,675,950	679,507,450
English Tweets	1,086,233,182	768,940,902	369,452,361
Active Users	1,868,150	1,315,313	1,125,968
Tweets (*)	1,248,300,919	880,889,333	375,741,789
English Tweets (*)	562,134,366	406,719,99	256,330,241

Table 5.1: The second column shows the full data crawled in 2011. The 3rd and 4th column show information of users who tweeted in *both* 2011 and 2013. From rows 2 to 4, we find information about active *and* inactive users. From rows 5 to 7, we find information of the active users *only*. Active users are those considered to have tweeted in English more than 55 and less than 1540 times. The (*) means that it is based on active users.

We then proceeded to collect all of the tweets posted in English by the original 55K users as well as their followees during 10 weeks starting from the second half of March 2011. By crawling the information of the followees, we attempt to target the typical accounts twitterers like to follow. It is mostly on these users and the 55K seed set that all our results are concerned.

Additionally, during the 10 weeks, we also crawled all the tweets containing the screen names of any of the previously geo-located users (which will help us to calculate some of the popularity metrics). In total, we obtained 8M geolocated users who tweeted around 2.4B tweets. We then crawled 10 weeks between October and December in 2013 looking for the same filtered users in 2011 and found that around 4.3M users tweeted at least once also in 2013. After the end of the crawling period, we identify the language in which tweets are written. We then proceed to classify as *active users* those who tweeted at least 55 and less or equal than 1540 tweets in English during 10 weeks to exclude inactive users and bots.In total we found around 538K users tweeting within this range in both years. We chose this range as to set a threshold of 1 tweet per working day (5 per week) and a maximum of 22 per day. The maximum limit was chosen based on a marketing study by Zarrella[1] where he finds that those with the most followers tweet an average of 22 times a day. With this we attempt to include users likely to be

[1] http://www.slideshare.net/HubSpot/the-science-of-timing.

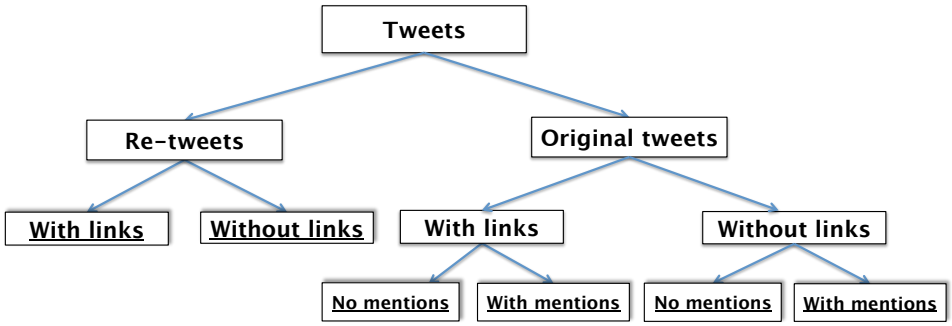


Figure 5.1: This classification tree represents the tweet formats used to classify users in different groups. The top groups include the tweets in the subsequent levels. The underlined nodes (leaves of the tree) are used in the clustering process (6 types).

engaged with the platform excluding those with an abnormal activity (*i.e.*, advertisers or bots). Table 5.1 shows the summary of the dataset used for the experiments. It is interesting to notice a higher proportion of active user among those users who tweeted in *both* 2011 *and* 2013 (the 5th row of the 3rd and 4th column) than those who tweeted in 2011 but not necessarily in 2013 (the 5th row of the 2nd column).

5.4 Methodology

As previously discussed in the related work section, some researchers argue that everybody has an *imagined audience* in a communicative act even if that act involves social media [92]. Given the various ways people consume and spread tweets, it is virtually impossible for Twitter users to account for their potential audience, although we often find users tweeting as if these audiences were bounded. For instance, the use of the @ sign before a user login name allows to "poke" that user which may trigger a reply and start dyadic conversations (through mentions) which are visible at the same time to others as well. In fact, Marwick and boyd [92] found, through interviews to twitterers, that sometimes users are "conscious of potential overlap among their audiences (*i.e.*, friends, family, co-workers, etc)." The authors report cases where users tweet to themselves, to fans, to fellow nerds, to super users, etc.

We propose a language-independent taxonomy of tweet types. The proposed types are based on the conventions established by Twitter such as

	Full DS 2011	2011	2013
Original tweets	77.30%	76.94%	74.77%
With URLs	14.93%	14.62%	18.74%
with mentions	6.39%	3.46%	4.16%
without mentions	11.36%	11.16%	14.58%
Without URLs	62.37%	62.32%	56.03%
with mentions	35.18%	35.36%	27.44%
without mentions	27.19%	26.96%	28.59%
Retweets	22.70%	23.06%	25.23%
With URLs	6.29%	6.75%	8.60%
Without URLs	16.41%	16.31%	16.63%

Table 5.2: Tweets from active users in 2011 and 2013, and the corresponding percentage of tweets that belong to each type.

the mention symbol @, the retweet flag and the URLs, *imagining* an audience through the combination of these symbols. Figure 5.1 shows these categories.

We start by classifying two main groups of tweets: retweets (RT) and original tweets (OT). Retweets refer to those tweets forwarded from other users. We hypothesize that a retweet targets the user who created the forwarded tweet and the followers of the user forwarding the tweet. Next, original tweets refer to tweets posted by users themselves and the audience could vary between the followers and the users themselves. For the RT and OT sets, we make two other distinctions: tweets with URLs and without URLs. We hypothesize that URLs target audiences who are willing to obtain information from the links posted and generally interested in exogenous stimuli. For tweets without URLS, users want to transmit a self-contained idea in maximum 140 characters. For the OT set we make yet another distinction, for the tweets with URLs and without URLs we divide them between tweets containing a mention (conversational) and those without a mention (textual). A OT containing a link with a mention implies that a user calls the attention of another user to open the link shared in the tweet. We do not make this last distinction (mention and link) for the RT set given than all retweets already refer to another user. In this study, we focus on the tweet types at the deepest level of each branch (6 in total): a) re-tweets with links, b) re-tweets without links, c) original tweets with links and no mentions, d)

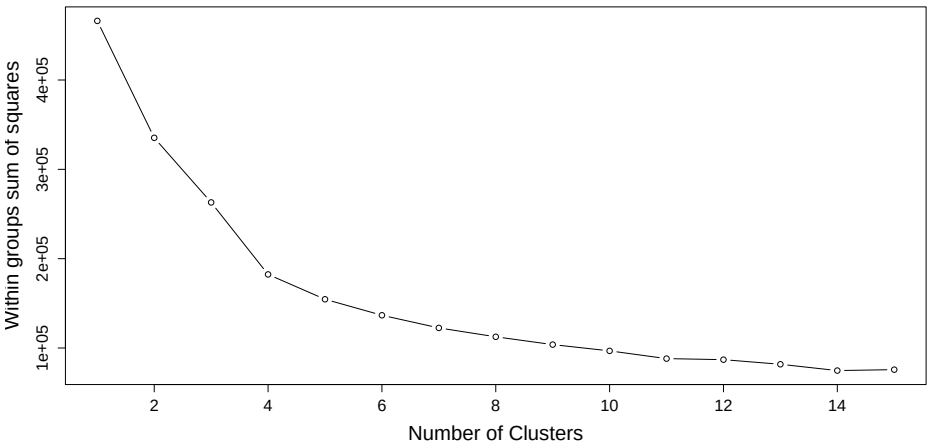
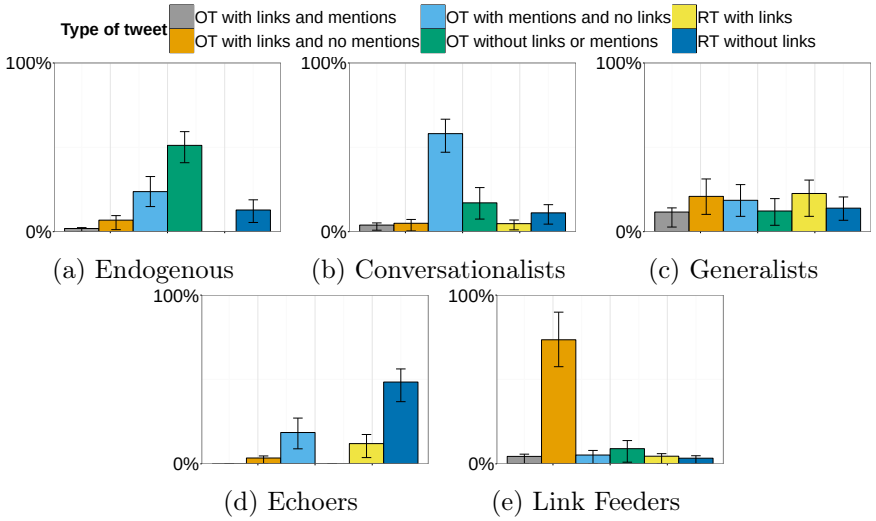


Figure 5.2: Elbow method for clustering: the *bend* lingers between 4 and 5.

original tweets with links and mentions, e) original tweets without links and no mentions and finally f) original tweets without links and mentions.

Based on this scheme, we classify the tweets of the *active* set of users (538 K) in 2011 and 2013 and find a slight increase in tweets with URLs in 2013 (from 14.62% to 18.74%). Table 5.2 has the percentage of tweets in each category for *active* users.

Furthermore, for each *active* pair (user, year) we calculate the percentages of tweets belonging to each of the tweet types. Each pair (user, year) is represented by a 6-dimensional vector, *6* being the number of all numerical features (the percentages) used to describe the objects to be clustered. We use the well-known k-means algorithm for clustering. To decide the k points in that vector space, we used the so called *elbow method*. This is a visual standard method [100] that runs the k-means algorithm with different numbers of clusters and shows the results of the sum of the squared error. The value of k is chosen by starting with $k = 2$ and increasing it by 1 until the gain of the solution drops dramatically, which will be the bend or elbow of the graph. This is the k value we want and is chosen visually. We found that the *bend* lingered between 4 and 5 (see Figure 5.2 in Appendix). We analyzed both cases and chose $k = 5$ because we observed that it best encapsulates interesting and distinctive patterns of tweeting behavior.



(a) Endogenous (b) Conversationalists (c) Generalists

(d) Echoers (e) Link Feeders

Figure 5.3: Clustering based on 6 tweet types posted by *active* users during 10 weeks in 2011 and 2013. The clusters appear from left to right according to their size in descending order. Each bar shows the average percentage of that tweet type. Error bars represent the interquartile range. Clusters (a) and (d) do not contain tweets of all types.

5.5 Results

We now proceed to the results and study how users have changed their tweeting behavior through time. Figure 5.3 shows the average composition of tweet type vectors in the clusters. The clusters are ordered by size and the bars indicate the interquartile range for each case. Note that we have abbreviated some of the names in the captions due to space concerns. We observe that each cluster has a dominant tweet type except for the third cluster (*Generalists*) that reports a balance among the tweet types.

We discuss now each of the identified patterns of tweeting behavior and relate them to the concept of the *imagined audiences* discussed in the previous section.

Endogenous: Users in this cluster mostly post and forward messages not linked to external information. Users in this cluster are supposed to use Twitter more as a social network than as a news media. The dominant type of tweets are self-contained posts created by the user herself without mentioning other users such as quotes, thoughts or even futile information.

In second place we observe original tweets with mentions which is a sign of conversation with other users.

Conversationalists: Users following this pattern are characterized mostly by tweets containing mentions with no links. Similarly to the *Endogenous* type, users in this cluster are also supposed to use Twitter more as a social network but with an emphasis on interacting with other users more than sharing self-contained ideas.

Generalists: This cluster groups users who use Twitter without a distinctive tweet type. It is interesting to notice that in this cluster, retweets with links and original tweets with links are slightly above the rest which may suggest an inclination to audiences interested in obtaining external information.

Echoers: These are users characterized by forwarding other people's tweets with no links. These users are mostly inclined to read what others have to say, indicating in a way that they make part of the audience of other users's original ideas (being these informative or not). An example of such users are those who follow accounts posting jokes, positive thinking, quotes, etc. The second dominant category in this cluster involves tweets with mentions, which most likely mean that users reply or chat with others.

Link Feeders: This cluster involves all those accounts that mostly tweet messages containing external links. In 2011 [143] found that around 50% of URLs posted in tweets came from media producers. We expect then that the owners of these accounts are mainly news media, journalists, link builders, SEO specialists, etc. Since these are tweets that contain no mentions, the expected target audience is then a general public that aims to obtain information through these accounts (*i.e.*, followers of news papers).

The clustering process was based on the tweets of active users in both 2011 and 2013. Figure 5.4 shows the number of users falling in one of the clusters for each year.

5.6 Change in Tweeting Behavior

Here we study how users have changed their tweeting behavior in 2013 with respect to 2011. Based on the active users only (those who remained active in 2011 and 2013), we plot these groups into a Sankey diagram in Figure 5.5 to observe the proportion of users moving from one cluster to another.

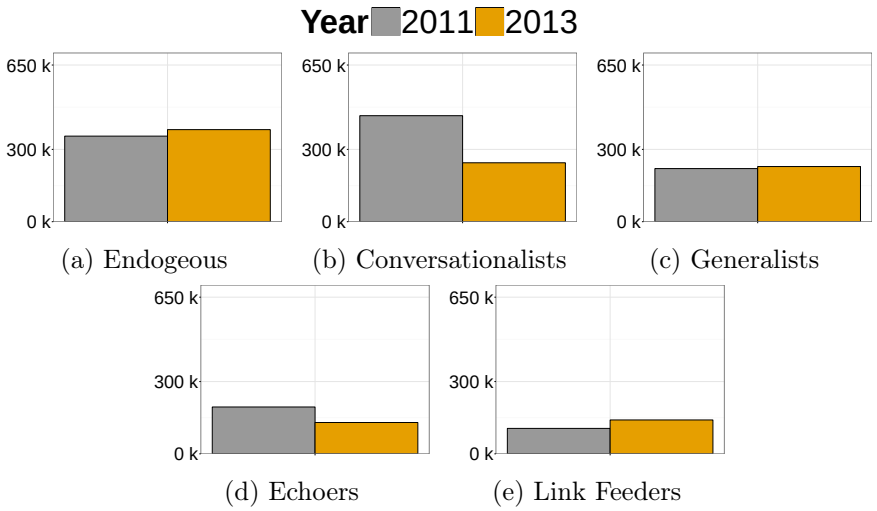


Figure 5.4: Number of active users in each cluster for 2011 and 2013.

We observe that in general around half of these active users remain in the same cluster in both periods, except for the Echoers. On the other hand, we observe an increase in 2013 of the *Generalists* and *Link Feeders* cluster with respect to 2011. The increase in the *Generalists* cluster is expected since our dataset contains users who have remained in Twitter for more than two years. These users have matured with the platform and most likely learned to use it for multiple reasons (chat, share information, retweets, etc). Moreover, the increase in the *Link Feeders* cluster goes along with Table 5.2, which also shows an increase in the percentage of tweets with URLs. Nowadays, Twitter automatically shortens URLs using the t.co service [22] which makes it easier for users to share links without the need to visit other URL shortener sites. This was not the case in 2011. Additionally, an increasing number of external sites allow to automatically post on Twitter with their link included. It is expected then that by 2013 users share more URLs than before.

On the other hand, we see a decrease in 2013 of the *Conversationalists* type. It seems that some users who used to chat a lot are evolving to chat less and be more *Endogenous* (posting their own tweets with no links or mentions) and *Generalists*. Mature users would have quickly realized that it was hard to continue conversations once the chat channel has passed in Twitter. On top of that, cross-platform instant messaging services more oriented to conversation purposes (*e.g.*,WhatsApp) have become increasingly popular. Neverthless, in 2013 Twitter made it easier to follow conversations in the

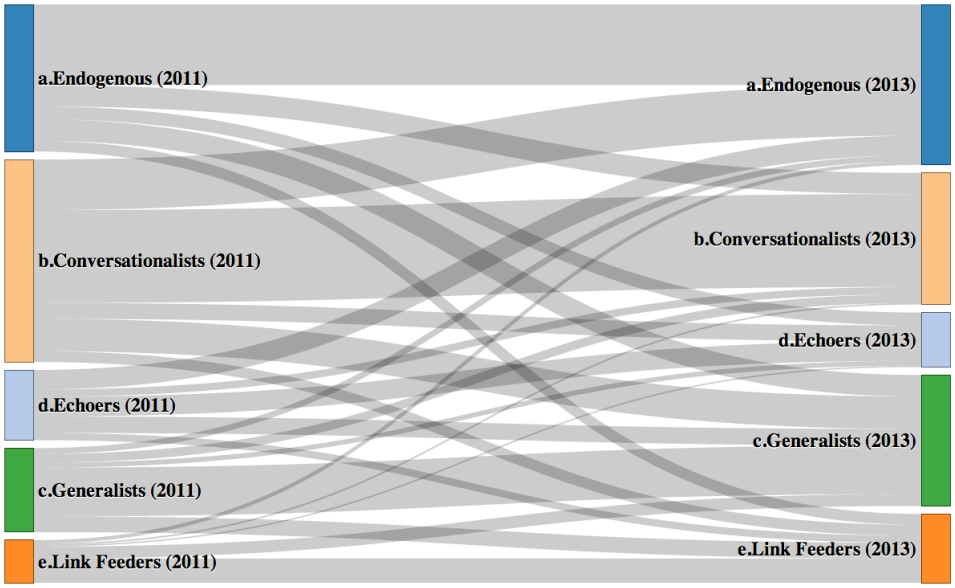


Figure 5.5: The Sankey diagram represents how users have changed the way they tweet in 2013 with regard to 2011. We observe that some users have stayed in the same cluster whereas others have moved to different ones. Inactive users are not considered.

timeline [70]. Perhaps, we will witness an increase in conversations after 2013.

Finally, the decrease in the *Echoers* cluster from 2011 to 2013 shows that users who tend to forward other people's ideas most of the time have evolved to generate more content themselves, moving to the *Endogenous* or *Generalist* clusters.

For a better readability of the evolution of active users' behavior, we did not include in the Sankey diagram the proportion of users who were filtered out of the active set in 2011 and moved to any of the clusters in 2013. We include this information in Table 5.3 in percentages (of around 4.3 M users) and show in Table 5.4 the corresponding absolute values. We observe that the majority of users from any cluster in 2011 become inactive in 2013. Similarly, inactive users tend to remain as such even two years later. Interestingly, the majority of hyperactive users move to one of the clusters but we also observe a significant percentage (26.71%) becoming inactive in 2013.

2011/2013	Endog.	Conver.	Gener.	Echoers	Link F.	Inactive	Hyper./Bots
Endog.	22.38%	5.89%	5.96%	3.56%	3.02%	58.33%	0.86%
Conver.	11.33%	20.79%	7.26%	3.54%	2.41%	53.80%	0.87%
Gener.	2.67%	3.88%	21.78%	2.17%	7.02%	62.07%	0.41%
Echoers	9.93%	3.72%	8.31%	9.93%	3.65%	63.62%	0.84%
Link F.	3.38%	1.47%	11.11%	1.25%	22.59%	59.45%	0.75%
Inactive	6.64%	3.30%	3.12%	2.38%	2.31%	82.00%	0.26%
Hyper./Bots	28.13%	17.42%	8.15%	6.48%	4.91%	26.71%	8.19%

Table 5.3: Percentage of users who changed clusters from 2011 (rows) to 2013 (columns). Some users passed from inactive or hyperactive/bot to other clusters and *vice versa*.

2011/2013	Endog.	Conver.	Gener.	Echoers	Link F.	Inactive	Hyper./Bots
Endog.	79,472	20,900	21,159	12,657	10,705	207,108	3,036
Conver.	49,832	91,429	31,945	15,570	10,616	236,624	3,807
Gener.	5,886	8,542	47,997	4,784	15,479	136,813	903
Echoers	19,308	7,235	16,149	19,306	7,105	123,704	1,640
Link F.	3,573	1,548	11,736	1,315	23,855	62,781	794
Inactive	194,636	96,641	91,391	69,684	67,769	2,403,596	7,481
Hyper./Bots	29,275	18,131	8,484	6,745	5,109	27,803	8,529

Table 5.4: The absolute number of users who moved across clusters from 2011 (rows) to 2013 (columns). Some users passed from inactive or hyperactive/bot to the other clusters and *vice versa*.

These findings go along with Liu *et al.* [86], who found a massive percentage of inactive accounts by the end of 2013. As Twitter users mature, many also choose to move to other platforms and to be less active.

5.7 Changes in Popularity

To gain more insights from our findings, we associate the popularity of users with their change of user behavior (clusters) between 2011 and 2013. But deciding how to measure popularity is not a straight forward task. Many researchers have proposed different ways of calculating a user popularity in Twitter, here we choose two of the most basic popularity metrics and associate them with the 5 clusters previously found.

Number of followers: It is the number of users following a twitterer and this information is provided by Twitter in each user's profile. From the tweets crawled in 2011 and 2013, we use the last reported number of followers for each user within the 10 weeks their tweets were crawled. The number of followers as a metric for reputation has been criticized by some [24] but other metrics have been shown to be worse (*i.e.*, the ratio followers/followees). Since we are interested to associate a form of popularity to our clusters rather than discussing which metric is the best, we choose the one provided by Twitter by default (followers).

Mentioners: It is the number of users who mentioned anyone in our dataset. The tweets include re-tweets, replies, conversations, etc. This metric is not provided by default. In order to obtain it, we crawled the tweets with a mention of any user from our set as explained in the *Dataset* section and counted the number of users posting these messages.

Since our dataset for both metrics does not follow a normal distribution and to avoid the influence of heavy outliers, we do not use the value of each measure, but the corresponding popularity rank. We then sort users according to their rank in 2011 and 2013 and use them to calculate the popularity gain across time.

First, we calculate the number of users who moved from one cluster to the other between 2011 and 2013 (already done in the previous section). Second, we compare the *ranking position* of users in 2011 with the one in 2013 for each pair of clusters and calculate the percentage of users who have improved their ranking. Table 5.5 shows the percentage of users who gained in popularity. The cells with a percentage of users higher than 50% are highlighted. From the results, it is interesting to observe that the majority of users moving from the *Inactive/bot* group to any of the 5 clusters, improve their ranking in *both* metrics. Similarly, the majority of users moving from any of the clusters to the *Inactive/bot* group decreases their ranking also for both metrics. It is then evident that tweeting too little or too much (*Inactive/bot*) affect users popularity *if* they stop being *active* and become *inactive/bot* and vice versa. We find an exception with respect to the *number of mentioners* in the transition of *Link Feeders* to *Inactive/bot* accounts. This makes sense to us because many news media accounts share a lot of information with their followers who in turn forward their tweets.

In these results, we do not include celebrities who tend to tweet little across time (passive in 2011 and 2013) and gain an increasing amount of followers

In number of followers

2011 / 2013	Endogenous	Conver.	Generalists	Echoers	Link F.	Inact./bots
Endog.	40.87%	**51.98**%	46.87%	**50.28**%	33.52%	24.23%
Conver.	43.84%	**52.66**%	33.42%	42.83%	43.13%	26.28%
Generalists	39.60%	48.06%	33.82%	48.28%	43.54%	25.33%
Echoers	42.17%	42.28%	29.16%	**50.96**%	41.78%	21.70%
Link Feeders	44.39%	**53.99**%	37.69%	**53.81**%	**53.71**%	26.29%
Inactive/bots	**59.81**%	**59.65**%	**59.76**%	**65.46**%	42.87%	NA

In number of mentioners

2011 / 2013	Endogenous	Conver.	Generalists	Echoers	Link F.	Inac./bots
Endogenous	36.33%	38.33%	30.50%	37.77%	35.30%	40.54%
Conver.	37.28%	36.60%	26.76%	25.43%	19.83%	38.72%
Generalists	32.86%	30.12%	30.23%	26.33%	21.26%	36.53%
Echoers	47.29%	38.19%	25.39%	40.44%	26.34%	29.36%
Link Feeders	**52.87**%	43.27%	**50.86**%	48.84%	35.34%	**56.98**%
Inactive/bots	**81.48**%	**78.85**%	**78.04**%	**83.19**%	**82.30**%	NA

Table 5.5: The matrix shows the percentage of users who gained in popularity (*i.e.*, followers or mentioners) after making a transition from a cluster in 2011(rows) to a cluster y in 2013 (column). We also included the transitions from/to the *Inactive/bot* group. The transitions higher than 50% are highlighted.

and mentions nor those who constantly tweet too much. These users remain in the *Icantive/bot* group in both 2011 and 2013. They are in the cell marked with a *NA*. We have decided to leave this cell blank since its analysis extends beyond the scope of this study.

5.8 Discussion

In this chapter we have carried out a study in Twitter between 2011 and 2013. We propose a taxonomy of 6 tweet types and found that users fall into 5 clusters of behavior: Endogenous (those who mostly tweet without links or mentions), Conversationalists (those who mostly converse with others), Generalists (those who post different type of tweets), Echoers (those who re-tweet more) and Link Feeders (those who share URLs most of the time). We then observed the evolution of users across clusters between these years and noticed a general tendency to become inactive or maintain the same type of behavior over years, with the exception of *echoers* who show to be active in a year full of controversial events. We also observed a decrease

of *conversationalists*, likely due to the maturation of users, the emergence of instant message services and the difficulty of chatting in Twitter before 2013. We also found more Link Feeders and Generalists in 2013. In the past, Twitter has been described as hybrid platform, being a social network and a news media at the same time [77]; our results, with the increase in news feeders and decrease in conversationalists, suggest that the main usage of the service by mature users is shifting towards the latter: a news media.

After completing this study, there are several complementary projects ahead. For instance, we plan to look closely at the behavior of the inactive and hyperactive users and bots. We also plan to study the lexical variation in dyadic conversations across time. Furthermore, it would be interesting to analyze if users tweeting in several languages differ in tweeting behavior for each language. Finally, we plan to compare this evolution to the change in user popularity.

Chapter 6

Cross-country Comparison of Microblogs Usage

6.1 Introduction

As seen in the previous sections, Twitter has become the most widely used microblogging service, and the messages people have posted on it have in many ways reflected real life events– from the revolutions in Tunisia and Egypt, to natural disasters such as the Chilean and Japanese earthquakes. Twitter users, however, post and share all kinds of information, ranging from personal opinions on important political issues, to mundane statements that may have little interest to most, except for their closest friends.

Given the range and scope of the service, and the fact that most user profiles and tweets are public, creates a huge opportunity to gain insights into not just how that particular service is used, but also into questions that are relevant in a social system at a particular point in time. This includes how news spread, how people communicate, and maybe how they influence each other, and many other aspects that roughly fall into the realm of Computational Social Science: a field that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors [80]. Interest in understanding such behaviors, however, goes well beyond the social sciences, as understanding those patterns can have applications in business and marketing decisions, and also in the development, and deployment of new products and services. These questions become particularly important when such services are provided on a world-wide scale and become so widespread that they may have important social impact.

Given this context, two key questions in the study of social media are how its use differs across cultures and countries, and whether any patterns revealed reflect behavioral differences and similarities between different groups. In spite of a long tradition and a lot of research in cultural anthropology [53], sociology [60], and other fields that address cultural differences, very little work has been carried out in the new Computational Social Science taking into account large data sets specifically examining differences across different countries.[1]

In this chapter we present the results of analyzing a large data set from Twitter in order to examine possible differences and similarities in several aspects of the use of the service. In particular, we focus on examining a year's worth of Twitter data for a large number of "active" users in the ten countries that tweet the most. We report on differences in terms of level of activity (number of tweets per user), languages used per country, the happiness levels of tweets, the content of tweets in terms of re-tweets, mentions, URLs, and the use of hashtags, and finally, we report on differences and similarities in terms of the network structure. While some of our findings are surprising (*e.g.*, levels of English used in some countries; differences in network structure), others confirm stereotypical perceptions (*e.g.*, that Brazilians might be happier than others).

The main contribution of our work is a series of insights on how tweeting behavior varies across countries (in terms of the factors described above), and on possible explanations for such differences. To the best of our knowledge, this is the largest study done to date on microblogging data, and the first one that specifically examines differences across different countries.

The rest of this chapter is organized as follows. In section 6.3 we describe our data set. In section 6.4 we describe the distribution of languages used in each of the ten most active countries in our data set, and the main findings of our analysis on the level of happiness in each country. Section 6.6 focuses on the content of the tweets and network structure, and we conclude by summarizing our main findings in sections 6.7.

This section is based on [41] and Poblete *et al.* [110].

[1] It is out of the scope of this chapter to provide an in-depth review, but here we refer specifically to analyzing differences in social media *across different countries*, as there has been of course a lot of work on social network analysis of large data sets.

6.2 Related Work

Researchers have focused on several aspects, including network structure, influence propagation, tweet topics, and several others. In this section we briefly mention work that we consider particularly relevant on comparing users from different countries.

Flickr pictures about the same thing but in different regions. Yanai *et al.* used state-of-the-art object recognition techniques to find representative geotagged photos related to a given concept and then studied how photos related to a concept change across countries [145]. They found, for example, that pictures of wedding cakes in US are much taller than those in Europe.

Travel destinations derived from Flickr posting. Based on where pictures are taken, Kling *et al.* derived travel patterns of a large number of Flickr users across countries [73]. They then used clustering methods to determine the extent to which any given pair of countries is related. They found that residents in Brazil and Chile have common travel destinations, for instance.

Color preferences in Instagram pictures. Hochman *et al.* extracted colors from pictures and found notable differences across pictures of different countries [58]. For instance, hues of pictures in New York are mostly blue-gray, while those in Tokyo are characterized by dominant red-yellow tones.

Download times of research publications. Wang *et al.* collected real-time data on which publication was downloaded at which time from the Springer Verlag website for 5 weekdays and 4 weekends [138]. Upon the resulting data set of 1,800,000 records, they found that downloads during weekends were the most common in Asian countries, and the least common in Germany.

Expression of emotions in Twitter status updates. Golder and Macy studied the 500 million English tweets that 2.4 million users produced during almost 2 years. Based on their hour-by-hour analysis, they found that offline patterns of mood variations also hold on Twitter: mood variations were associated with seasonal changes in day length, people changed their mood as the working day progressed, and they were happier during weekends [51].

Query logs. Baeza-Yates *et al.* studied the geographic locations of users who clicked on 759,153 Internet hosts. They found that users tended to click on hosts in other countries in which people speak the same language, and clicks coming from countries with similar human development index tend to end up into the same countries [7].

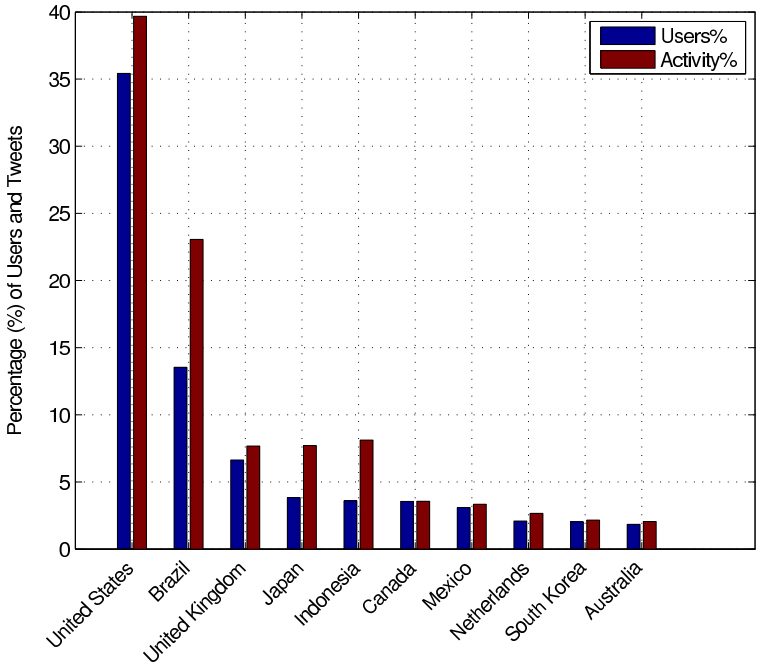


Figure 6.1: Distribution of users (%) in the dataset for each Top-10 country and their activity (%).

To summarize, in spite all of the recent work on Twitter, at the time of this research, we were not aware of any other study that focuses specifically on analyzing differences across different countries. We analyze the content of the tweets, the language used, and the happiness levels, but specifically compare how these differ in the ten most active countries examined. Similarly, we analyze network structure, again focusing on the differences between countries.

6.3 Data Set

As seen in Section 2.2, Twitter allows users to choose between keeping their profiles public or private. We limit our research only to information provided by users with public profiles in Twitter.

The focus of our research is mostly on characterizing large on-line social networks, based on user geographical location, for which Twitter provides limited information. Therefore, we perform an initial filter of users based on

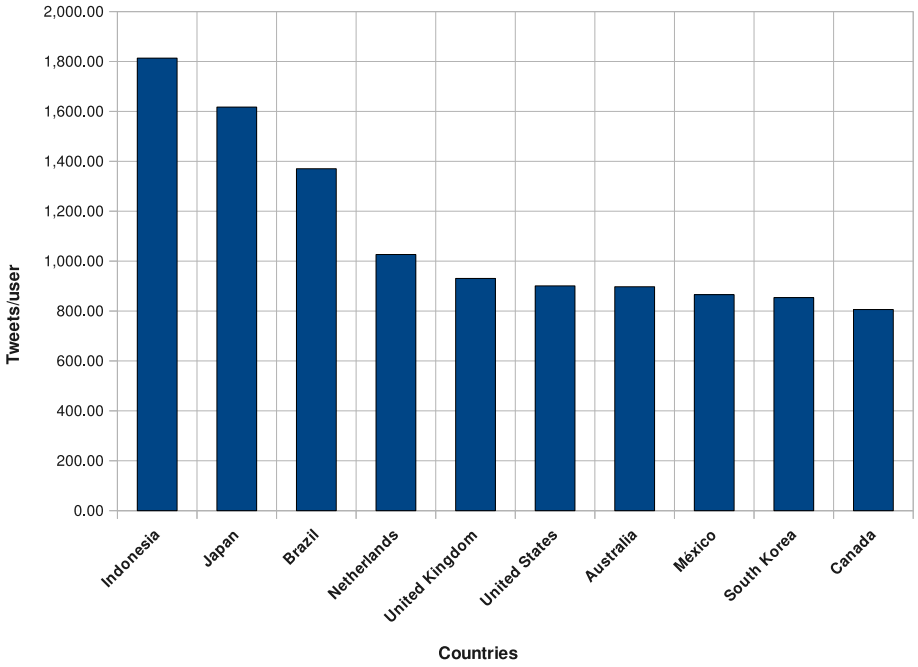


Figure 6.2: Tweet/user ratio for all of the top-10 countries.

activity and profile information. First, we choose users which we determine to be *active*. For this we examine a 10-day continuous time window of user activity, selecting day-1 randomly from the year 2010. Then, we consider *only* as active, users which generated tweets during this time frame. Secondly, we filtered the resulting users to keep only active users which had also entered a *valid location* into their profiles during this same time period as in Section 2.4. It should be noted, that we performed a more or less static analysis, so we did not consider user mobility during this period. We did not process location information which was automatically generated for the user with a GPS device on their client application, based on the fact that GPS location changes continuously with the user. Since in this work we are interested more in characterizing geographical communities of users, we decided to use the location which reflects more accurately the user's *home country*.

Using this criteria, we obtained a set of $6,263,457$ active users with valid location information, which were divided into 246 different countries. For the rest of our analysis, we selected the Top-10 countries with more activity

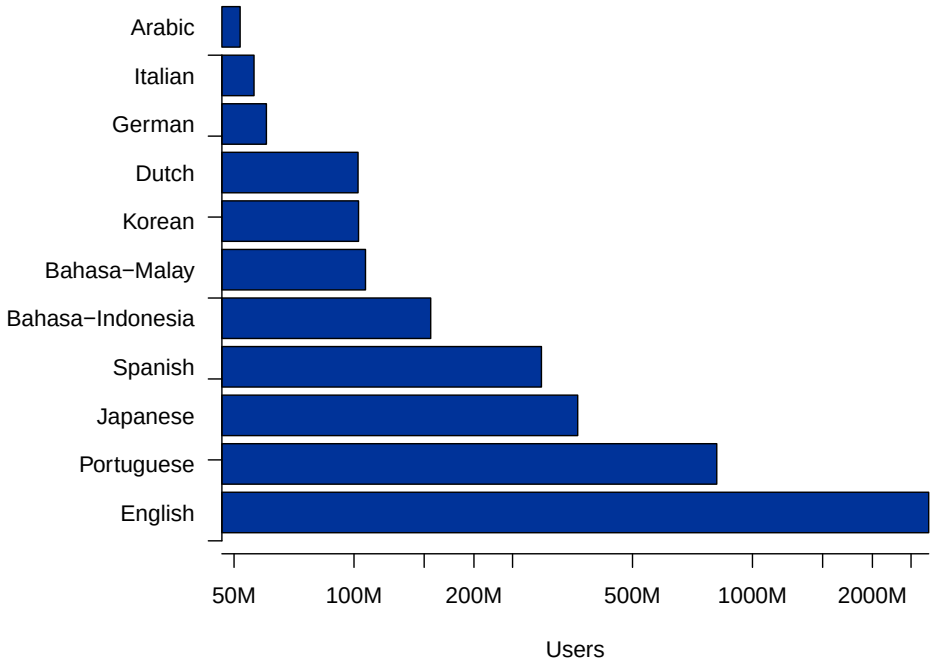


Figure 6.3: Most commonly used languages in all of the top-10 countries.

and gathered all of the tweets generated by these users for the entire duration of 2010. In total our working dataset consisted of 4,736,629 users (76% of the initial 10-day user sample), and 5,270,609,213 tweets. Figure 6.1 shows the distribution of the users in our dataset into the top-10 countries, and the activity that they generated for 2010. Note that the amount of activity registered for each country is not necessarily proportional to the number of users. This is explicitly shown in Figure 6.2, which displays the tweet/user ratio for each country. This ratio is independent of the number of users in each local network.

6.4 Languages

To analyze the language in which tweets are written, we cleaned them by removing URLs and non-alphanumeric characters. Then we used proprietary software to classify the language for each of the 5,270,609,213 tweets. As a result, 99.05% of the tweets were classified into 69 languages. The 10 most popular languages are shown in Figure 6.3. As expected, English is

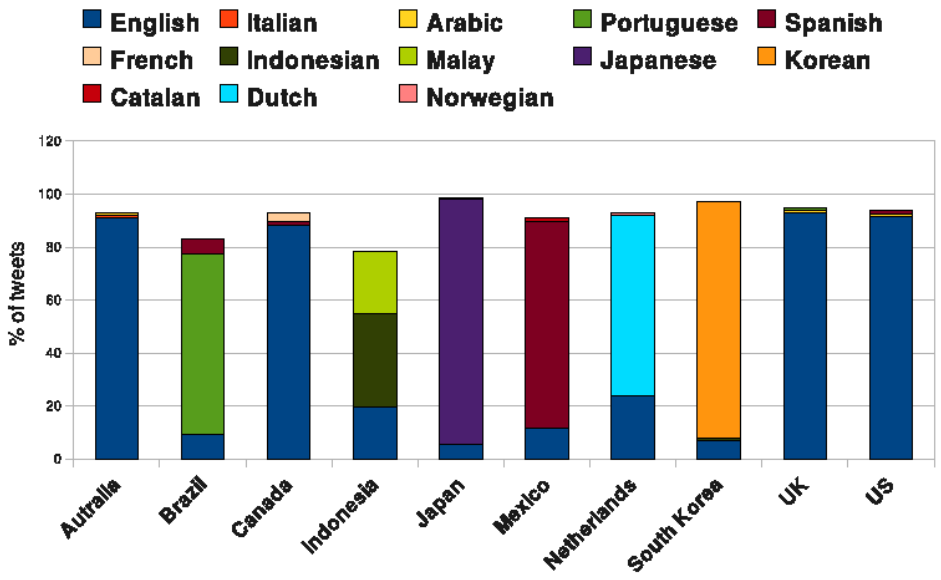


Figure 6.4: Three most popular languages for tweets in each of the top-10 countries.

the most popular language for Twitter updates, and it corresponds to nearly 53% of all the tweets. Additionally, Figure 6.4 shows the three most common languages for each of the top-10 countries, as well as the percentage of tweets which correspond to these languages. It is worth noting that English is one of the three most frequently used languages for these countries, and for the Netherlands, Indonesia, and Mexico more than 10% of tweets are in English, while for Brazil it is 9%. Additionally, Italian, Catalan, Arabic and Norwegian tweets (less than 3% of the total tweets of our data set) appear in Figure 6.4 in very small proportions. This is a bit suspicious given the fact that Italy is not considered in the top-10 countries of our study nor any Arabic country, as well as the number of people who speak Catalan and Norwegian world wide is relatively small. By sampling the tweets for Catalan language in Mexico (1.5%) and Italian in Australia (1.1%) we find that many of them correspond to false positives given by our classifier, since they actually correspond to Spanish and Portuguese. For example, the tweet "*Mexico no hay que llegar primero... si no que hay que saber llegar*" was labeled as Catalan although it is in Spanish and "*Um pequeno e valente*

guerreiro na luta contra o sono" was classified as Italian although it is in Portuguese (migration from Brazil has increased since the early 2000s in Australia[2]). Similarly, we observe a small proportion of Norwegian tweets in The Netherlands (0.98%) and found that some of these tweets were actually in Dutch. A possible explanation is that some Norwegian words are quite similar to the Dutch ones. The high resemblance of these languages, in addition to the common use of slang, along with misspellings, makes automatic language identification particularly challenging in these cases. In the case of the Arabic tweets in Australia (0.99%), UK (0.85%) and US (0.83%), we found tweets written indeed in Arabic. We explain this due to the high immigration of Arabic speaking people into these countries.

6.5 Sentiment Analysis

We also analyzed the sentiment component of tweets, for this we use the measure of *happiness* as coined by Dodds *et al.* [32], which is also more commonly referred to as *valence*. This value represents the psychological reaction which humans have to a specific word, according to a scale which ranges from "happy" to "unhappy". In particular, we analyze the happiness levels for each of the top-10 countries, considering only tweets classified as English and Spanish. To achieve this, we used the 1999 Affective Norms for English Words (ANEW) list by Bradley and Lang [18] for English tweets, and for Spanish, we used its adaptation by Redondo *et al.* [111]. The ANEW list contains 1,034 words and each word has a score in a 1 to 9 range, which indicates its level of happiness. The scores for the individual words were obtained by asking participants of a study, to rate them from: 9 which is "completely happy" to 1 which is "completely unhappy, annoyed. For example, the word "loved" has an average happiness value of 8.64 and its equivalent in Spanish ("amado") has a value of 7.99.

For each top-10 country, we computed the "weighted average happiness level", based on the algorithms of Dodds *et al.* [32; 31], as follows:

$$happiness(C_L) = \frac{\sum_{i=1}^{N_L} w_i f_{i,C_L}}{\sum_{i=1}^{N_L} f_{i,C_L}} = \sum_{i=1}^{N_L} w_i p_{i,C_L} \tag{6.1}$$

where $happiness(C_L)$ represents the weighted average happiness level for a country C, based on all of its tweets in language L (English or Spanish),

[2] https://www.dss.gov.au/sites/default/files/documents/02_2014/brazil.pdf.

during 2010. Therefore C_L represents all of the tweets registered for the country C which are expressed in the language L. Additionally, N_L represents the number of words in the ANEW list for the language l, while w_i is the score for the i-th word in the ANEW list for l, and f_{i,C_L} corresponds to the frequency of this word in the collection C_L. Finally, we denote p_{i,C_L} as the normalized frequency of each sentiment scored word in C_L.

The results of this sentiment analysis for a) English and b) Spanish, are shown in Figure 6.5. These results agree with those reported by Dodds *et al.* [32]: the values are between 5 and 7 for both languages and there is also a general increase in happiness towards the end of the year. It's interesting to note that Brazil has the highest values almost every month, even though we are not particularly considering Portuguese. Nevertheless, after August the happiness level in Brazil decreases until November. Also, in December all countries show an increase in their happiness level. Indonesia has high increase this month with scores that are even higher those of Brazil. South Korea also presents a strong increase this month, almost scoring the same as Brazil.

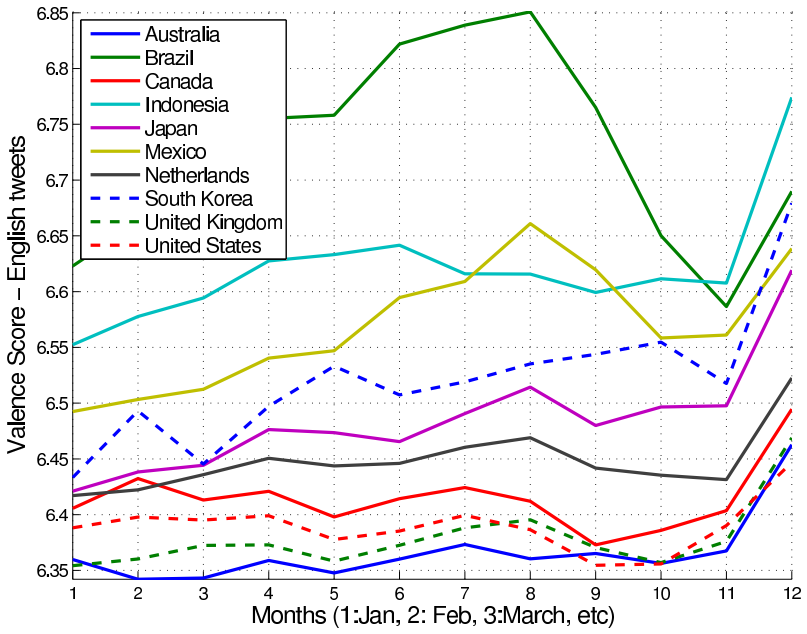
Some differences can be appreciated in the results for Spanish tweets, Figure 6.5.b. The number of tweets in Spanish is disproportional as 7 countries account for less than 1% of the tweets, while Mexico, USA and Brazil together account for almost 98% of the total. Nevertheless, USA and Mexico have happiness patterns that are similar to most countries. Only Brazil and Indonesia results which differ from the rest: there is a strong increase in happiness from June to July for Brazil and Indonesia. Interesting drops in levels happen in Indonesia during the months of May and August. Brazil has clearly the highest values for all months, but it also presents higher ups and downs.

6.6 Content and Network Structure

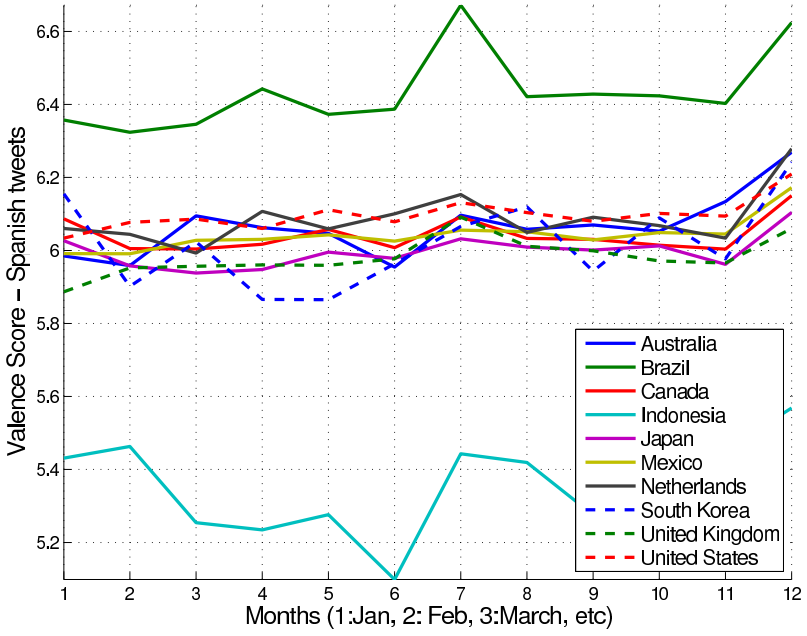
6.6.1 Twitter Conventions

In this part of our study, we analyzed briefly certain features of tweets provided by users for each top-10 country. These features have also been used in prior work, such as [21]:

- #: indicates whether a tweet contains a "#" symbol, used to give a tweet a particular topic.



(a)



(b)

Figure 6.5: Average *happiness* level per month for each country: (a) English and (b) Spanish.

Country	$\frac{Tweets}{Users}$	$(URL)\%$	$(\#)\%$	$(@)\%$	$(RT)\%$
Indonesia	1813.53	14.95	7.63	58.24	9.71
Japan	1617.35	16.30	6.81	39.14	5.65
Brazil	1370.27	19.23	13.41	45.57	12.80
Netherlands	1026.44	24.40	18.24	42.33	9.12
UK	930.58	27.11	13.03	45.61	11.65
US	900.79	32.64	14.32	40.03	11.78
Australia	897.41	31.37	14.89	43.27	11.73
Mexico	865.70	17.49	12.38	49.79	12.61
S. Korea	853.92	19.67	5.83	58.02	9.02
Canada	806.00	31.09	14.68	42.50	12.50

Table 6.1: Average usage of features per user for each country

- *RT*: indicates whether a tweet contains the keyword "RT" used to indicate a *re-tweet* or re-post of a message originally posted by another user.
- @: indicates whether a tweet contains an "@" symbol, used preceding a user name and which indicates a mention to that user.
- *URL*: Denotes whether a tweet contains a URL or not.

We computed the average per user for each country as follows:

$$AVG(symbol) = \frac{\sum_{i=1}^{N} \frac{T(symbol)_{u_i}}{T_{U_i}}}{\sum_{i=1}^{N} U_i} \tag{6.2}$$

Where $AVG(symbol)$ is the average number of tweets per user of a particular country containing a feature denoted by *symbol* (*e.g.*, #, RT, URL, @). Also, N is the total number of users for a particular country and $T(symbol)_{U_i}$ is the total number of tweets containing that feature for user U_i.

Table 6.1 shows the average per country as well as the ratio $\frac{tweets}{user}$. In our analysis, the appearance of the feature *symbol* in each tweet, was only counted once, that is, if a user used two hashtags in one tweet we counted it as one. The countries are ordered according to the ratio $\frac{Tweets}{User}$. Results show that Indonesia ranks first in tweets per user, followed by Japan and Brazil. It is interesting also to see that Indonesia and South Korea have the highest percentage of mentions in contrast to Japan that has the lowest,

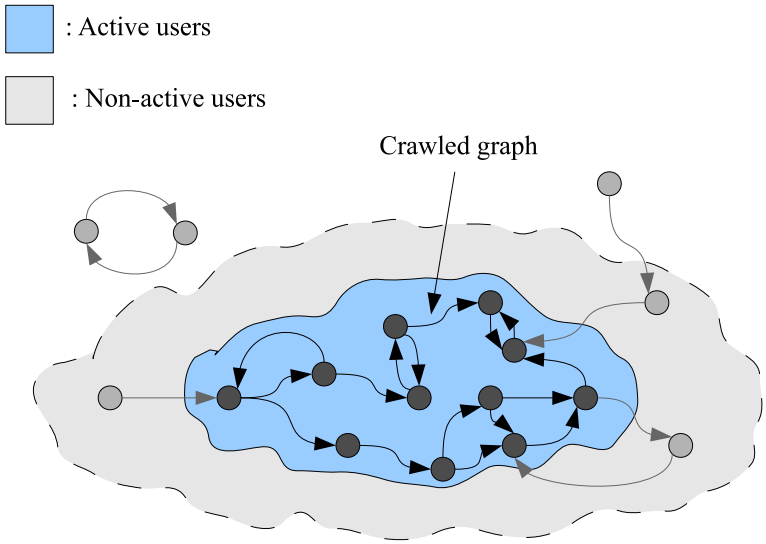


Figure 6.6: Collection strategy for the Twitter social graph, we considered only *active users* and the edges between them.

and it seems also to be the country with the fewest re-tweets in our data set. This indicates a higher use of Twitter for conversation than in other countries. The Netherlands is the country with the most hashtags per user, while the US seems to be the country with most mentions of URLs per user. At first glance, this could indicate that the US uses Twitter more than other countries for formal news dissemination, given that they are citing external sources.

6.6.2 Network

The Twitter microblogging service, provides also a social network structure for its users. This is, users connect to each other through directed links, therefore relationships are not necessarily reciprocal, as in Facebook.[3] Users can choose to *follow* other users, by subscribing to their updates. These connections between users can be viewed as a large directed graph.

In this section we focus on the analysis of the Twitter social network graph for each top-10 country and its active users (as defined in Section 6.3). In order to obtain this graph, we extracted user relationships using the public

[3] http://www.facebook.com.

Country	Users	Cov.(%)	Links	Cov.(%)	Reciprocity(%)
USA	1,616,702	12.47	11,310,538	12.46	18.91
Brazil	688,427	5.31	4,248,259	4.68	13.49
UK	286,520	2.21	1,370,699	1.51	17.22
Japan	133,536	1.03	408,486	0.45	**32.01**
Canada	132,240	1.02	553,726	0.61	**26.11**
Indonesia	130,943	1.01	199,704	0.22	**26.97**
Mexico	112,793	0.87	399,409	0.44	17.27
Netherlands	86,863	0.67	354,021	0.39	22.11
South Korea	80,381	0.62	499,261	0.55	**28.14**
Australia	67,416	0.52	299,556	0.33	23.51

Table 6.2: General summary of network statistics per country (high reciprocity values in bold).

Twitter API (4J), collecting the list of followers/followees for each user. In this particular graph, connections between users are highly dynamic, so we worked with a snapshot of the graph, which was crawled between November 25 to December 2, 2010. This crawl resulted in $12,964,735$ users and $90,774,786$ edges. We cleaned this data set to keep only edges and users which corresponded to our *active user set*. Figure 6.6 illustrates the crawling strategy, the inner cloud represents the social network considered in our study, and the outer layer represents the set of discarded users and edges. Prior work [95] has shown that analysis of partial crawls of social networks can underestimate measures like degree distribution, but continue to preserve accuracy for other metrics, such as density, reciprocity and connectivity. We believe that by preserving the active component of the graph we are analyzing the most relevant part of the social structure.

Table 6.2 shows a summary of each countries' network graph statistics. In particular, for each local network analysis, we considered only connections between users in the same country. The second and third columns in Table 6.2 show the node and edge coverage of each country with respect to the entire graph. We also show the percent of reciprocity, which is the fraction of ties between users which are symmetric (*i.e.*, "a" follows "b" and "b" follows "a"). Overall, the top-10 most active countries cover 25.73% of the total of active users in the social graph. Additionally, these countries cover the 21.64% of the total number of edges in the global network. Table 6.2 also shows that for some countries reciprocity is very significant, in comparison to others. These values are in particular high for Japan, South Korea, In-

Country	Avg. δ	Density	Avg. Clus. Coef.	Strongly CC
USA	8.95	0.56 E-04	0.0645	9,667
Brazil	7.55	1.09 E-04	0.0711	4,813
Indonesia	2.12	1.62 E-04	0.0618	7,942
United Kingdom	6.05	2.11 E-04	0.0933	14,818
Japan	4.36	3.26 E-04	0.0603	6,052
Mexico	4.44	3.91 E-04	0.0826	6,885
Canada	5.73	4.33 E-04	0.1001	6,630
South Korea	8.61	10.67 E-04	0.0879	3,864
Netherlands	5.39	6.16 E-04	0.1017	4,626
Australia	5.83	8.52 E-04	0.0959	3,423

Table 6.3: Summary of network density statistics per country.

donesia and Canada. The symmetric nature of social ties affects the network structure, increasing connectivity and reducing the diameter, as we will see in the remaining of this work.

Table 6.3 shows a summary of graph density statistics, such as average degree (δ), density and average clustering coefficient. USA and South Korea are the countries with the highest averaged degree per node, meaning that users tend to concentrate more followers and followees than in other countries. Indonesia, on the other hand, presents a very low degree (only 2.12 edges per node on average) in spite of being a very active community. The second column in Table 6.3 shows each local network's density values. Density is computed as $\frac{m}{n(n-1)}$, where n is the number of nodes and m is the number of edges. The density is 0 for a graph without edges and 1 for a fully connected graph. In our study, South Korea displays the highest density of all countries. Additionally, density increases as the network becomes smaller, *e.g.*, USA has the lowest density, and the three highest values correspond to South Korea, Netherlands and Australia. Therefore, smaller communities are more well connected to each other globally, within their own country.

The third column in Table 6.3 shows the average clustering coefficient. We compute the clustering coefficient for each node counting the number of triples (non-directed triangles) in the graph which include the target node. Then, we compute the clustering coefficient, as the fraction over the total number of possible triples that exist. Values in Table 6.3 represent average values of the clustering coefficient, computed for all users in each country. We can observe that communities with high clustering coefficient and less

Country	Modularity	Number of communities
USA	0.418	2,954
Brazil	0.462	2,896
Indonesia	0.537	3,358
United Kingdom	0.397	2,486
Japan	0.458	1,998
Mexico	0.358	1,406
Canada	0.568	1,269
South Korea	0.312	756
Netherlands	0.412	936
Australia	0.452	634

Table 6.4: Summary of graph modularity statistics per country.

reciprocity may indicate more hierarchical-type relationships between users (*i.e.* two users who share a reciprocal tie follow a same third user who does not reciprocate). The fourth column of Table 6.3 shows the number of strongly connected components that exist in each country.

In table Table 6.4 values obtained when measuring the *modularity* of each social network graph. We obtain these values by computing the degree of separation between each node to very other node its network. We use the modularity coefficient as defined by Girvan and Newman [49], which evaluates how well a graph can be partitioned. A value of 0.4 or greater is generally considered meaningful. In our analysis we can appreciate that Indonesia and Canada display high modularity, which indicates that the communities found in these countries are more compact and closed than in other countries. On the other hand, Mexico, South Korea and United Kingdom indicate less separation between their communities.

In Table 6.5 we summarize some general network distance measures per country. We consider network diameter, *i.e.* the maximal distance between all pairs of nodes. Also, we compute the average path length of each network, and the number of shortest paths. Table 6.5 shows that Indonesia presents the highest diameter, indicating that this network is very partitioned, which agrees with its high modularity coefficient. Several countries register diameter values in the range of 16-18. The lowest diameter is found for the South Korea network. We can also see that average path lengths are proportional to diameter values. Also, in general, the number of shortest

Country	Diameter	Avg. Path length	Shortest paths
Indonesia	35	9.69	33,940,227
USA	18	6.49	5,746,903,535
Brazil	16	6.37	2,147,483,647
Mexico	16	5.27	50,512,898
Japan	18	5.26	86,348,633
United Kingdom	15	5.19	402,698,573
Netherlands	17	4.81	33,628,133
Canada	16	4.71	77,645,673
Australia	14	4.52	22,271,542
South Korea	11	4.02	33,517,802

Table 6.5: Summary of graph distance measures per country.

paths is proportional to the number of edges in the graph. Notice that, for example, the three graphs with the highest edge coverage values (USA, Brazil and United Kingdom) are also the three countries with the highest shortest path values (see Table 6.2). Additionally, Figure 6.8 shows a visual comparison of the networks of a) Indonesia and b) Australia, in which we can appreciate the differences in diameter and density. Indonesia has lower density and a larger core component, which increases path lengths between nodes.

We analyze the existence of a direct relationship between average path length and diameter with reciprocity. Intuitively, we would expect that shorter paths and diameters would result from networks with high reciprocity. Nevertheless, we do not observe any apparent relationship, as shown in Figure 6.7. In Figure 6.7 we display average path length and diameter ordered by increasing reciprocity. On the contrary, several countries show significant reciprocity and at the same time large diameters. The most noticeable case is Indonesia, which shows the largest diameter and also high reciprocity. This suggests that the graph structure strongly influences the relationship between reciprocity and diameter. Given our previous observation, which was that Indonesia had high modularity (see Table 6.4), this corroborates more the idea that this country has very compact and isolated communities of users. On the other hand, Canada also shows a very significant modularity value but its diameter and average path length values are very similar to countries that do not show a community structure. The main difference

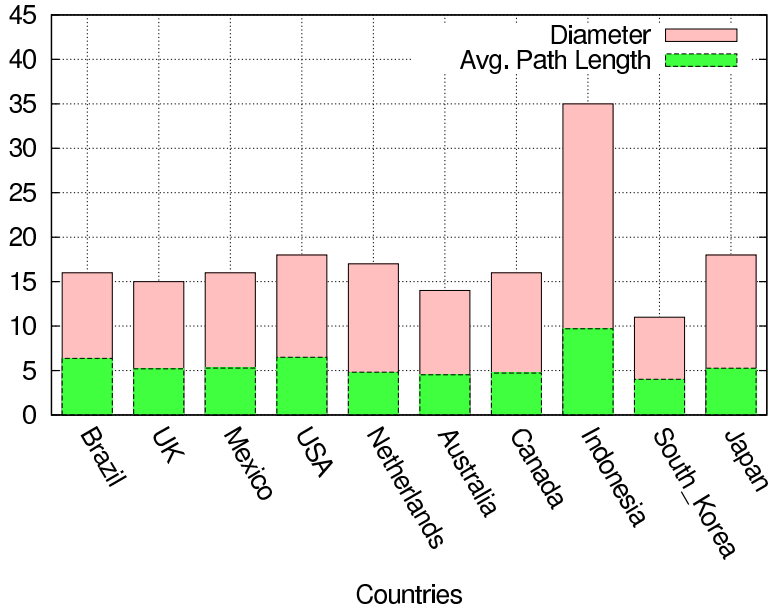
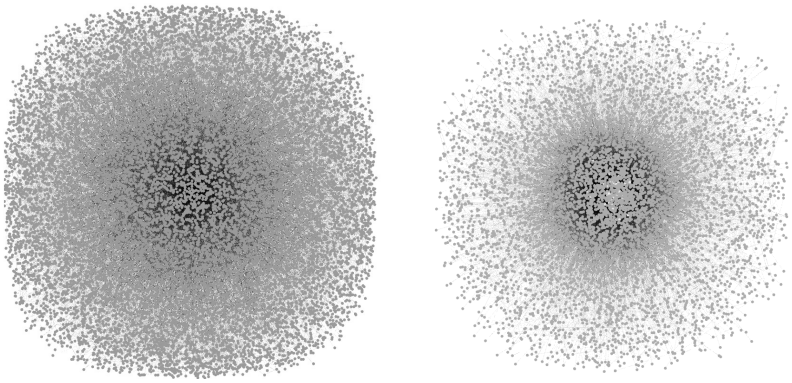


Figure 6.7: Average path length and diameter per country, sorted by reciprocity in increasing order from left to right.

between Indonesia and Canada from our observation is that the first has a much lower clustering coefficient and density than the second. This might indicate that Indonesia has more users than Canada which do not participate in large communities.

We also examine the graph structure of each network by considering node degree distribution. Degree distributions of many social networks have been shown power laws behaviors. This kind of networks are networks where the probability that a node has degree k is proportional to k^{γ}, where γ is known as the power law coefficient. Figure 6.9 shows the out-degree and in-degree cumulative distribution function for the entire graph.

As Figure 6.9 shows, both distributions display a power law behavior. Additionally, power law distributions are observed for every country. This can be interpreted as the power law behavior of users being independent of geographical differences. Overall, users have few followers / followees ties, and only few users register significant number of followers and followees. In Table 6.6 we show the power law coefficients for each country and their assortativity values. Out-degree coefficient values are greater than in-degree



(a) (b)

Figure 6.8: Social networks for Twitter communities in a) Indonesia, and b) Australia.

Country	In-degree Power Law	Out-degree Power Law	Assortativity
USA	9.51	13.62	-0.19
Brazil	7.56	12.42	-0.17
Indonesia	6.21	9.48	-0.06
United Kingdom	7.31	10.89	-0.18
Japan	7.48	9.68	-0.07
Mexico	5.91	9.26	-0.21
Canada	8.31	9.30	-0.11
South Korea	6.36	8.21	-0.27
Netherlands	6.67	8.64	-0.17
Australia	7.72	8.12	-0.11

Table 6.6: Summary of graph degrees and assortativity statistics per country.

values, indicating that in general it is more likely to be followed by a user than to follow a user.

We also study structural properties of each sub-graph by exploring the joint degree distribution (JDD). We approximate the JDD by analyzing the average neighbor connectivity for every node with a given degree, coefficient known as k_{nn} and formally defined as follows:

$$k_{nn} = \sum_{k'} k' P(k' \mid k),$$

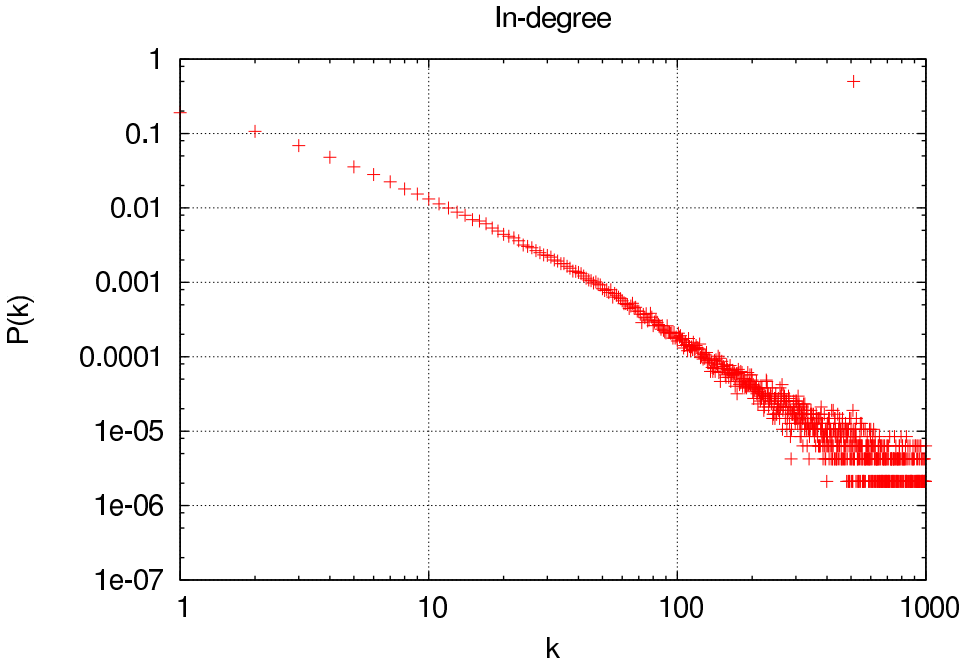
where $P(k' \mid k)$ represents the conditional probability that an arc of node degree equals to k points to a node with degree k'. Intuitively, if this function is increasing, high degree nodes tend to connect high degree nodes. Alternatively, a decreasing function characterizes a disassortative network, in which high degree nodes are connected to low degree nodes. We calculate the Pearson correlation coefficient of degree between pairs of nodes, also known as the assortativity coefficient. Positive assortativity values indicate relations between nodes with similar degrees and negative assortativity values indicates relationships between nodes with different degrees. Figure 6.10 shows the k_{nn} coefficient per node degree per country. Each plot also shows its assortativity coefficient, listed in Table 6.6. We can observe that Twitter networks appear to be disassortative, which indicates that low degree nodes tend to connect to high degree nodes, illustrating a preferential attachment behavior.

Besides looking at properties of each country in isolation from the rest, we have also analyzed the number of in-links and out-links from one community to another. These results are presented in Figure 6.11. In this ilustration it is interesting to observe that all countries direct the majority of their external out-links to the US. Nevertheless, several countries concentrate their most significan amount of links towards themselves, with the exceptions of Canada, Australia and UK, which connect to the US almost as much as to themselves.

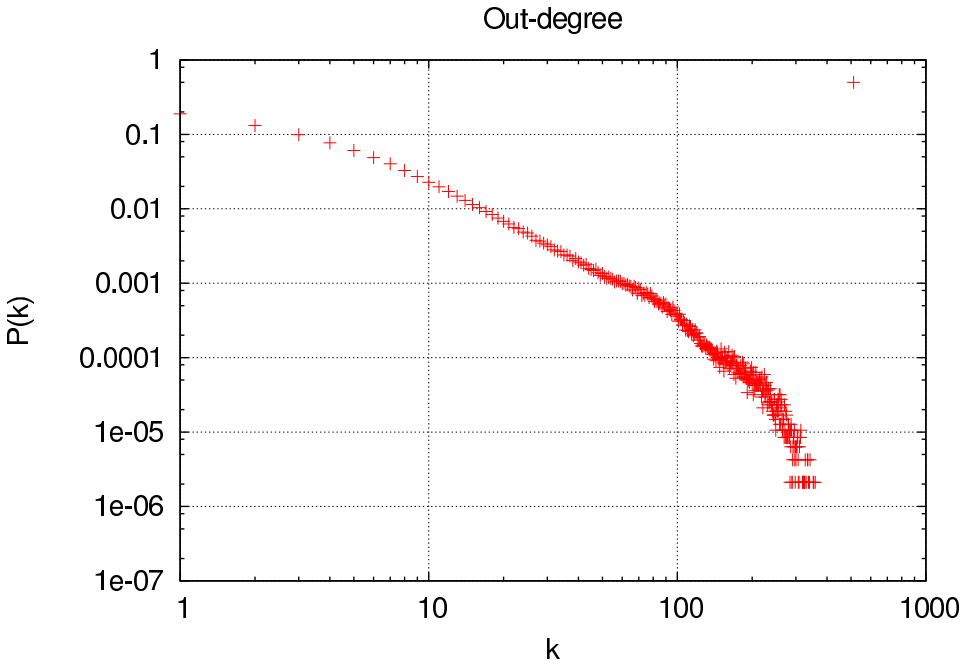
6.7 Discussion

We have presented a broad study of the Twitter on-line social network. We have segmented our analysis into the top-10 countries with most activity and collected data from a representative sample of users for one year. We analyze several aspects, such as language, sentiment, content and network properties.

In particular some countries standout, based on their peculiar characteristics such as the level of reciprocity in the network. Network reciprocity tells us about the degree of cohesion, trust and social capital in sociology [56]. In this context, the equilibrium tendency in some human societies is to have reciprocal connections. It is said that asymmetric ties are unstable. Therefore,



(a)



(b)

Figure 6.9: Degree distributions for the full data set: in-degree (top), and out-degree distribution.

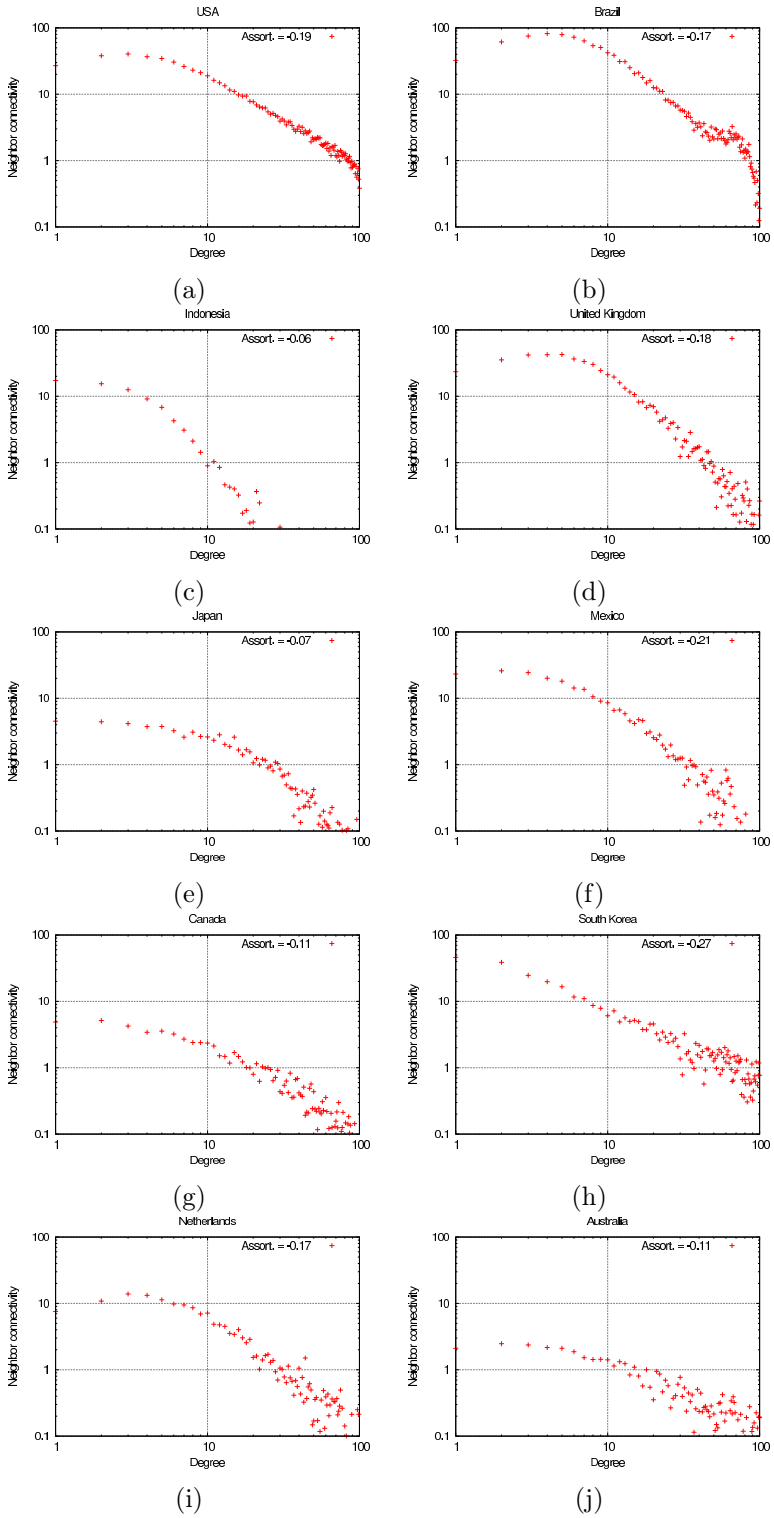


Figure 6.10: Log-log plots of neighbor connectivity versus node degree. Assortativity values are also shown, suggesting the presence of disassortative networks.

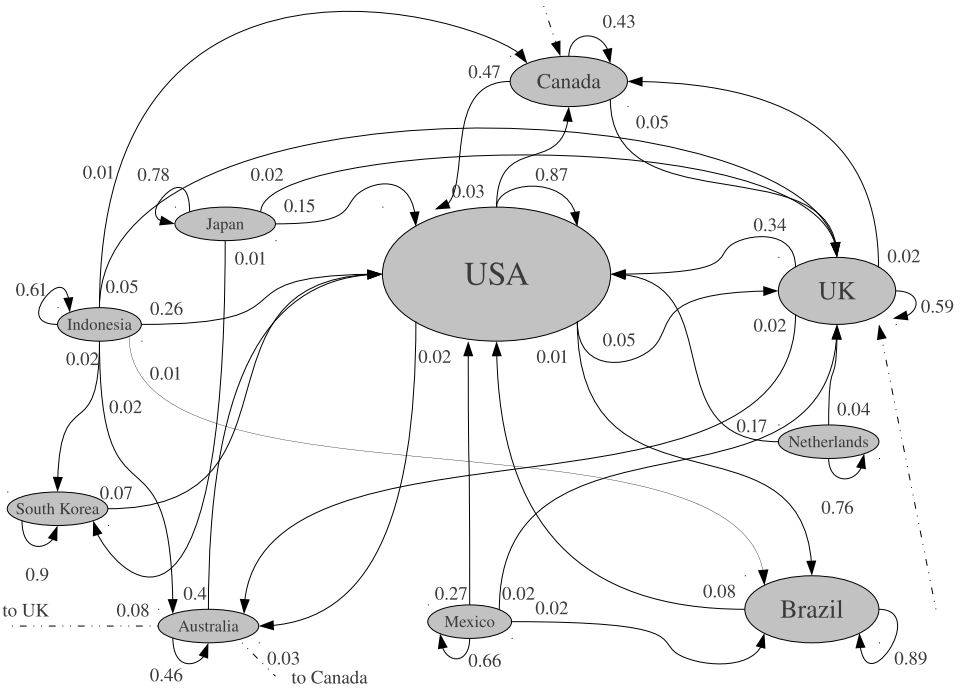


Figure 6.11: Ties between countries, sizes are proportional to the size of the community. Weights represent fraction of ties calculated over the total number of ties of a given country.

reciprocal is more stable or equal. Nevertheless, Twitter networks seem to work towards an equilibrium which is not reciprocal and more hierarchical. Therefore, from our perspective, it tends more to follow a model in which we have authorities which receive many ties but do not reciprocate. Nevertheless, we detected that some smaller networks display high reciprocity and more conversation. This indicates a high conversational use within compact communities, as opposed to broad dissemination of news information.

Furthermore, we observe that countries which have high reciprocity tend to have a higher tweets/user ratio. Additionally, smaller networks also have a tendency to have higher reciprocity. Indicating more local communities. Additionally, we see that communities which tend to be less hierarchical and more reciprocal, also displays *happier* language in their content updates. For example, Brazil has higher scores, which overall indicate a more positive language use. Interestingly, Brazil also shows important fluctuations, which

could indicate stronger polarity in in expressiveness. In this sense other countries with high conversation levels (@), shown in Table 6.1, such as Indonesia and Mexico, show higher levels of *happiness* too. This is reasonable, if we think that higher conversational levels can imply that users privilege more informal communication, as opposed to formal news dissemination. Following this reasoning, we can hypothesize that these users, use Twitter more as a conversation channel than formal information source. Therefore, their interaction is highly conversational with friends, as opposed to being more formal. which is the case of the US.

High reciprocity does not imply that these countries are more well connected overall, in cases such as Indonesia, reciprocity is contained within very compact and closed groups. This increases the overall diameter of the network, as seen in Table 6.4 and Figure 6.7. In particular Figure 6.8 illustrates this showing that Indonesia has a more disperse core than Australia, which makes its diameter larger.

By analyzing Table 6.3 in Section 6.6.2, we can observe that smaller communities, tend to present more reciprocal ties between users in general, along with high density and average clustering coefficient (not always, but this is the tendency). Pointing towards more reciprocal relationships. Additionally, high reciprocity creates high activity in communities such as Indonesia and Japan.

On the other hand, we can see in Table 6.3 that countries with high density and high clustering coefficient, such as South Korea, Netherlands and Australia, contain users which participate in small and compact communities. Other countries, such as USA, have low clustering coefficient and density, which indicates that many users do not participate in a small compact community. Additionally, high clustering coefficient, but low reciprocity may also be an indicator of more hierarchical communities. In the sense that nodes are quite connected in communities, but not reciprocal, therefore most of the in-degree connection are towards a few nodes, which do not reciprocate to their followers.

In the next chapters, we explain the differences in user behavior across countries by overlapping our results with previous anthropological studies of culture as well as economic and sociological indicators.

Part III

Culture in Microblogs

In this part we present two studies based on the Cultural Models explained in Sections 3.2 and 3.3. First we explore how culture influences the way we use social media by considering *time, individualism* and *power*. Second, we analyze how culture and socio-economic indicators influence the way we *communicate* in microblogs.

Chapter 7

Time, Individualism and Power

7.1 Introduction

Researchers have found that the ways people perceive and accept power differences, interact with each other, and perceive time, drastically differ across countries. For example, in certain countries (*e.g.*, Japan), direct disagreement is synonym of confrontation, while speaking one's mind is a virtue in others (*e.g.*, USA). Also, cultures with a fast Pace of Life (*e.g.*, Germany, Switzerland) tend to give more importance to punctuality and have less flexible schedules; by contrast, cultures with a slower Pace of Life (*e.g.*, Brazil) are more flexible and give importance more to human interactions than to keeping the schedule [84].

Cultural variations across countries have been empirically studied using small-scale experiments and surveys in the real world. As we saw in Chapter 3 Geert Hofstede administered opinion surveys to IBM employees in over 70 countries [61]. This data, with over 100,000 questionnaires, were one of the largest cross-national databases that existed in 1971. By analyzing it, Hofstede discovered that there were significant differences between cultures: he found that five main factors explained most of the variance in the data and called those factors cultural dimensions, and two of those have been widely studied. The first is *Power Distance* and reflects the extent to which people (especially those less powerful) expect and accept that the power is distributed unequally (*e.g.*, employees would rarely contradict their managers). The second dimension is called *Individualism vs. Collectivism*

and reflects the extent to which social relationships are loose (*e.g.*, people look after themselves and are likely to have friends outside their immediate families) as opposed to relationships integrated in strong and cohesive groups (*e.g.*, friends are likely to be within families).

Similarly, we saw in Chapter 3 the concept of *Pace of Life*. Robert Levine run different experiments to capture Pace of Life in a variety of countries. In 31 countries, he and his students measured the time it takes for people to walk 100 meters in coffee shops, for post clerks to send a parcel, and they also kept track of the accuracy of clocks in public spaces (*e.g.*, in post offices). That resulted into ranking those countries by what they then called Pace of Life [84].

Individualism, Power Distance, and Pace of Life have been found to determine how people behave differently in the same situations in the real world. The main goal of this work is to assess the extent to which such differences can also be captured from online interactions. We will see that these differences matter because they are associated with the economic aspects of Gross Domestic Product (GDP) per capita, income inequality and education expenditure.

To go beyond small-scale experiments and surveys, we consider Twitter, a microblog massively used worldwide, and set out to answer the following research question: Does national culture determine the temporal randomness with which Twitter users post, or the extent to which they mention, follow, recommend and befriend others? We crawl more than 2.34 million user profiles (starting from 55K seed users), their tweets during 10 weeks from March to May 2011, their geographic locations, and corresponding time stamps (Section "Data Description"). Upon this data covering the 30 most represented countries in our sample, we test three main hypotheses associated with the three cultural aspects and, in so doing, we make four main contributions:

- We test whether the higher a country's Pace of Life, the more predictable its citizens' temporal patterns (Section 7.4). The link between Pace of Life and temporal predictability comes from the finding that countries with higher pace of life tend to schedule their time in more predictable ways [84], [142]. To test this on Twitter, in our period of ten weeks, we divide each working day into 5 segments and compute the extent to which each user tweets or mentions others in the same daily segments. We aggregate all users in each of the 30 coun-

tries, produce a country-level temporal predictability, and correlate it with the country's pace of life. The correlations are $r = -0.62$ for tweets' temporal unpredictability, $r = -0.68$ for user mentions', and $r = -0.58$ for tweeting activity within working hours. These consistent results confirm that countries with higher pace of life tend to be more predictable not only offline but also online.

- We also test whether people in collectivist countries interact more with each other than those in individualistic countries (Section 7.5). We do so by computing the percentage of users who mention each other. We find that the correlation between Individualism index (one of Hofstede's cultural dimensions) and the extent to which users mention each other is as high as $r = -0.55$.
- We test whether users in countries comfortable with unequal distribution of power (high power-distance countries) will follow, recommend, and accept recommendations preferentially from users who are more popular (Section 7.6). To this end, we consider three types of relationships: a) who follows whom; b) who recommends whom; and c) who starts to follow whom upon a recommendation. For each relationship, we compute the difference of followers between the pair of users in the relationship, and call that power imbalance. We then correlate country-level imbalance with corresponding Power Distance (another one of Hofstede's cultural dimensions). We find that the correlations are $r = 0.62$ for "who follows whom" relationships; $r = 0.33$ for "who recommends whom" relationships, and $r = 0.42$ for "who starts to follow whom" relationships.
- We finally show that those three cultural dimensions are associated with the three economic indicators of GDP per capita, income inequality and education expenditure. We find correlations as strong as $r = 0.60$.

These strong correlations suggest that cultural differences are not only visible in the real world but also emerge in the way people use social media. To show why these cultural dimensions matter, we will study their relationships with socio-economic indicators, including GDP per capita. We conclude by discussing the theoretical and practical implications of this work (Section "Discussion").

There has not been any work on how cross-country variations of language independent features (*e.g.*, predictability, mentions and subscription activity)

in a general-purpose platform (*e.g.*, in Twitter) are associated with indicators well-established in anthropological studies (*e.g.*, cultural dimensions, pace of life). That is why we run such a study next. This section is based on [43].

7.2 Related Work

Our goal is to study variations of Twitter use across countries. In a similar way, researchers have already analyzed how a variety of aspects of the online world change across countries. Some of them were already presented in Chapter 6 and other studies, more specific to this chapter, on country-variability cover the following aspects:

Scheduling. Reinecke *et al.* studied how the use of the web scheduling tool varies across 211 countries [112]. They did so by relating activity features (*e.g.*, consensus, availability) to Hofstede's Colllectivism vs. Individualism dimension, and with Inglehart Survival and self expression values. They found that users of the tool in Germany tended to schedule far ahead of time (around 28 days in advance, while those in Colombia schedule up to 12 days in advance).

Applications. Oh *et al.* [104; 103] also presented a model that integrates a *Cultural Dimensional Model* and empirical data based on phone applications. First, they collected information about mobile-phone applications downloaded in several countries and classified each application according to its content (Finance, Entertainment, etc). Second, they defined a group of dimensions inspired in Hofstede as a basic framework for analysis. Then they used a Delphi Survey with 5-point scales to assign values to each application according to the framework established by them. Next, for each phone application they assigned a *Cultural Dimensions Score of Content* based on the Delphi survey. Finally, they used this score to calculate a *Cultural Index Score for Country* [104] which represents the country's inclination to a given cultural dimension. Several techniques are used to present similarities and differences of these indexes at the national level. Probably due to limited access to free downloaded phone applications, their results showed different characteristics from what Hofstede's.

7.3 Data Set

From the Twitter stream API, we randomly selected 55K users who tweeted at least once in March 2011 and that had an outdegree and indegree in the range $[100, 1K]$. This choice is imposed by API restrictions but has the side benefit of filtering away less legitimate (*e.g.*, spam) users: the majority of spam users tend to have outdegree and indegree outside the range $[100, 1K]$ [81]. We select users with a geo-location (in Section 2.4 we explained how to find them) resulting into 12.6K seed users that are geo-located. For these seed users, we collected their outbound links (followees) and found that 1.96M of them had location information. Since one of our hypotheses require the study of recommendations (which are made using the follow friday hashtag *#ff* or *#followfriday* in Twitter), we also collected the recommendations (*i.e.*, users to follow) that were made by the followees (outbound links) during the subsequent 10 weeks. This resulted in 362K recommended users with valid geolocations. Overall, we will study 2.34M users (12.6K+1.96M+362K), their timestamped tweets (considering all different timezones), and their locations.

The way we sample our users is convenient and easy to interpret but might be biased by our particular choice of seeds. To partly address this concern, we only considered the top-30 countries in our sample. We choose 30 because it is the highest number of countries in which presence on Twitter highly correlates with presence on the Internet (Figure 7.1), and in which the number of per country users is always more than 5K, ensuring statistically significance of our results. Figure 7.2 plots the number of users in our sample as a function of the number of Internet users. Most of the countries follow a straight line. USA deviates considerably from it simply because of its high Twitter penetration rate.

Next, we consider those users and their countries and study their specific cultural aspects in sections 7.4, 7.5 and 7.6.

7.4 Pace of Life

Having this data at hand, we can now start with the first dimension of our analysis: Pace of Life. This differs across countries: for example, Levine found that USA's Pace of Life is higher than Brazil's [84]. One could order countries by the value residents give to time and would see that Sioux Indians do not have a notion of time (they even do not have a word for it); Brazilians have a 'relaxed' notion of it (*e.g.*, Levine fount that students defined 'being

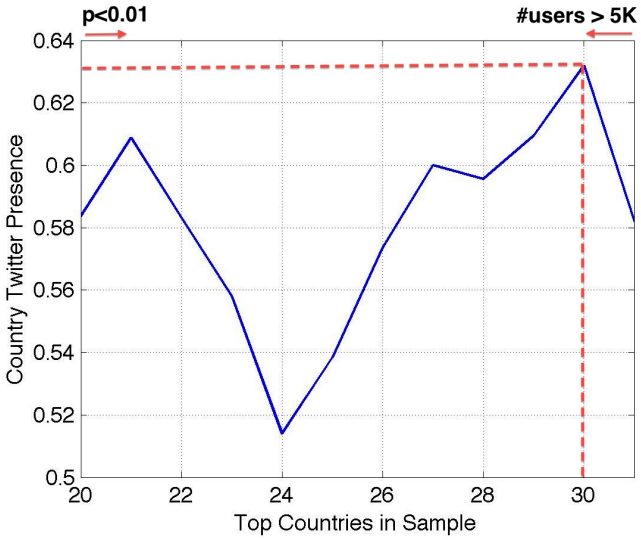


Figure 7.1: Plot of country-level presence on Twitter vs. the number of countries *in our sample*. The highest number of countries for which Twitter presence is significant is around 30. That is, by considering the top 30 countries by number of users, we strike the right balance between representative presence on Twitter and number of countries under study.

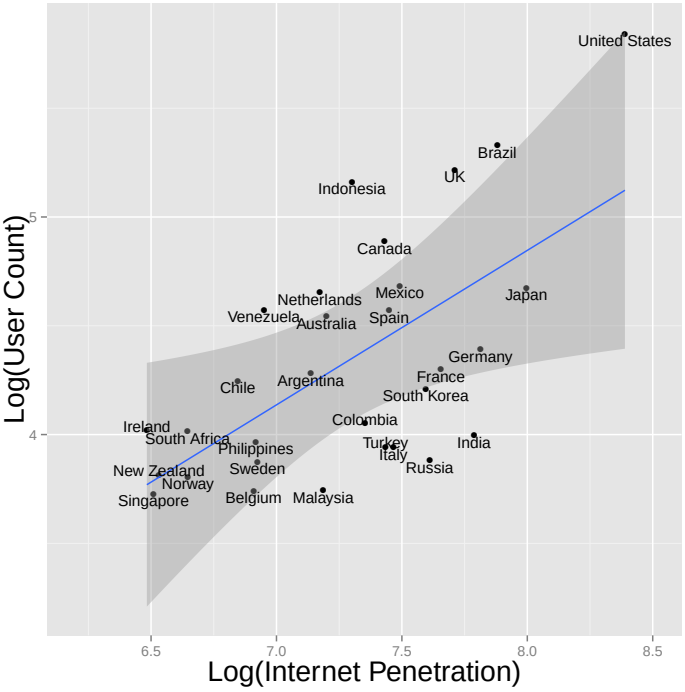


Figure 7.2: Number of users in our sample versus number of Internet users in a country. Both quantities are log-transformed.

late' as being 33 minutes late on average); and people in USA give high importance to time up to the point of associating it with money (*e.g.*, people experience considerable levels of stress if deadlines are not met).

As seen in Chapter 3, another time system was proposed by Hall [55]: monochromic and polychromic time. Monochronic countries are considered more predictable (*e.g.*, United States, Germany, Switzerland, and Japan) and polychromic less predictable (*e.g.*, France, Italy, Greece, Mexico and some Eastern and African countries). The problem is that Hall did not provide any country scores we could use for this study but "Levine's Pace of Life research has been indirectly linked to the observations of Hall (1983) to suggest that polychronicity and Pace of Life are negatively related" [27], and that insight was used in the study of the scheduling tool of Doodle [112].

To paraphrase these ideas in the context of Twitter, we hypothesize the following relationship:

[$H1.1$] *The activities (*e.g.*, mentions, status updates) of users in countries with higher Pace of Life are more temporarily predictable.*

To test this hypothesis, since there are several factors that influence people's routine during weekends, we leave them out and analyze activities during working days, during which the differences between monochronic and polychronic cultures are more salient [23]. After adjusting for the different time zones, we divide each day in five time intervals: sleeping time (00:00 - 05:59), rising time (6:00 - 8:59), working hours (9:00-17:59), dinner (18:00-20:59) and late night (21:00-23:59). This division allows us to separate working hours from the rest of the day and effectively mark changes of activities [51]. Then, to capture each user's predictability, we compute the user's entropy in those five intervals, and we choose entropy because it is often used to characterize unpredictability in time series [122]. Specifically, we consider a measure of entropy proposed by [76; 122] called *temporal-uncorrelated entropy* and adapt it to our context. The temporal-uncorrelated entropy calculates the tweeting randomness across time intervals for a given user and is defined by:

$$-\sum_{j=1}^{N_i} p_i(j) \log_2 p_i(j) \tag{7.1}$$

where $p_i(j)$ is the historical probability that user i posted in time interval j and N_i is the number of distinct time intervals in which user i posted his/her

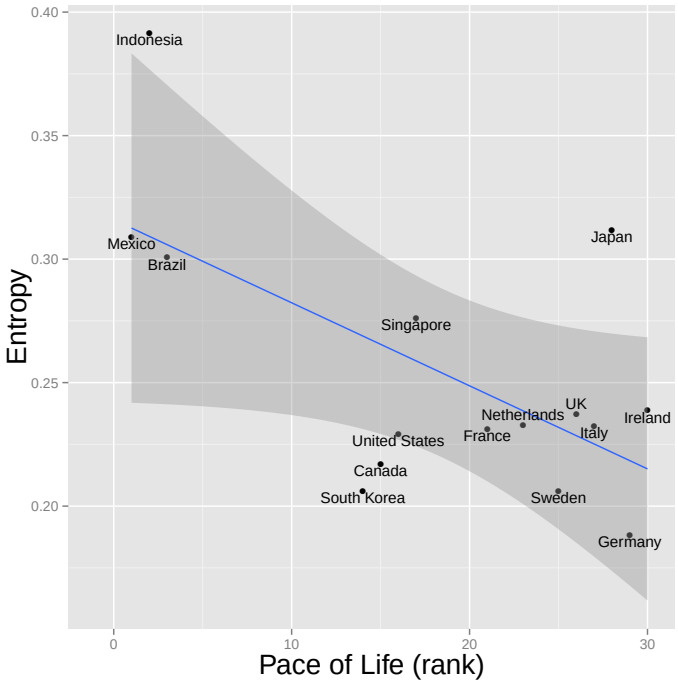


Figure 7.3: Entropy of posting and mentioning activities versus Pace of Life. Countries with high pace of life tend to be temporally predictable.

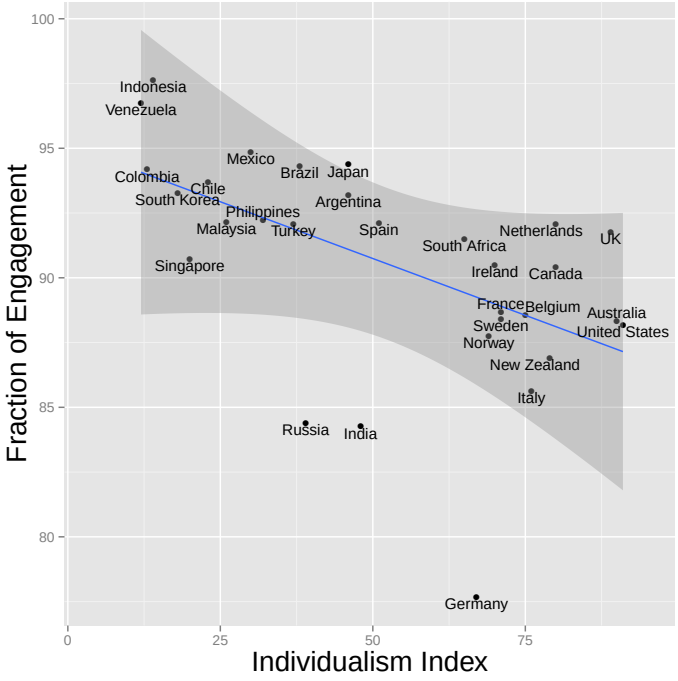


Figure 7.4: Fraction of users engaged with others versus Individualism. In countries with low Individualism, users tend to engage with each other more.

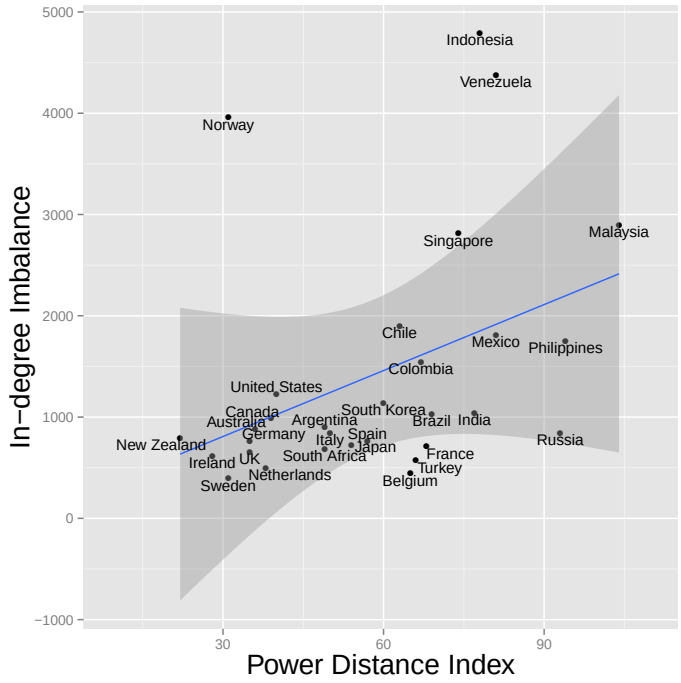


Figure 7.5: In-degree imbalance between user-followee versus Power Distance users have stronger in-degree imbalance.

tweets. The whole sum reflects the (un)predictability of user i posting across all $j's$ intervals.

This metric is computed for the two main activities of posting updates tweets and mentioning others. After obtaining these entropies for all users for these two activities, we compute the Pearson product-moment correlation between the geometric average of the country-level entropy and its corresponding Pace of Life rank. Pearson's correlation $r \in$ [-1, 1] is a measure of the linear relationship between two random variables, whereby 0 indicates no correlation and +1(-1) perfect positive (negative) correlation. Table 7.1 and Figure 7.3 summarize the results. The higher the Pace of Life (monocronic countries), the lower the tweets' temporal unpredictability ($r_{(15)} = -0.62$) and user mentions' ($r_{(15)} = -0.68$).

Figure 7.3 shows the negative relationship between unpredictability and Pace of Life. Thirteen countries follow this relationship but two do not: Japan and Indonesia. Japan's (JPN) Pace of Life is "one of the most de-

	Entropy (Tweets)	Entropy (Mentions)
Pace of Life (overall)	-0.62**	-0.68**
Pace of Life (walking speed)	-0.56**	-0.61**
Pace of Life (post office)	-0.44	-0.50*
Pace of Life (clock accuracy)	-0.45*	-0.51*

Table 7.1: Pearson correlation coefficients between the entropy of the activity in twitter and three measures of the pace of life, p-values are expressed with *'s: $p < 0.05$ $(***)$, $p < 0.05$ $(**)$, and $p < 0.1$ $(*)$.

manding on earth" [84], after Switzerland (SWE), Ireland (IRL) and Germany (DEU), and one would thus expect to find predictable (monochromic) temporal patterns for it. Instead, we find high unpredictability, and that matches what Hall found more than 20 years ago [54]: Japan is an outlier, in that, it mixes high Pace of Life with strong polychronic characteristics, not least because of, Hall suggested, the importance attributed to social relationships. Also Indonesia (our second outlier) shows considerably higher unpredictability than the remaining countries, and that matches what Levine found when he went to one of Jakarta's post office to buy stamps: "It took us considerably longer than in many other countries to find this out." The postal clerk was more interested in conversing about Levine's life rather than fulfilling his request.

Next, we focus on working hours only. Since people in countries with high Pace of Life schedule their time in a linear way, we expect hat they would tweet less during working hours, in proportion, to avoid any interruption:

[$H1.2$] *The percentage of a country's users who have tweeted during working hours negatively correlates with the country's Pace of Life.*
The daily fraction of users in a country who tweet during working hours does indeed negatively correlate with Pace of Life ($r_{(15)} = -0.58$).

7.5 Individualism vs. Collectivism

In addition to pace of life, also human relationships differ across cultures. In high collectivist cultures, users tend to focus more on the community to which they belong: for example, peers tend to unconditionally support superiors' opinions. Such countries (*e.g.*, Indonesia) are characterized by "in-groups", and their members are expected to look after each other. By

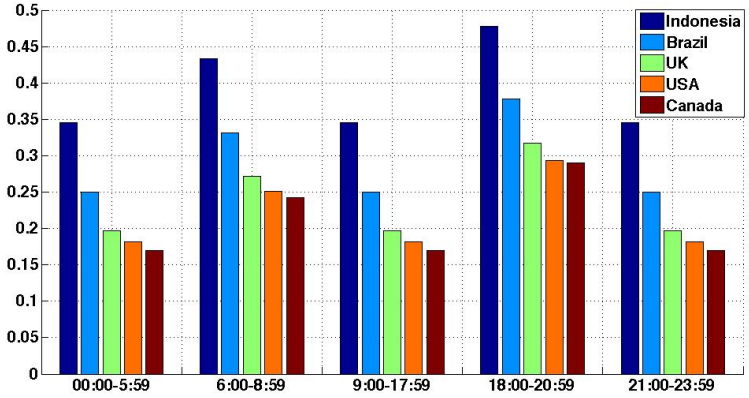


Figure 7.6: Fraction of users engaging with others at different times of the day. Users in Indonesia and Brazil (collectivist countries) engage with others more than those in UK, USA, and Canada (individualistic countries), and they consistently do so throughout the day.

contrast, people from high individualist countries like the U.S. are in a more loosely knit social network, and are generally expected to look after themselves or only after immediate family members [61].

Another characteristic that differentiate collectivist countries from individualist ones is that the former tend to adopt high-context communication as opposed to low-context. In high-context cultures, people tend to emphasize interpersonal relationships. According to Hall, these cultures prefer group harmony and consensus to individual achievement: "flowery language, humility, and elaborate apologies are typical" [55]. Also, the way people acquire information also varies between cultures. According to Hofstede *et al.*, the primary source of information is one's social network in collectivist countries, while it is (news) media in individualistic countries [61]. Finally, the right to privacy is relevant in many individualist societies, while letting one's in-group invade one's private life is acceptable in collective societies.

Based on these studies, one should thus expect that people in collectivist countries will engage into public conversations *more* than what people in individualist countries do. We thus hypothesize that:

[*H*2] *The fraction of users who mention (engage in a conversation with) others negatively correlates with Individualism.*

Using Pearson coefficients, we correlate each country's fraction of users engaging in conversations with the country's Individualism index reported by [61]. We find that users in individualistic countries mention others far less than those in collectivist countries (the correlation coefficient is as high as $r_{s(30)}$ =-0.55 ($p < 0.005$). This consistently holds at different times of the day, and Figure 7.6 exemplifies that by contrasting two high collectivist countries (Indonesia and Brazil) with three high individualistic countries (USA, UK, Canada). Figure 7.4 then shows the correlation between lack of engagement and Individualism, which is high for all countries except for Germany (DEU). This result matches that of a previous study on microblogs: German tweets received the least number of mentions out of the 10 most common languages in Twitter [63]. Also, in Germany, few comments are left in blogs, and users react to comments lower than what users in less individualist countries such as in Russia (RUS) [90].

7.6 Power Distance

Hofstede defines Power Distance as the "extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally." [61]. In countries comfortable with Power Distance, subordinates expect to be told what to do: employees tend to prefer to have a boss who decides autocratically [14]. As such, hierarchy in organizations and inequalities are expected and desired, and that applies not only to work environments but also to schools and families.

The number of a user's followers (indegree) does not necessarily reflect influence but does reflect popularity [8; 24; 77]). Therefore, the power relationship between a pair of users is leveled, if their numbers of followers are comparable; while it is imbalanced, if the numbers of followers greatly differ. Based on this observation, we posit that:

Pace of Life	Correlation
[$H1.1$] *The activities (*e.g.*, mentions, status updates) of users in countries with higher Pace of Life are more temporarily predictable*	$r_{(15)} = -0.62^{**}$ $r_{(15)} = -0.68^{**}$
[$H1.2$] *The percentage of a country's users who have tweeted during working hours negatively correlates with the country's Pace of Life*	$r_{(15)} = -0.58^{**}$
Individualism	Correlation
[$H2$] *The fraction of users who mention (engage in a conversation with) others negatively correlates with Individualism index*	$r_{s(30)} = -0.55^{***}$
Power Distance	Correlation
[$H3$] *In countries comfortable with Power Distance, a pair of users who engage in any type of relationship is likely to show indegree imbalance*	$r_{(30)} = 0.62^{***}$ $r_{(30)} = 0.33^{*}$ $r_{(30)} = 0.42^{**}$

Table 7.2: Pearson correlation coefficients: (*H1.1*) between Pace of Life and the temporal predictability of users' activity (mentions and tweets); (*H1.2*) between Pace of Life and the percentage's of a country's users tweeting during working hours; (*H2*) between Individualism and the fraction of users engaged with others; and (*H3*) between Power Distance and in-degree imbalance shown in three types of relationships ("who follows whom", "who recommends whom" and "who starts to follow whom"). p-values are expressed with *'s: $p < 0.005$ $(***)$, $p < 0.05$ $(**)$, and $p < 0.1$ $(*)$.

[$H3$] *In countries comfortable with Power Distance, a pair of users who engage in any type of relationship is likely to show indegree imbalance.*

We correlate country-level indegree imbalance with corresponding Power Distance. We find imbalance online and Power Distance offline go together for all three types of relationships (Table 7.2): the correlations are $r = 0.65$ for "who follows whom" relationships; $r = 0.33$ for "who recommends whom" relationships, and $r = 0.42$ for "who starts to follow whom" relationships. Figure 7.5 shows the correlation for the "who follows whom" relationship. Norway (NOR), Venezuela (VEN) and Indonesia (IDN) are outliers. It is difficult to see why this is the case for Norway and Venezuela as there is no previous study for them in this matter. For Indonesia, instead, we found that our results match those on blogs: 27% of all blog trends in this country are about pop and celebrities, which may result in Indonesian users following more celebrities than users in other countries [118].

Critics might argue that the number of followers does not necessarily reflect one's popularity, not least because there are "spammers" who accumulate followers by subscribing to random users profiles. This was the case especially in the early years of Twitter [97]. To counter that criticism, in addition to indegree as proxy for popularity, we consider the ratio in-degree/out-degree, correlate the corresponding country-level popularity imbalance with Power Distance, and obtain the same correlation as when considering indegree ($r = 0.67$).

7.7 Why It Matters

We have found strong correlations between country-level behavioral patterns on Twitter and the three cultural aspects of pace of life, individualism and power distance. Those correlations translate into being able to track these three aspects at fine-grained temporal levels - one does not need to wait for the next 10-year effort that replicates Levine's study or Hofstede's; on the contrary, by simply tracking behavioral patterns on Twitter, one could predict the three cultural aspects to a considerable extent for countries that are well represented on Twitter. However, before doing so, one may well wonder why these three aspects matter at all. To see why it would be important to use Twitter to track them, one should consider that the three cultural dimensions have been found to be associated with three main economic indicators: GDP per capita, income inequality and education expenditure [61; 140]. We now test whether these economic indicators do also correlated with our three Twitter features: temporal predictability, activity levels during working hours, engagement with others, and popularity imbalance. To ease explanation, we collate the results in Table 7.3 and comment them next.

GDP per capita. High collectivism was found to be related to countries with low national wealth [61]. To test whether this holds also for our Twitter features, we get hold of the Gross Domestic Product values for our 30 countries (these values reflect purchasing power normalized by population) and correlate them with our four Twitter features. We find that GDP is associated with three features in a statistically significant way (first row in Table 7.3): low-GDP countries tend to be temporally unpredictable ($r = 0.55$), be active during working hours ($r = -0.57$), and feel comfortable with popularity imbalance ($r = -0.48$). Figure 7.7 shows this relationship and associated outliers - Japan and Singapore (SGP). The result for Japan (JPN) is explained by what we found in the Section "Pace of Life." The result

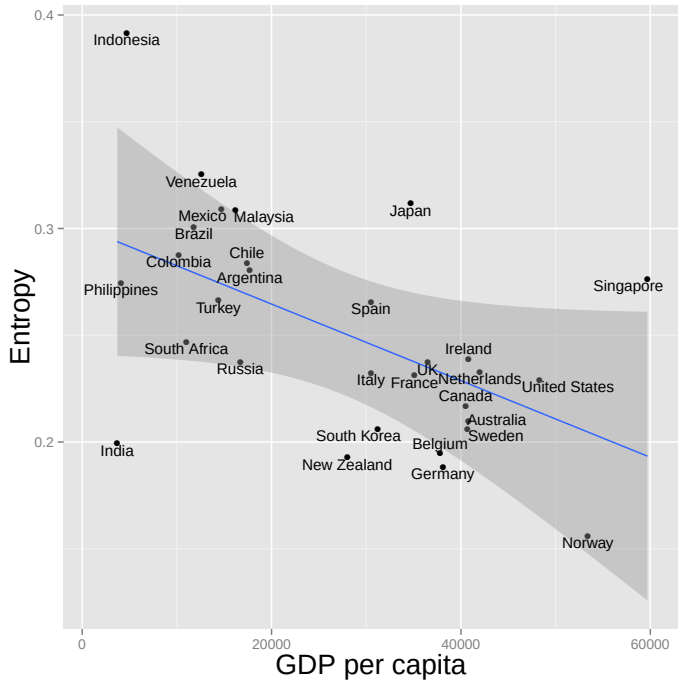


Figure 7.7: Entropy vs GDP: the relationships between Twitter features and socio-economic indicators.

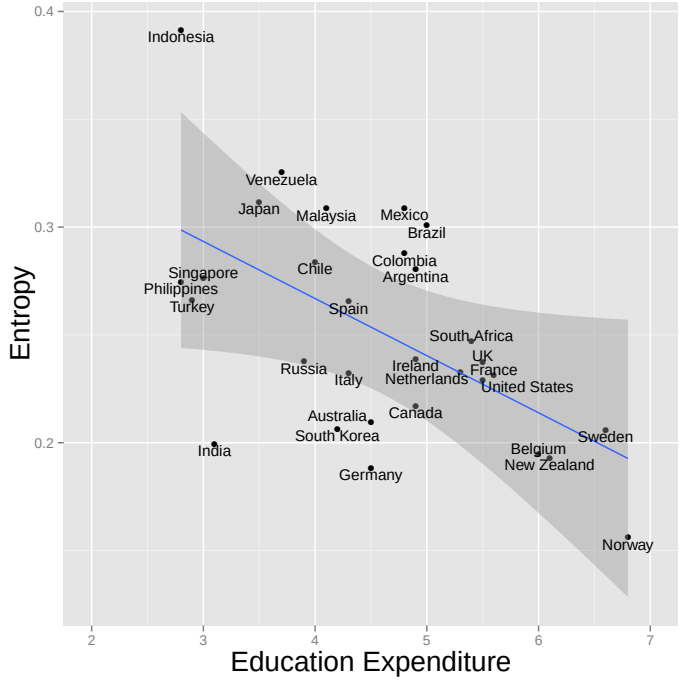


Figure 7.8: Entropy vs Education Expenditure: the relationships between Twitter features and socio-economic indicators.

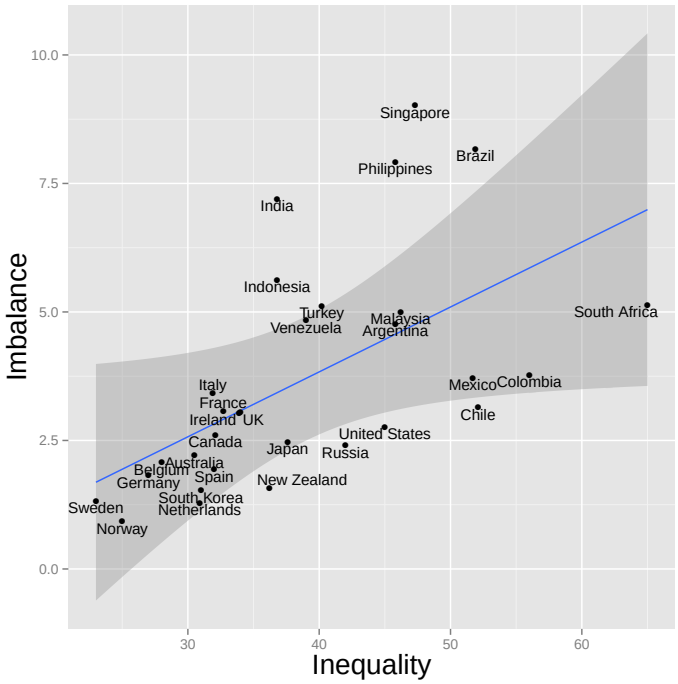


Figure 7.9: Indegree vs. Inequality: The relationships between Twitter features and socio-economic indicators.

for Singapore is explained by considering that the country has faced rapid economic growth rate and is thus "highly developed and enjoys remarkably open and corruption-free environment, stable prices, and a per capita GDP higher than that of most developed countries".[1] At the same time, however, it preserves high collective characteristics typical of most Asian countries, and that explains the association of higher entropy with GDP.

Education Expenditure. We correlate education expenditure (as percentage of GDP) with our four Twitter features and find that countries with low education expenditure are characterized by the same features as countries with low GPD, even if expenditure is normalized by it. They are (second row in Table 7.3): temporally unpredictable ($r = 0.58$), be active during working hours ($r = -0.51$), and feel comfortable with popularity imbalance ($r = -0.60$). Figure 7.8 depicts high correlation for most countries. Again, Indonesia is an outlier, as one would expect from the previous results.

[1]Information taken from the Central Intelligence Agency of USA

Indicator	[$HP1.1$] Predictability	[$HP1.2$] Users (%) in working hours	[$HP2$] Mentions	[$HP3$] Imbalance
GDP per capita	$\mathbf{r_{(30)} = 0.55}^{***}$	$\mathbf{r_{(30)} = -0.57}^{**}$	$r_{(30)} = -0.41^{*}$	$\mathbf{r_{(30)} = -0.48}^{**}$
Education	$\mathbf{r_{(30)} = 0.58}^{***}$	$\mathbf{r_{(30)} = -0.51}^{***}$	$r_{(30)} = -0.24$	$\mathbf{r_{(30)} = -0.60}^{***}$
Inequality	$\mathbf{r_{(30)} = -0.53}^{***}$	$\mathbf{r_{(30)} = 0.49}^{**}$	$r_{(30)} = 0.39^{*}$	$\mathbf{r_{(30)} = 0.58}^{***}$

Table 7.3: Pearson correlation coefficients of three socio-economic indicators (first column) with: predictability (second column), activity in working hours (third column), mentions (fourth column) and in-degree imbalance (fifth column). p-values are expressed with *'s: $p < 0.005$ $(***)$, $p < 0.05$ $(**)$, and $p < 0.5$ $(*)$.

Income Inequality. Power distance was found to be related to the use of violence in domestic politics and to income inequality [61]. One widely-used way to measure income inequality is the *Gini* coefficient. This measures the degree of inequality in the distribution of family income in a country [140]. The lower its value, the more equal a society is. We find that unequal countries tend to be (third row in Table 7.3): temporally unpredictable ($r = -0.53$), be active during working hours ($r = 0.49$), and feel comfortable with popularity imbalance ($r = 0.58$). It should come as no surprise this last result: that the strongest predictor of income inequality is popularity imbalance (popularity inequality) in Twitter. Figure 7.9 shows that India (IND), Indonesia (IDN), Philippines (PHL), Singapore (SGP) and Brazil (BRA) are outliers. That is because these countries are characterized by disproportionately high levels of inequality [140].

7.8 Discussion

Social media sites often assume that people from different countries use their services in very similar ways. By contrast, we find that the use of Twitter considerably changes across them. Fortunately, these changes are not random but are *predictable* so much so that simple country-level behavioral features derived from Twitter strongly correlate with cultural dimensions. Users in monochronic countries tend to be temporarily predictable, those in collectivist countries considerably talk with each other, and those in countries uncomfortable with power distance will not preferentially engage only

with popular users. These findings might not only have theoretical implications for future cross-cultural studies but might also have practical implications, including the prediction of country-level economic indicators at fine-grained temporal level, and the design of culture aware recommender system.

7.8.1 Theoretical Implications

Twitter is a distal communication modality (distal in the sense that users are separated in space and time), and it has been argued that it is not a social-networking tool but a broadcasting platform of, for example, news and opinions [77]. Yet, our cultural analysis suggests that Twitter enjoys social-networking features, and that engagement is predominant among users in collectivist countries. This study not only has suggested the extent to which Twitter use is associated with specific culture dimensions, but also points to the possibility that social media sites could be used to run large-scale cross-cultural studies and could ultimately become tools that promote computational social science. This is is a new discipline that aims at using large archives of naturalistically-created behavioral data (of, for example, emails, tweets, Facebook contacts) to answer social science questions [80; 38].

7.8.2 Practical Implications

Our findings could also be used to design:

Culture-aware engagement tools. In collectivist countries, users do engage with each other by exchanging messages and recommending others. One could design country-tailored tools that: promote interactions with strangers in individualistic countries, and with strong ties in collectivist countries; rank status updates based on interestingness in small-power-distance countries, and on popularity in large-power-distance; targets ads in specific time of the day for monochronic countries, and in user-tailored for polychronic countries.

Culture-aware people recommender. One increasingly important feature in Twitter is its people recommender system, which suggests people one might know. This tool makes suggestions based on structural features (*e.g.*, common followers) and on content features (*e.g.*, matching one's topics of interest). However, the tool might well benefit from cultural dimensions as well: recommending strangers is fine in individualist countries but not in

collectivist ones; or users in large-power-distance are likely to preferentially follow highly-popular users.

7.8.3 Limitations and Future Work

Despite the strong correlations, this study suffers from five limitations. First, our sample was collected in a specific time frame. Critics might rightly say that our findings may be co-founded by the days data was crawled. However, the sample spans 10 weeks and, as such, it might be large enough to capture the normal routine of users. Second, we might run the risk to promote stereotyping of individuals based on their countries of origin. This study is about 'mean behavior', and one should consider that there is high variability across individuals in the same country. Third, we naively equated use of mentions with "engagement with others", but that might not be necessarily the case. That is why, in the future, it might be beneficial to propose a taxonomy that will distinguish one's purposes when mentioning others (*i.e.*, conversational, informative, attribution). Fourth, this has been an exploratory study in which causal inference has not been established (and it was not the aim of the study). However, there are two remarks to be made: a) many of the observed relations on Twitter confirm those that are already known in the real world; and b) some of the correlations are weak, but others are very strong, suggesting a dose-response form *from* country characteristic *to* behavior on Twitter. Fifth, we have focused on language-independent features. In the future, we will explore how the use of language changes depending on cultural dimensions [108]. For example, do individualistic countries use more singular first-person pronouns (*e.g.*, I, my, mine, yo, eu, moi)?

Chapter 8

Communication: Cultural and Socio-economic Factors

8.1 Introduction

The rise of the Internet and social networks have lead some researchers to hypothesize that "distance is dead" [19] or is not longer important to make social contacts. At the very conception of online networking pundits predicted the loosening of the "grip of geography" [20], foreseeing the strengthening of the bonds between people with the same interest in different parts of the world, and globalization of both the workforce and the scope of governmental considerations. Nevertheless, empirical studies have shown that distance still matters in online communication, including email [123; 96] and instant messages [82], with these new modes of communication reinforcing the strong ties we make in person. However, recently other factors were shown to mediate the effect of distance, including language, air travel frequency [128], and culture [123]. For instance, countries sharing cultural features have a higher affinity in international email exchanges, and can be effectively clustered into "civilizations", as suggested by Samuel Huntington in "The Clash of Civilizations" [65].

Finding whether Internet users are trapped in socio-economic or cultural "bubbles", despite the supposed freedom and multi-cultural nature of the web, is a first step to identifying the blind spots in our communication. Specifically, cultural dimensions have long been studied by sociologists. To measure cultural values, as they relate to personal behavior, we use Hofstede's culture indexes [59]. We bring these into the realm of social media

analysis by relating the international communication flows in Twitter to the extent to which countries share these cultural characteristics and various other country-specific attributes.

Recent wide adoption of Twitter has fostered a global network of relatively weak ties based on user interests. As defined by Granovetter [52], the strength of a tie is "a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie". Since Twitter messages are short (140 characters), and are broadcast publicly, easy to both read and to ignore, Twitter provides a perfect platform for the establishment of weak ties. Also, the connections need not be reciprocal, and users are free to 'follow' (subscribe to) any other user with a public profile in order to see the posts, or status updates, of that user in their timeline. Furthermore, users are free to contact others (being their followers or not) by simply mentioning their users names. The value of such characteristics make Twitter a useful tool for exploring communication in online social media, going beyond the strong ties of personal e-mails or Facebook. Indeed, we find major differences between the importance of economic and cultural factors in Twitter communication as compared to e-mail, as described in [123].

Here, we explore how various factors (distance, social, economic, and cultural dimensions) shape the cross-country communication through the lightweight social networking services. Specifically, we address two questions: (1) To what extent does distance determine the informal communication of users from different nations? and (2) To what extent do social, economic and cultural factors mediate/impede this communication? To tackle these questions, we study user mentions among 13 million geolocated users during a 10 week period from March to May 2011. Using this data, covering 111 countries, along with country-specific statistics gathered from outside sources (CIA, World Banks and World Values Survey), we make two main contributions:

- We employ the *gravity model* [75], which uses node population and physical distance, to construct a baseline communication network, and test to which extent it estimates cross-country Twitter communication. We use the Haversine distance between two countries and two population proxies: country population and the Internet penetration, with the latter showing moderate correlation with the number of mentions and retweets of 5,932 pairs of countries ($r = 0.68$ for unique mentions, and $r = 0.66$ for unique retweets).

- We build a regression model that uses economic, social and cultural country attributes, along with the gravity model to predict communication volume between pairs of countries. We find that the complete model performs well with *Adjusted* $R^2 = 0.80$, illustrating the importance of social economic and cultural variables in bilateral online communication.

We conclude by discussing the design implications of these findings in the realms of collaborative work, software design, and recommendation systems. This chapter is based on [45].

8.2 Related Work

A number of studies have used confidential communication to examine the social connections between individuals across the world. A well known study by Leskovec and Horvitz [82] uses the private messages to build a "planetary scale" social network of 180 million nodes, and examines social phenomena, such as Milgram's "6 degrees of separation" [132] (finding that, indeed, the users of the service had an average path length of 6.6).

Specialized communication has also been considered. A community of travelers on CouchSurfing.com was studied by Lauterbach *et al.*, [79] who attempted to predict the trust the users display toward one another. They show that, among more personal variables (such as whether the users have met in person), whether users are from the same country affects the chances of one user vouching for another. Olson *et al.* [105] carried out empirical studies of remote work, both in the field and in the laboratory, concluding that distance impacts the quality of end result, regardless of the technology used. More recently, Takhteyev [127] discussed examples of successful collaboration over long distances by looking at how several cultural and geographic constrains were negotiated in the face of increasingly "global" knowledge and technology.

Across social media, geographical distance has been shown to play a major role in human connections. Scellato *et al.* [115] show that, among the users in Brightkite, Foursquare, and Gowalla communities, 40% of links are made in a radius of under 100km. Similar results were found of a 2,852-user sample of the Twitter network in 2009 by Takhteyev *et al.* [128] with 35% of links being under 100km but they also find that other variables, such as the commonalities in language and the extent of air travel to be more pre-

dictive of Twitter communication than physical distance. They speculate that air travel may stand as "a proxy for other kinds of pre-existing connections between places, which in turn influence formation of electronic ties". Inspired by this, we examine social, economic, and cultural factors in international Twitter communication. However, a marked difference between these previous studies and one described here is the network construction process. Instead of using follower or followee edges (subscriptions), which do not necessarily imply active communication or attention [134] with 25% of Twitter users never tweeting at all [12], we use geolocated user mentions in nearly 3 billion posts.

Though first, we use the *Gravity Model* to capture the effect of distance. Inspired by Newtonian physics, it models the importance of physical distance in communication between two populated nodes, using a proportion of the population sizes of the two nodes to the distance between them [75]. For example, it has been applied to modeling road and airline networks [10; 69], phone calls [75], and flows of passengers in a London metro system [121].

An email communication study by State *et al.* [123] examines the extent to which inter-national communication flows according to the civilization of users, as defined by Samuel Huntington in *The Clash of Civilizations* [65]. By including geographic, economic and cultural factors in their regression model (including the first four of Hofstede's cultural indexes), they show that the membership in the same Huntington civilization to nearly double the pairwise communication density, increasing it by factor of 1.941. However, since the notion of civilization encompasses both geographic attributes of the countries, we find it unsuitable in our aim of separating geography from culture. Moreover, Lipi *et al.* [4] measured how culture affects nonverbal expressions in conversations. They proposed a parameter based model employing Bayesian network techniques where culture is connected to Hofstede dimensions which are also connected with nonverbal expressions. For each node in the Bayesian network, probabilities are assigned according to Hofstede's cultural dimensions and to results obtained from experiments. Basically, when a country is chosen, the model outputs the estimated probability of expressive parameters (Rigidness, Mirroring, etc).

In summary, to the best of our knowledge, we present a previously unattempted study of international Twitter communication which combines cultural information with geographic, economic, and social features, using a variety of outside sources from the CIA and World Bank. Unlike e-mail, Twitter mention graph goes beyond strong ties of inter-personal communi-

cation, potentially breaking down barriers of distance and culture. Also, unlike the previous studies on Twitter which use subscriptions instead of tweet content [114; 128], this study focuses on the active conversation and attention beyond the user's immediate follower/followee network. Finally, we use the gravity model and a variety of other predictors to build a communication model based solely on data independent of the particulars of Twitter data set.

8.3 Data Set

We model communication across countries in Twitter by observing mentions and retweets by users in one country involving users from another. Similar to [123], we say that a communication is established from country a to b when a message is posted by a user from a mentioning a user from b. A "mention" consists of any Twitter username preceded by the *at* symbol (@). So, for example, if user @Maria located in Spain creates a post "@BarackObama is the president of the United States," we know two things: a) BarackObama received a notification in his account (although unlikely to answer) and b) a communication was made from Spain to USA. The interpretation of this phenomena consists of both conversation and attention, in that, mentions and retweets may be used in a conversation between users, but may also signify an awareness of another user (as with BarackObama and the notification he received in the previous example). Thus, in this chapter, when we refer to the number of mentions or retweets as "communication", we do so loosely.

The data was collected similarly to Chapter 4. We first randomly selected 55K users who tweeted at least once on March 2011 and obtained their profile information. From this information, we selected users with out-degree and in-degree in the range of $[100, 1000]$ and crawled their corresponding followee network (for a user u, it is all users who u is following).

We then proceeded to collect all of the tweets posted by the original 55K users as well as their followees during 10 weeks starting from the second half of Mach 2011. We also collect all tweets containing a mention of any user of our sample (*i.e.*, identified by @username) and the user profile of who posted these tweets.

We continue by finding the geolocation of each user via the location field entered in their profiles as explained in Section 2.4, resulting in 13 million geo-located users. To alleviate any bias due to the selection of seed users

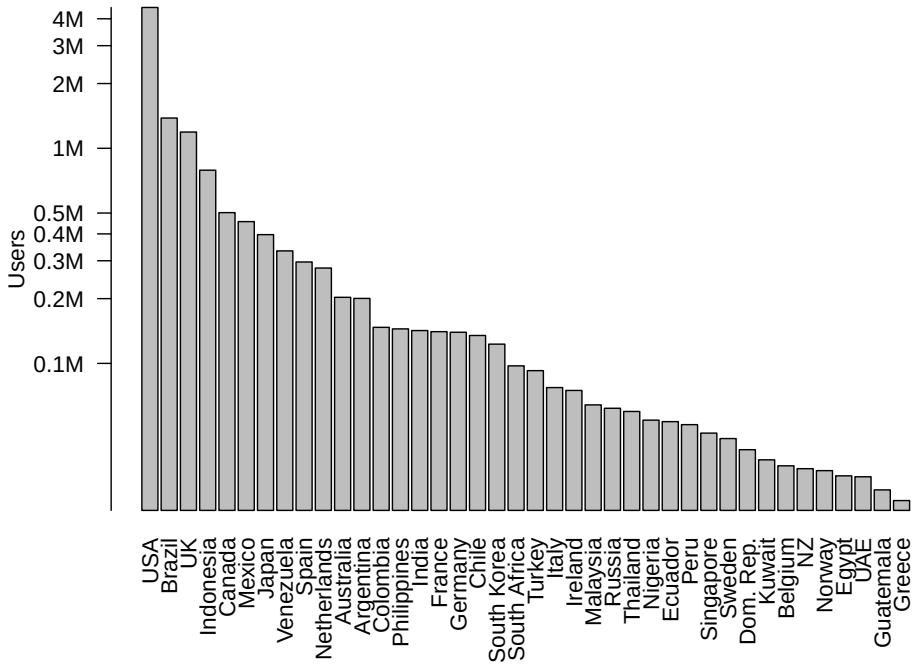


Figure 8.1: Number of users per country in the sample in logarithmic scale (showing top 40 countries).

and obtain representative samples, we only consider the countries with more than 1,000 users in our sample.

Figure 8.1 shows sample sizes, with the typical skew across the countries with USA having by far the largest share of Twitter users, followed by Brazil, United Kingdom, and Indonesia. Seven out of the top 10 countries in our sample overlap with the top 10 countries by site traffic in 2011[1] and we also find a *Pearson correlation* of 0.72 to the corresponding logarithm of the number of internet users in 2011 reported by the US's Central Intelligence Agency (CIA). The discrepancy can be attributed to our sampling method which favors users who are mentioned, thus promoting Mexico, Venezuela, and Netherlands in our ranking, excluding Germany, India, and Australia which appear in the traffic ranking.

[1] http://www.marketinggum.com/twitter-statistics-2011-updated-stats/.

Data	Total
# of geolocated users	13,139,763
# of tweets (with mentions)	2,924,398,138
# of mentions	534,868,476
# of unique mentions	258,534,246
# of countries with $> 1K$ users	111
# of country pairs with complete predictors	481

Table 8.1: Summary of the data set. We identified the geolocation of more than 13M users but considered only the countries with more than 1K users, which represents more than 90% of our sample users. In total, we obtained 481 country-pairs with no missing attribute values for regression analysis.

Table 8.1 shows statistics about the final data set, including the number of geolocated users and their tweets, and mentions found in those tweets. We count "unique" mentions per user, summing the number of unique accounts mentioned by each. On average, for each user-user conversation, there is one duplication, since it is common to mention a specific user more than once (same holds true for unique retweets).

We analyze the communication across countries by mapping the country of the mentioned users to the countries of those who posted the tweets, obtaining a country to country graph. Since we are interested in measuring the flow of information between countries and not the direction of it, we obtain an undirected graph of the inter-country communication by adding the bilateral number of mentions and retweets between a pair of countries. Furthermore, we discard self-loop edges since we are interested in communication across countries, not within. This resulted in $5,392$ country-country pairs.

Finally, to tackle our hypotheses and objectives, we obtain geographic, social, economic, and cultural features of these countries. We collect the number of direct flights between each of the countries,[2] as well as the spoken languages in each country, as reported by the CIA. Additional social, economic and cultural indicators came from the WorldBank API for R.[3] Each of the variables is explained in the *Social, Economic and Cultural Predictors* section. Since data was missing for some of the countries, we excluded the

[2] http://openflights.org/data.html.
[3] http://www.r-chart.com/2010/06/world-bank-api-r-package-available.html.

	Sample Size	γ	Internet Penet.	γ	Country Population	γ
Mentions	0.915	0.83	0.670	0.42	0.489	0.84
Unique mentions	0.919	0.83	0.679	0.43	0.501	0.84
Retweets	0.911	0.88	0.676	0.49	0.505	0.92
Unique retweets	0.904	0.87	0.663	0.48	0.492	0.91

Table 8.2: Pearson correlation between observed Twitter interactions and gravity model estimations using three different population masses ($N = 5392$ country pairs) and adjusted distance exponent (γ).

records with no values. This gave us a total of 481 pairs with complete information for each predictor variable in our model.

8.4 Gravity Model

In its simplest formulation, the Gravity Model posits that the gravitational interaction between two places is proportional to their mass and inversely proportional to the distance between [148] and it takes the form of:

$$I_{1,2} = k\frac{p_1^{\alpha} * p_2^{\beta}}{d_{1,2}^{\gamma}} \tag{8.1}$$

where $I_{1,2}$ is the volume of interaction between communities 1 and 2, k is a constant, p_1 and p_2 refer to the "population mass" (that is, community size) of communities 1 and 2, and $d_{1,2}$ refers to the distance between these communities. The exponents α, β, γ and the scaling factor k are adjustable parameters chosen to fit the data modeled. The pure gravity model is retained if the population exponents (α and β) are 1 and the distance exponent (γ) is 2; but the formula allows the exponents to be adjusted to finely tune the data being modeled.

It has been previously shown that the gravitational model is applicable to various phenomena such as telecommunication, email and transportation flow between countries, cities and within cities [120]. Since the gravity model can be used to account for any interaction or flow from one place to another, we apply it to estimate the volume of Twitter traffic between two countries and adjust the γ exponent to better fit our data.

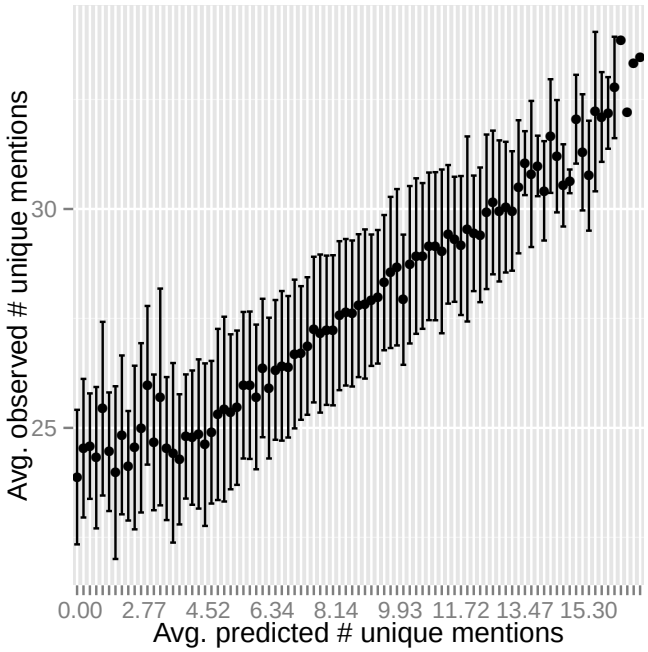
We employ several alternative proxies for the "population mass" in the gravity model: (i) sample size, (ii) internet penetration and (iii) country population; and as proxy for distance, we use the Haversine distance (distance between two points on a sphere).[4] We examine the correlation with Twitter bilateral interactions, measured as a) the number of mentions, b) unique mentions, c) retweets, and d) unique retweets. For the case of re-tweets (cases c and d), we counted only the original authors of a tweet ignoring all other mentions.

Table 8.2 shows the Pearson correlation between Twitter interactions and gravity model estimations using three different estimates for the population mass, along with the best values of γ for each. Sample size produces the strongest correlation with all four measures, at $r = 0.919$ with unique mentions, with no significant difference in communication flow across countries between re-tweets and mentions. In the following regression experiments, to make sure the dependent variable is not related to the predictors, instead of using sample size we use Internet penetration as a proxy for population.

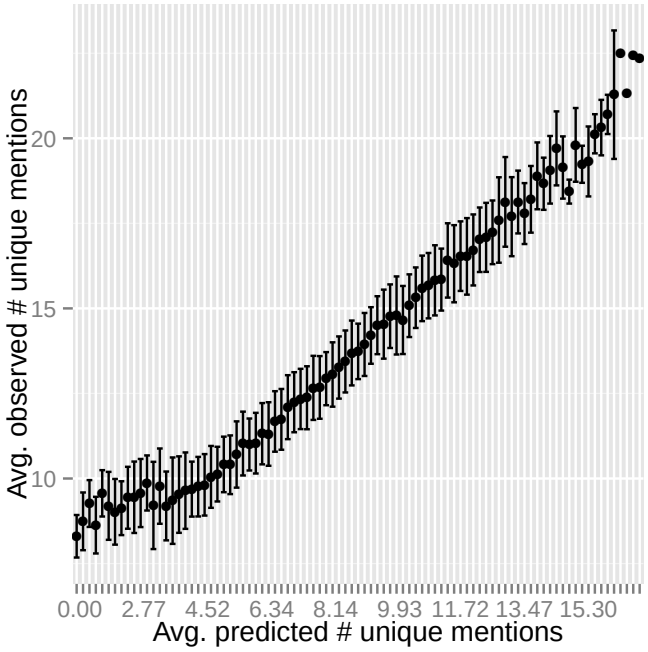
In Figure 8.2, we show the distribution of observed unique mentions vs. estimated mentions flows, using (a) internet penetration and (b) sample size as proxies for population. We see that both versions of gravity model provide estimates which correlate well with observed Twitter interactions. Sample size provides the best estimation, but the standard deviation tends to increase as communication increases.

Finally, we visualize the mention networks induced by these two measures by selecting the top 1000 strongest edges and top 50 edges (Figure 8.3). The nodes are positioned using force-directed algorithm using log-transformed edge weights, and colored according to the continent on which they reside. The countries in gravity model network are largely clustered according to their geography, with most populated countries (Brazil, US, UK) connecting them at the center. The network built using unique mentions also has UK and US as central hubs, however the geographical differences between the countries are less pronounced. Now, Spain is much closer to Mexico and its South American peers in language. In larger mentions network, countries with a smaller Twitter use are also pushed out into their own group in the upper right, including countries from Africa, Eastern Europe, and Middle East. Although the major players remain the same (partially because we are using our sample size in the gravity model), the geographic separations are less pronounced in the Twitter-induced network. Next, we examine

[4] https://github.com/linkedin/datafu.



(a)



(b)

Figure 8.2: Unique mentions versus gravity model using (a) internet penetration and (b) sample size, with standard deviations of unique mentions. The country pairs are first binned by estimated flow, then we plot the mean estimated flow in each bin vs. the mean observed flow of the edges in each bin. The error bars show the standard deviation of the observed flows in each bin.

the extent to which international communication is explained using features other than physical distance.

8.5 Social, Economic and Cultural Predictors

The high correlation of retweets and mentions with the gravity model testifies to the importance of distance. Nevertheless, the standard deviation of the observations seen in Figure 8.2 show that there are other factors to take into account when studying cross country communication. We observe an acute tendency to overestimate communication flows. This is especially the case for European countries which are very close together and have a large number of people online, including Germany, The Netherlands, Poland, and United Kingdom. Same is true for China and Japan. Similarly, communication between countries which are far apart, such as United States and United Kingdom, and United States and Australia, is underestimated. This behavior suggests that the model does not take into account important information, such as culture and other international connections. We now proceed to study 16 variables that we hypothesize will impact communication. These variables are classified into social, economic and cultural.

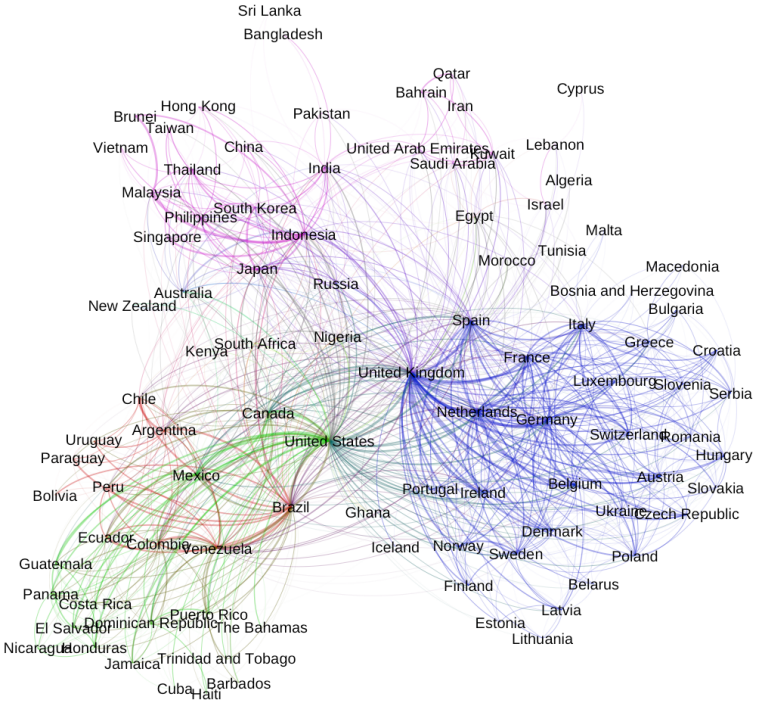
8.5.1 Economic Indicators

Does difference in income divide people? Despite the fact that the so called "liberation technologies" have and continue to alter information propagation across countries during crises, the boundaries separating high- from low-income countries affect the daily real world interactions between people, and therefore affect interactions online [93]. In fact, where income differences are bigger, social distances are bigger and social stratification more important [140]. We use predictors (in American dollars) that account for economy described as follows.

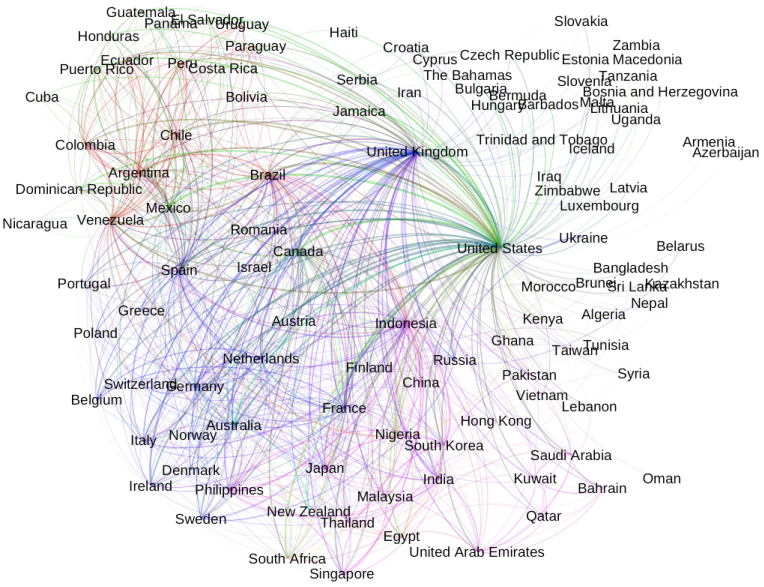
Income: We take the GDP per capita for each pair of countries and multiply them. A high product stands for the combination of two wealthy nations.

Similarly, the trade relationships between countries has been shown to be affected by ease of communication [66]. For this reason, we assume the trade between two countries should also be taken as a predictors of communication and we do so under three perspectives (metrics obtained from the World Bank):[5]

[5] http://wits.worldbank.org/.

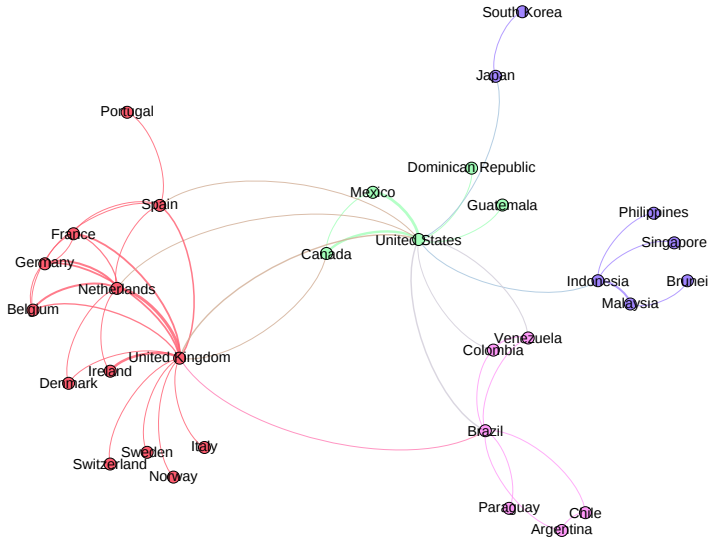


(a) Gravity Model

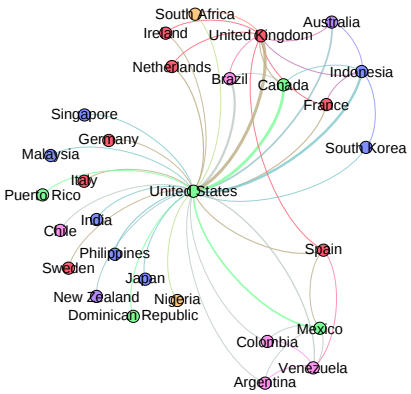


(b) Unique Mentions

Figure 8.3: Cross-country communication network, 1000 most prominent edges, color-coded by continent.



(a) Gravity Model



(b) Unique Mentions

Figure 8.4: Cross-country communication network, 50 most prominent edges, color-coded by continent.

Export importance: We propose a metric that measures the importance of the exports between two countries with regard to the overall exportations of both countries − we add their pair-wise exports and normalize by the total sum of all their exports.

Trade intensity index: The value of trade between two countries on the basis of their importance in the world trade. This metric is defined as the share of one country's exports going to a partner divided by the share of world exports going to the partner. To obtain one value per country pair, we multiply their corresponding trade intensity index.

Trade market share of total exports: It measures how much of the world import demand is covered by the country's exports. Similarly, to obtain a single value per country pair, we multiply the share of total exports for each country.

8.5.2 Social Indicators

We present four social variables expected to affect information flow related to migration and air travel.

The term "transnational migrants" refers to the extent to which immigrants keep cross-border ties when sharing political or religious ideas as well as maintaining cross-border activities of travel, remittance flow and telephone communication with their home-country. These interlocking networks across national boundaries are even more evident with people from border-free travel zones where individuals can work and live in a different country and travel regularly to their home-countries without major bureaucratic barriers.

We propose four migration metrics:

Net migration rate: the difference between the number of persons entering and leaving a country during the year (per 1000 persons). A positive value indicates an excess in immigration and a negative number an excess in emigration. We calculate the absolute difference between these values for each country.

Emigration: obtained by summing the number of emigrants from one country to the other divided by the total number of emigrants for both countries.
Migration: obtained by summing the number of emigrants from one country to the other divided by the total.

Direct flights: The availability of direct fights has been proven to mediate distance when measuring social interactions [128]. Besides simplifying the

process of travel of immigrants to their home-country, it fosters interactions between tourists, visitors and business partners. We consider the number of direct flights between each pair of countries.

8.5.3 Cultural Indicators

We use Hofstede's cultural dimensions (see Chapter 3 and take the absolute value of the difference between the cultural indexes of a pair of countries to measure the cultural differences between their inhabitants.

To this, we have also added *Racial intolerance* as one more dimension that can strongly affect communication between people from different countries. Researches have shown that there is a causal relationship between well-being, economic freedom and tolerance [13]. Using the "World Values Survey", racial intolerance was measured in more than 80 countries by asking participants what kinds of people they would not want as neighbors answered and calculating the percentage of those who answered "people of a different race" option. We calculate intolerance as the maximum percentage reported by the survey between a pair of countries: the highest intolerance will determine the level of communication with people from the other country.

Finally, we add *Language* to this category because it defines a culture, through the people who speak it and what it allows speakers to say. Many immigrants and tourists choose to travel to places where they can communicate, as we tend to establish social ties with people who can speak the same language. The CIA provides a rank ordering of spoken languages per country. We set the binary variable *language* to 0 if there is no common language between two countries and to 1 if there is.

8.6 Regression

To verify the predictability power of the gravity model, as well as the economic, social and cultural variables (summarized in Table 8.3), we run a regression analysis, and build a model to predict the normalized volume of mentions between the countries. To avoid an excess of variables versus data points, we only consider the pairs with no missing values (resulting in 481 pairs). Finally, to account for the violations of normality exhibited by the distributions in Table 8.3, every variable is log transformed and then standardized.

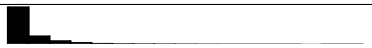
	Distribution	Max
Unique Mentions		58,214,512
Gravity Model		2,731,487
Economic Variables		
Income		802,604.5
Exports		0.35
Trade Intensity		395.8
Trade Market Share		92.6
Social Variables		
Routes		6.68
Emigration		0.83
Migration		0.05
Migration Rate		39.9
Cultural Variables		
Language		1
Intolerance		86
Power Distance		82
Individualism		84
Masculinity		90
Uncertainty avoidance		104
Long Term Orientation		88
Indulgence vs. Restrain		97

Table 8.3: Statistics of regression variables: unique mentions (dependent variable) and 17 independent variables, collected for 5,392 country pairs. The distributions begin at zero and end at the adjacent maximum. Language and income group are categorical variables converted to numeric factors. There are 481 pairs having values for all the predictive variables.

We define communication strength as the communication volume between two countries, as measured by the number of unique user mentions between users of two countries. We choose this measure as it encompasses both conversations and unsolicited mentions of users. We normalize the raw unique mention volume between countries to a scale of $\{0, 1\}$ in order to represent the communication flow strength of a pair of countries in comparison to the the rest. The transformation was made by:

$$s_{i,j} = \frac{m_{i,j} - min_m}{max_m - min_m} \tag{8.2}$$

where $s_{i,j}$ is a normalized mention volume, $m_{i,j}$ is the number of unique mentions from i to j and vise versa, and min_m and max_m are the minimum and the maximum observed unique mentions between any pair of countries in the data set.

We use multiple linear regression to predict our dependent variable. Consequently, we model communication strength as a linear combination of the predictive variables and the gravity model:

$$cs_{i,j} = \alpha + \beta_1 G_{i,j} + \beta_2 R_{i,j} + \beta_3 D_{i,j} + \epsilon_{i,j} \tag{8.3}$$

where $cs_{i,j}$ is the communication strength between the i-th and j-th country, $G_{i,j}$ is the gravitational model variable, $R_{i,j}$ is the vector of remaining predictive variables (classified into social, economic and cultural), $D_{i,j}$ represents the pairwise interactions between all the predictors, and $\epsilon_{i,j}$ is the error term.

Multicollinearity. Before applying the model, we check for multicollinearity among the model's variables. We employ Variance Inflation Factors (VIF), which measure the extent to which errors of the estimated coefficients are inflated by the existence of correlation among the predictor variables in the model [133]. We detected two groups of variables for which VIF was high (above 4): one dealing with trade: Trade Intensity at 7.3 and Trade Market Share at 12.3, and with migration: Migration at 50.2 and Emigration at 48.6. One way of eliminating multicollinearity is to remove one of the violating predictors. Thus, we exclude Trade Intensity and Emigration from the analysis, with the resulting model showing VIFs of under 4.

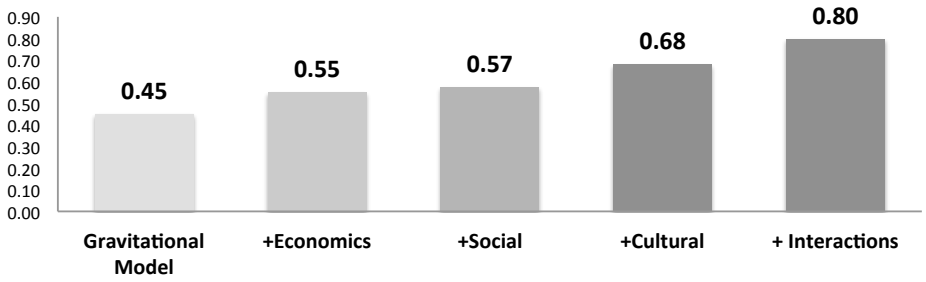


Figure 8.5: Adjusted R^2 as new dimensions are added to the model. Modeling interactions between dimensions results in substantial performance boost.

8.7 Results

When predicting the normalized communication volume, the complete model fits the data very well, with an *Adjusted* $R^2 = 0.80$ at $p < 0.001$ and a standard error of 0.087. This error interval is of less significance for countries with communication close to 1. Note that if we use sample size as a proxy for population, the full model achieves *Adjusted* $R^2 = 0.95$, suggesting overfitting. As mentioned in *Gravity Model* section, we instead use Internet population in order to remove the effects of our sampling method. As part of a regression routine, we have looked for normality of residuals in a QQplot and observed that they follow approximately a normal distribution except for the top and bottom of the line. These outliers are understandable in this situation and therefore should not be considered as evidence for instability of the model [46].

Figure 8.5 summarizes the model's performance for communication volume broken down by four predictor groups. There are notable gains when adding economic and cultural predictors to the model, but it is the interaction term that is responsible for boosting the performance to *Adjusted* $R^2 = 0.80$.

Figure 8.6 visualizes the predictive power of the four dimensions as part of communication volume. For this figure, we have not included interactions in order to analyze each dimension individually (recall that we have controlled for the multicollinearity). The weight of a dimension is calculated by summing the coefficients of the variables belonging to it as in [48]. As described in more detail later, cultural factors (most prominently language) and the gravity model play the largest role (also, Figure 8.5 shows that

indeed cultural features are more important in boosting the results).

Next, we add interaction factors to our model. Table 8.4 presents the coefficients of the top 12 predictive variables ordered by a) beta coefficient and b) *t-value*. Trade, cultural dimension of Masculinity vs. Femininity (MAS), and gravity model and its combinations show the highest significant coefficients. Gravity model alone, as well as in combination with the economic variable of trade, exports, and cultural variable of language is high on the coefficient ranking. However, by the magnitude of the coefficient (at 0.165) Trade Market Share proves to be an even better predictor. Among the cultural variables, we see MAS to have the highest coefficient, followed by intolerance with a negative coefficient at -0.054. Language in combination with the gravity model proves to be a more significant predictor than language alone.

If we consider the *t-values*, which signify a variable's importance in the presence of other variables (the left column of Table 8.4), we find three significant combinations of cultural attributes. The most significant is the interaction between intolerance and the cultural dimension of Long-Term Orientation (LTO) (*t-value* $= 3.66$) and intolerance and Uncertainty Avoidance (UAI) (*t-value* $= 2.83$). The dimensions of LTO and UAI are linked to tradition, nationalism, and the fear of the unknown [59]. For example, studies show that in Japan (ranking high in LTO and UAI), studies show that people avoid communication with non-Japanese for fear of failing to understand and interact with strangers from different cultures which reflects the way of how strangers are treated [35]. UAI and LTO combined with the intolerance variable, although not explicitly studied by Hofstede, shows that they are indeed related. Also, language, in combination with Masculinity vs Femininity (MAS) (*t-value* $= 2.57$) is more significant than language alone, suggesting its importance in the cultural domain.

The most prominent economic factor is trade market share of total exports – the share of the world's import that is covered by the two countries' exports – with a coefficient of 0.165 – eclipses the direct measure of income groups (at -0.03 not included in the top 12 predictive features), showing trade to be a better indicator of communication than per-capita GDP. Trade agreements are organized over various historic events and through geo-political considerations, thus it is interesting to see them play such an important role in determining every-day online communication. A connection between political climate and communication would be an enticing potential future direction of this research.

Variable	β	**t-value**	**p-value**	**Variable**	β	**t-value**	**p-value**
Trade Market Share	0.165	3.90	***	Gravity Model	0.072	7.17	***
Exports	-0.151	-1.48		Gravity Model x Trade Market Share	-0.067	-4.67	***
Exports x Language	-0.110	-1.45		Trade Market Share	0.165	3.90	***
MAS	-0.102	-2.76	**	Gravity Model x Language	0.060	3.70	***
Gravity Model x Exports	0.098	2.73	**	Intolerance x LTO	0.022	3.66	***
Gravity Model	0.072	7.17	***	Migration x PDI	0.041	3.24	**
Language	-0.070	-1.70	.	Trade Market Share x Exports	0.016	2.95	**
Gravity Model x Trade Market Share	-0.067	-4.67	***	Gravity Model x MAS	0.031	2.93	**
PDI	0.061	1.63		Intolerance x UAI	0.023	2.83	**
Gravity Model x Language	0.060	3.70	***	MAS	-0.102	-2.76	**
Intolerance	-0.054	-2.11	*	Gravity Model x Exports	0.098	2.73	**
Income group x Migration Rate	-0.051	-2.41	*	Language x MAS	0.042	2.57	*

Table 8.4: The top 12 predictive variables in the final model (including interaction factors) ordered by beta coefficients (columns 1-4) and *t-value* (columns 6-9). The gravity model was calculated by using internet penetration as a proxy for population. Significance: *** $p < 0.0001$, ** $p < 0.001$, * $p < 0.01$, . $p < 0.05$.

Finally, the importance of language (here, considered a cultural feature) is mitigated when we add the interactions, with trade and gravity model playing a more important role than language. This is interesting if we refer to recent studies on cooperative work in software [127] where it was found that English is almost always the working language of such communities, even if their mother tongue is not English and hence reducing the importance of common language other than English.

Figure 8.7 compares the model's prediction to communication volume. The figure shows a higher accuracy at high communication volumes with worse performance as the communication decreases. In the next section we discuss these results and look at some of the most difficult to predict cases. Finally we look at practical significance of the findings and its limitations.

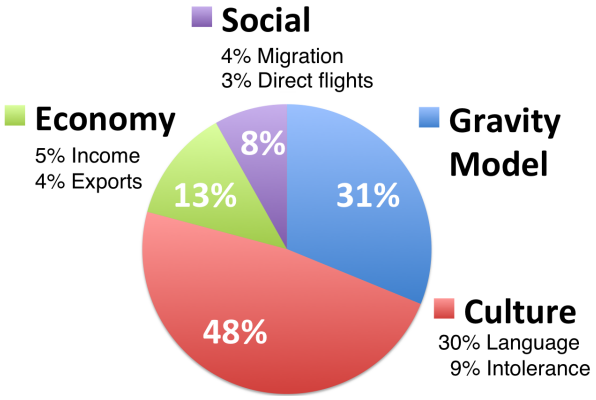


Figure 8.6: The predictive power of the four dimensions with three most important variables. A dimension's weight is computed by summing the absolute values of the coefficients belonging to it.

8.8 Discussion

Is distance dead? We show that no, it is still predictive of international communication in Twitter, but cultural and socio-economic factors, especially that of language, also play an important role. Linguistic and physical separation has been considered a major obstacle in international communication and collaboration. In 2000 Olson *et al.* [105] argued that distance impacts the effectiveness of collaborative work, with language, trust and cultural differences endanger the quality of project results, despite technological enablement of international communication. However, we show that the language barrier is strongest in combination with cultural factors dealing with intolerance and the fear of the unfamiliar. Finding a common culture, thus, may present a way of overcoming language barriers. For instance, a recent study by Takhteyev *et al.* [127] describes successful international collaborative projects in Open Source software development. This takes place, authors argue, when contributors follow an agreed "common" culture and communicate mostly in English. For example, *Lua*, a programming language developed in Brazil and used in the development of several well-known projects such as *World of Warcraft* and *Angry Birds*, was adapted by the global collaborative circles, such that the manuals were in English rather than in Portuguese, fostering widespread international partnerships. To improve collaboration in a culturally-diverse setting, Kittur *et al.* [72] propose several strategies, including observing behavior of other workers,

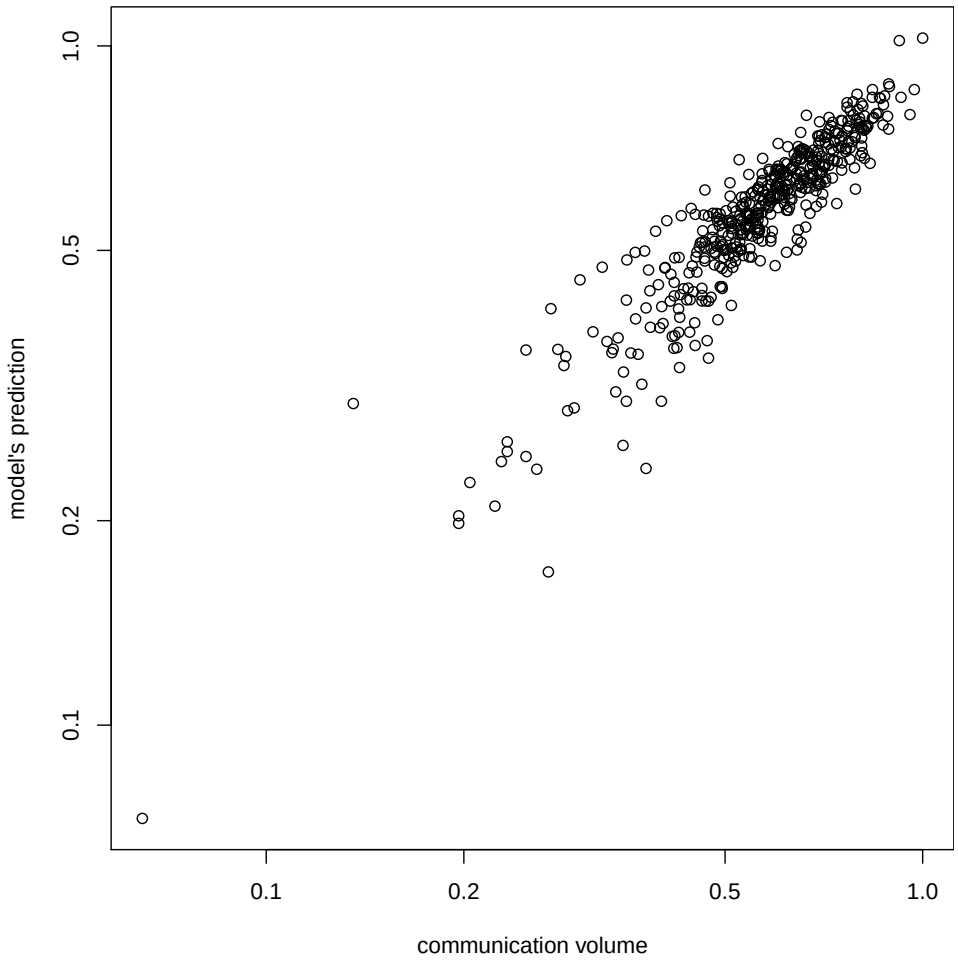


Figure 8.7: Observed unique mention volume versus the model's predictions.

electing leaders, and passing knowledge to others. Figure 8.4 (b) shows that many of the strongest ties lie between countries with different native language, such as United States and Japan, Indonesia and South Korea, Spain and United Kingdom, which, although geographically remote, may be connected by common cultural attributes. Thus, our findings show that finding a common culture could be an important barrier which software designers, in particular those who mean to enable international conversation and collaboration, must overcome.

Similarly, State *et al.* [123], find the concept of *civilization* – countries that share the same religion and continent – having a strong positive effect on the private email exchanges. If one ranks the significant coefficients of their model, one finds Colonial Link and Hungtington's Civilization as the top and third most important predictors, respectively. Language, physical distance, and population size also appear at the top (as second, fourth, and fifth in the ranking). For Twitter communication, we show that even though people can subscribe to the majority of users without authorization or reciprocity (unlike the definition of links in [123]), active interactions (through mentions) are still aligned along culture and physical distance.

Will distance, economic and social constrains impact communication online forever? Or will eventually ubiquitous internet access, new social platforms and globalization open the door for unrestricted communication between countries despite their economic and social differences? The awareness of these boundaries is prerequisite in our understanding of the kinds of information residents of these countries are likely to consume, and of the constraints on the world-wide information propagation.

8.8.1 Practical Implications

Our findings have several implications for the design of social media software. First, we find language and culture to be substantial barriers to Twitter communication. It is noted by Nardi *et al.* [99] that people have to adjust to the technology from other cultures. For instance, Japanese have adapted their writing style to the horizontal typewriter-style word processors, and spelling of words in languages having letters not included in the standard English keyboard has been adjusted accordingly. In the communication network shown in Figure 8.3, the strongest links are with the United States – the country in which Twitter originates – and the culture of which would drive the design of the software. However, Twitter is already making an effort to diversify its service to embrace non-English languages by providing support for a variety of character sets and automatic translation.[6]

One of the major cultural factors we found impacting communication is intolerance, implying that the users of countries associated with higher intolerance would be less likely to communicate with other nations. As discussed by Borning and Muller [17], designers must not assume that some cultural views and values, including those on gender, age, and speech, are universally

[6] https://support.twitter.com/articles/434816-about-the-twitter-translation-center.

held [36]. Some of these values could be learned by gathering user behavior data using built-in tools and interaction logs, combined with the cultural attributes extrapolated from user location information and other personalized data, providing potentially better culture-aware services.

However, using a user's cultural preferences, we may want to instead enrich her experience and broaden the reach of information. Including economic, cultural, and social factors, we plan to enhance the recommendations presented to the user. In particular, we'll consider the task of finding people that will most likely re-tweet posts and reach larger audiences [137] not only locally, but internationally.

8.8.2 Limitations

Although we attempt to reduce the effect of multicollinearity in our model and exclude some of the variables, it is impossible to find a complete, and yet altogether independent set of real-world variables. Further, a different selection of country-specific variables would likely somewhat change the observed results. For example, the importance of trade market share may imply that certain economic and trade policies may also be important, as well as official social policies and visa requirements between countries. Also, due to unavailability of some of the predictive variables, the data set for regression was quite limited compared to original data. The missing values are a source of potential selection bias. This happens because the data extracted from the World Bank and CIA does not distinguish between countries that do not report their trade statistics and country pairs with no bilateral trade (resulting in zero values). A more complete data set would expand the scope and accuracy of such analysis.

Even though Twitter may not be a representative sample of the world's population, our study shows barriers even among the relatively more well-off, technically savvy communities. As a new kind of light-weight, public communication, it is a platform which encourages weak ties between its users. More longitudinal studies are needed to determine whether the development and change of these barriers can be detected in other, more personal or established, means of communication like e-mail and texting.

Finally, we have not considered in our regression inputs that capture how automative or repressive a countries's regime is, nor the imposition of internet censorship.[7]

[7]Twitter was banned in Iran and China before 2011.

We are continuing this line of investigation. In this study, we have unveiled the dynamics of flow between countries, but have not considered directionality and the focus of international attention. Of particular interest would be a deeper study of the language used among residents of different countries as well as the topics discussed in their interactions. By tracking these topics over time, we could detect major shifts in public attention and opinion, especially around crises or other major events.

We are also looking to invert the focus of this research, and attempt to predict socio-economic factors of populated areas, such as city districts, via their online communication. For this, we will attempt to increase the international coverage of our data by including sites like Weibo (an equivalent of Twitter in China) in our data set.

Chapter 9

Conclusions

This thesis main focus was to study the behavior of microblogging users and differentiate their collective behavior across countries. Though this thesis builds on previous works, our work added the use of social sciences theories of culture to this literature.

The path towards understanding user behavior through the lenses of cultural theories started by analyzing user-generated content in Twitter (Part II). First, we conducted an analysis of user-generated broadcast recommendations of who to follow. We explored why users accept and reject these type of recommendations and build a ground-truth dataset of acceptances and rejections (Chapter 4). Next, we used the same dataset to analyze other type of posts (*i.e.*, purely textual, with mentions, with URLs, etc.) and proposed a taxonomy of tweet types. We then clustered active users accordingly finding 5 user types with a certain predominant type of tweets. Moreover, we studied the evolution of these users and how it is related with their popularity (Chapter 5). Finally, we also proposed to analyze the differences and similarities from users of different countries. To target this, we grouped users by country and analyzed how they tweet, the sentiment of their words and the structure of their network of friends.

Based on the insights acquired during the analysis of user behavior patterns in Part III, we identified the need of using social sciences theories to better understand users. For this reason, we presented in Chapter 3 anthropological studies on culture that are suitable for understanding the behavior of users online. Among those, we highlighted Hofstede's dimensions of national culture and Levine's pace of life ranking. We explained how these theories could be used in conjunction with microblogging data. In Chap-

ters 7 and 8 we applied some of these models to explain user behavior and communication.

9.1 Main Results

In this section we answer the research questions drawn up in Chapter 1, based on the findings shown through the chapters of this thesis.

- **Social link recommendations made by current friends have a measurable effect on link formation and the accepted recommendations have more longevity than other links (Chapter 4).** To show that, we calculated the rejections and acceptances of *Follow Friday* recommendations (recommendations of who to follow made by users themselves). Furthermore, for the ties formed after an accepted *Follow Friday* recommendation, we measured the number of weeks (up to 12) they lasted in the receiver's network.

- **As users mature, they evolve to adopt microblogs as a news media rather than a social network (Chapter 5)**. Mature users engage less in conversations, share more links and do not have a predominant type of tweet.

- **The collective behavior of users from some countries standout, based on their special characteristics (Chapter 6).** For example, some countries show to have users considerably more engaged in conversations, with higher reciprocity than others. Others have users with more globalized and hierarchical communities.

- **National culture determine the temporal randomness with which Twitter users post, or the extent to which they mention, follow, recommend and befriend others (Chapter 7)**. We test three main hypotheses associated with power distance, individualism and pace of life and find that activity predictability negatively correlates with pace of life, tweets with mentions negatively correlate with individualism and power imbalance in relationships (based on number of followers).

- **In addition to distance, socio-economic and cultural features also impact international communication (Chapter 8).** We show that by adding socio-economic and cultural features to distance,

the accuracy of the prediction of international communication flow across countries increases significantly.

9.2 Applications

The conclusions of this thesis have several implications for marketing campaigns, community managers and advertisers:

- **Beneficial for marketing campaigns.** We have showed that users do follow the recommendations of their social ties, To make this recommendations possible in Twitter, users use hashtags: in 2011 we analyzed the use of *Follow Friday* hashtags, measure acceptances and rejections, and show that accepted recommendations last longer than other added social links and present different features that could influence acceptances. What is interesting about this phenomena is that these recommendations were created by users themselves.[1] In 2014 people witnessed in a much greater scale a similar phenomena in marketing campaigns of charity organizations such as the Cancer Research UK [2] and the ASL Association. The big difference is that the trending hashtags of these campaigns are related to a drastic increase of donations instead of the number of friends. Similarly to the *Follow Friday* hashtag, the hashtags #nomakeupselfie (for Cancer Research) and #icebucketchallenge or #alsicebucketchallenge (for ASL) were originated by users and became viral thanks to social media but most importantly *because* users directly challenged others to donate. Our studies in Chapters 4 and 5 could be used as a model to study these type of successful campaigns and track down features that lead to successful donations (*i.e.*, using mentions and hashtags).

- **Monitoring international communication**. To monitor how well a social media site is doing in connecting users from different countries, *community managers* could use the features listed in Chapter 8 and observe how well they can predict the international communication flow in the social media sites they are managing.

- **Culture-aware design of social media sites**. In Chapters 6 and 7, we show that culture influences the way people use social media and

[1] http://www.theguardian.com/voluntary-sector-network/2014/aug/20/ice-bucket-challenge-hashtag-charity-macmillan.

[2] http://www.cancerresearchuk.org/.

how users interact with others. Designers can explain these differences through the models presented in Chapter 3. Understanding why users are inclined to certain behavioral patterns (*i.e.*, chat more with others in microblogs, self-promotion, etc.) is crucial to understand the needs of users from different cultures and consequently the success of a social media site. Failing to target these needs and understand the reasons (*i.e.*, cultural dimensions, pace of life, etc.) may lead users to move to other social media platforms.

Moreover, in terms of theoretical implications, the importance of culture should be explored in information diffusion and recommender systems. Explore for example why news from certain cultures/countries are spread faster worldwide than others. Moreover, studying how culture influence the success of a recommendation in social media sites is a research arena that has not been explored in length and that could benefit from this thesis.

9.3 Future Work

There are different types of social media sites. Kaplan and Haenlein [71] suggest that there are six types: collaborative, projects, blogs and microblogs, content communities, social networking sites, virtual games and virtual communities. According to them, Twitter is the representative of the type *blogs and microblogs* and the provider of all the datasets used for our quantitative studies. However, many of the topics covered in this thesis should also be replicated in other types of social media datasets in order to have a better perspective of how user behavior change across platforms. Independently of the social media type, emphasis should be given to cultural differences because it has a major influence on users' opinions and actions. Nowadays, we are facing a rapid increase in Internet and mobile penetration as well as social media engagement of users from all over the world. The Internet is rapidly facing new users from countries different than the ones in North America and Western Europe, where most of the research on social media is done. In particular, the number of mobile broadband subscriptions as well as mobile penetration are significant in Central/Eastern Europe and East/Southest Asia.[3] For this reason, more studies should be done in different platforms to motivate social media sites to be cultural aware in the design of their sites and apps as to better satisfy the demands of users from different nationalities.

[3]Based on the agency "We are Social" (http://wearesocial.net/).

Likewise, another interesting topic explored in this dissertation is related to user-generated recommendations in social media sites. We have seen that users do follow friendship recommendations and that they tend to last longer than social ties added by other reasons. This type of research can be extended to situations where there is a monetary advantage in the recommendations of ideas or products to friends. As pointed out in Section 9.2, we have recently witnessed that the success of charity campaigns like #nomakeupselfie (Cancer Research UK) and #alsicebucketchallenge (ASL Association) were possible thanks to the role of friends in social media when challenging (*i.e.*, recommending) others to donate. What is even more interesting about these phenomena is that these type of recommendations/challenges were originated from users themselves instead of the own charity institutions. There are plenty of room to study this type of phenomena in user-generated recommendations and it would be particularly interesting to analyze it with a cultural perspective. For instance, study if recommendations from friends are followed more in certain countries than others and study the reasons behind the acceptances or rejections (*i.e.*, power distance, individualism, etc.).

Furthermore, with the available data there are also paths for future work in each chapter on top of those already discussed. For example, in Chapter 4, we saw that explicit recommendations are not specifically addressed to a user but rather broadcasted in one's "wall" which causes users to miss many of these potential good recommendations. What if we would collect these recommendations and display them in a list every time the user logs in? We plan to do an offline experiment that captures social explicit broadcasted recommendations and build a recommender system of *who-to-follow* in Twitter derived from selected tweets posted by the target user's network of followees and evaluate our system not only by the acceptance rate but also by the permanence (tenure) of these recommendations in the target user's network. Furthermore, the features that we plan to evaluate differ from previous studies in that we not only focus on social network characteristics and influence but also on the culture of users. This offline study is useful to capture features that make a social recommendation "good" which can be employed in the second phase of this study which will consist of an online recommender.

On the other hand, in Chapter 8, we saw that international communication is affected not only by distance but also by cultural, language and socio-economic barriers. Among those, we observed that language is a high barrier impeding communication. Nevertheless, some social media platforms

have made efforts to motivate multilingual interactions by adding *machine translation* to posts and messages hoping to shatter the barrier. Facebook was the first in implementing it followed by Twitter and Google+.[4] However, the people interacting in these platforms often know each other already, and have a language in common (*i.e.*, friends). But what happens when machine translation is actually used to help multilingual interaction among strangers who perhaps have common interests but not a common language? How often and who pays attention to content outside their immediate reach and why?. Answer to these questions can shed light to the creation of tools that promote cross-cultural and multilingual interaction among strangers and help the world shift to a more integrated place.

Furthermore, Chapter 8 unveiled the dynamics of flow between countries, but we did not considered directionality and the focus of international attention. Of particular interest would be a deeper study of the language used among residents of different countries as well as the topics discussed in their interactions. By tracking these topics over time, we could detect major shifts in public attention and opinion, especially around crises or other major events. Likewise, we are also looking to invert the focus of this research, and attempt to predict socio-economic factors of populated areas, such as city districts, via their online communication. For this, we will attempt to increase the international coverage of our data by including sites like Weibo (an equivalent of Twitter in China) in our data set.

Finally, it is also important to move from big data to qualitative data to analyze culture in a different context. Although social media is a repository with great cultural value, it also comes with a lot of noise that sometimes is hard to detect. Nevertheless, subjects involved in obtrusive studies generally change their behavior and responses because they are aware that they are being watched. This greatly affect the validity of the data gathered during the experimental process [33]. For this reason, employing modern eye tracking methods to study human behavior might be a good solution. Eye tracking is unobtrusive and the eyes are quite easy to observe and their movements may tell us how the brain works [34]. If we were to present a specific visual task to two culturally different group of participants, *would we detect significant differences in their viewing behavior patterns*? If so, *would anthropological studies on culture help to explain why?* As a preliminary experiment, we present a controlled eye/tracking user study in Appendix A to detect differences in attention patterns when participants from Spain and the United

[4] https://blogs.law.harvard.edu/andresmh/category/internet-culture/.

Arab Emirates (U.A.E.) read search engine result pages (SERP). We find that U.A.E. participants stayed on the result pages longer, they read more results and they read each snippet in a more complete way than Spaniards.

Bibliography

At the end of each reference, we indicate the pages where it appears.

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *Proceedings of the 19th Conference on User Modelling, Adaptation and Personalization*, 2011. 29

[2] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship Prediction and Homophily in Social Media. *ACM Trans. Web*, 6, 2012. 29

[3] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012. 33

[4] Afia Akhter Lipi, Yukiko Nakano, and Matthias Rehm. A Parameter-Based Model for Generating Culturally Adaptive Nonverbal Behaviors in Embodied Conversational Agents. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*. Springer Berlin / Heidelberg, 2009. 108

[5] Andreas Auinger, Anna Maria Aistleithner, Harald Kindermann, and Andreas Holzinger. Conformity with User Expectations on the Web: Are There Cultural Differences for Design Principles? In *Design, User Experience, and Usability. Theory, Methods, Tools and Practice - First International Conference, DUXU 2011, Held as Part of HCI International 2011*. Springer, 2011. 155

[6] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the 2011 International Conference on Web Search and Data Mining*, 2011. 29

[7] Ricardo Baeza-Yates, Christian Middleton, and Carlos Castillo. The Geographical Life of Search. In *Web Intelligence*, 2009. 61

[8] Eytan Bakshy, Jake M. man, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 2011. 96

[9] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference companion on World Wide Web*, 2012. 33

[10] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2004. 108

[11] Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 1st edition, 2008. 33

[12] Beevolve. An Exhaustive Study of Twitter Users Across the World. http://www.beevolve.com/Twitter-statistics/, January 2012. 108

[13] Niclas Berggren and Therese Nilsson. Does Economic Freedom Foster Tolerance? Technical report, Research Institute of Industrial Economics, 2012. 119

[14] Sylwia Biatas. Power Distance as a Determinant of Relations Between Managers and Employees in the Enterprises with Foreign Capital. *Journal of Intercultural Management*, November 2009. 96

[15] Johan Bollen, Bruno Goncalves, Guangchen Ruan, and Huina Mao. Happiness is Assortative in Online Social Networks. *Computing Research Repository*, 2011. 13

[16] Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social Network Analysis and Mining for Business Applications. *ACM Trans. Intell. Syst. Technol.*, May 2011. 1, 2

[17] Alan Borning and Michael Muller. Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human factors in computing systems*. ACM, 2012. 127

[18] Margaret M. Bradley and Peter J. Lang. Affective Norms for English words (ANEW): Stimuli, Instruction Manual, and Affective Ratings. In *Technical Report C-1, The Center for Research in Psychophysiology*, 1999. 13, 66

[19] Jean-Francois Brun, Céline Carrère, Patrick Guillaumont, and Jaime De Melo. Has Distance Sied? Evidence from a Panel Gravity Model. *The World Bank Economic Review*, 2005. 105

[20] Frances Cairncross. *The Death of Distance: How the Communications Revolution is Changing our lLves.* Harvard Business Press, 2001. 105

[21] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. In *Proceedings of the 20tt International Conference companion on World Wide Web*, 2011. 67

[22] Twitter Help Center. Posting Links in a Tweet. *https:// support.Twitter.com/ entries/ 78124-how-to-shorten-links-URLs*. 53

[23] Centre for Good Governance. Soft Skills for Public Managers: Handbook of Time Management Skills, 2011. 91

[24] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, May 2010. 12, 13, 17, 30, 56, 96

[25] Patrick Y.K. Chau. Cultural Differences in Diffusion, Adoption, and Infusion of Web 2.0. *Journal of Global Information Management*, 2008. 154

[26] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selçuk Candan, Lexing Xie, and Aisling Kelliher. How Does the Data Sampling Strategy Impact the Discovery of Information Difussion in Social Media? In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010. 13

[27] Jeffrey M. Conte and Tracey E. Rizzuto. A Construct-oriented Analysis of Individual-level Polychronicity. *Journal of Managerial Psychology*, 1999. 91

[28] Anthony Cox and Maryanne Fisher. An Expectation-Based Model of Web Search Behavior. In *Proceedings of the 2th International Conferences on Advances in Computer-Human Interactions (ACHI'09)*, 2009. 155

[29] Pamela L. Cox, Barry A. Friedman, and Thomas Tribunella. Relationships among Cultural Dimensions, National Gross Domestic Product,

and Environmental Sustainability. *National Business and Economics Society*, 2011. 23

[30] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 1st edition, 2009. 42

[31] Peter Dodds and Christopher Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, August 2010. 18, 66

[32] Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLOS ONE*, 2011. 13, 18, 66, 67

[33] Sergio Raúl Duarte Torres. *Information Retrieval for Children: Search Behavior and Solutions*. PhD thesis, University of Twente, February 2014. 136

[34] Andrew Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer, 2007. 136

[35] Patricia M Duronto and Shin'ichi Nakayama. Japanese Communication. Avoidance, anxiety, and uncertainty during initial encounters. *Journal of East Asian Affairs*, 2005. 123

[36] Charles Ess. and Fay Sudweeks. On the edge: Cultural barriers and catalysts to IT diffusion among remote and marginalized communities. *New Media & Society*, 2001. 128

[37] Gabrielle Ford and Paula Kotzé. Designing Usable Interfaces with Cultural Dimensions. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction(INTERACT'05)*, 2005. 154

[38] National Science Foundation. *Rebuilding the Mosaic*. National Science Foundation, 2011. 102

[39] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*, 2012. 11

[40] Sandra Garcia Esparza, Michael P. O'Mahony, and Barry Smyth. On the real-time web as a source of recommendation knowledge. In *Proceedings of the 4th ACM Conference on Recommender Systems*, 2010. 29

[41] Ruth García-Gavilanes, Marcelo Mendoza, Barbara Poblete, and Ale-

jandro Jaimes. Microblogging without Borders: Differences and Similarities. In *Proceedings of the 2011 ACM Web Science Conference (WebSci'11)*, 2011. 60

[42] Ruth García-Gavilanes, Neil O'Hare, Luca Maria Aiello, and Alejandro Jaimes. Follow My Friends This Friday! An Analysis of Human-Generated Friendship Recommendations. In *The 5th International Conference on Social Informatics (SOCINFO)*, 2013. 28, 46

[43] Ruth García-Gavilanes, Daniele Quercia, and Alejandro Jaimes. Cultural Dimensions in Twitter: Time, Individualism and Power. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013. 88

[44] Ruth García-Gavilanes, Andreas Kaltenbrunner, Diego Sáez-Trumper, Ricardo Baeza-Yates, Pablo Aragón, and David Laniado. Who are my Audiences? A Study of the Evolution of Target Audiences in Microblogs. In *Proceedings of the 6th International Conference on Social Informatics (SocInfo 2014)*, 2014. 46

[45] Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. Twitter Ain'T Without Frontiers: Economic, Social, and Cultural Boundaries in International Communication. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'13)*, New York, NY, USA, 2014. 107

[46] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006. 122

[47] Jan E. Gewehr, Martin Szugat, and Ralf Zimmer. BioWeka - extending the Weka framework for bioinformatics. *Bioinformatics/computer Applications in The Biosciences*, 2007. 42

[48] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human factors in computing systems*, 2009. 122

[49] Michele Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 2002. 73

[50] Jennifer Golbeck. *Analyzing the Social Web*. Elsevier Science, 2013. ISBN 9780124058569. 1

[51] Scott A. Golder and Michael W. Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 2011. 61, 91

[52] Mark S. Granovetter. The Strength of Weak Ties. *American journal*

of sociology, 1973. 106

[53] Edward T. Hall. *Beyond Culture*. Anchor Books, New York, NY, 2nd ed edition, 1977. 60

[54] Edward T. Hall and Mildred Reed Hall. *Hidden Differences: Doing Business with the Japanese*. Anchor Press/Doubleday, 1987. 94

[55] Edward T. Hall and Mildred Reed Hall. *Understanding Cultural Differences*. Intercultural Press, 1990. 16, 22, 91, 95

[56] Robert Hanneman and Mark Riddle. *Introduction to social network methods*. University of California Riverside, CA, 2005. 77

[57] John Hannon, Mike Bennett, and Barry Smyth. Recommending Twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the 4th ACM Conference on Recommender Systems*, 2010. 29

[58] Nadav Hochman and Raz Schwarts. Visualizing Instagram: Tracing Cultural Visual Rhythms. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012. 61

[59] Geert Hofstede. *Culture's Consequences, International Differences in Work-Related Values*. Sage Publications, Inc, 1980. 105, 123

[60] Geert Hofstede. *Culture's Consequences : Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications, Thousand Oaks, California, 2nd ed edition, 2001. 60

[61] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and Organizations: Software of the Mind*. McGraw-Hill, 2010. 15, 16, 17, 18, 85, 95, 96, 98, 101, 159

[62] Lichan Hong, Gregorio Convertino, and Ed H. Chi. Language Matters In Twitter: A Large Scale Study. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 13

[63] Lichan Hong, Gregorio Convertino, and Ed H. Chi. Language Matters In Twitter: A Large Scale Study. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 96

[64] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational Tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, 2010. 46

[65] Samuel P. Huntington. The Clash of Civilizations. *Foreign Affairs*, 1993. 16, 105, 108

[66] William K. Hutchinson. Does Ease of Communication Increase Trade? Commonality of Language and Bilateral Trade. Technical report, Van-

derbilt, University Department of Economics, 2002. 115

[67] C.J. Hutto, Sarita Yardi, and Eric Gilbert. A Longitudinal Study of Follow Predictors on Twitter. In *Proceedings of the SIGCHI Conference on Human factors in computing systems*, 2013. 29

[68] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th ACM WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 2007. 12

[69] Woo-Sung Jung, Fengzhong Wang, and H. Eugene Stanley. Gravity model in the Korean highway. *EPL (Europhysics Letters)*, 2008. 108

[70] Jinen Kamdar. Keep up with conversations on Twitter. *https://blog.Twitter.com/2013/keep-up-with-conversations-on-twitter*. 54

[71] Andreas M. Kaplan and Michael Haenlein. Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons*, 53, 2010. 134

[72] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 2013. 125

[73] Christoph Carl Kling and Thomas Gottron. Detecting culture in coordinates: cultural areas in social media. In *International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web (DETECT)*. ACM, 2011. 61

[74] Anett Kralisch and Bettina Berendt. Cultural Determinants of Search Behaviour on Websites. In *The 6th International Workshop on Internationalisation of Products and Systems*, Vancouver, Canada, 2004. 153

[75] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D. Blondel. Urban Gravity: a Model for Inter-City Telecommunication Flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009. 106, 108

[76] Katherine Krumme. *How Predictable: Patterns of Human Economic Behavior in the Wild*. Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2010. 91

[77] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of*

the 19th international conference on WWW, Raleigh NC USA. ACM, 2010. 12, 30, 45, 58, 96, 102

[78] Haewoon Kwak, Hyunwoo Chun, and Sue Moon. Fragile Online Relationship: a First Look at Unfollow Dynamics in Twitter. In *Proceedings of the SIGCHI Conference on Human factors in computing systems*, 2011. 29, 36

[79] Debra Lauterbach, Hung Truong, Tanuj Shah, and Lada Adamic. Surfing a Web of Trust: Reputation and Reciprocity on CouchSurfing.com. In *The 12th IEEE International Conference on Computational Science and Engineering (CSE'09)*, 2009. 107

[80] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, *et al.* Computational Social Science. *Science*, February 2009. 59, 102

[81] Kyumin Lee, Brian Eoff, and James Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 30, 89

[82] Jure Leskovec and Eric Horvitz. Planetary-scale Views on a Large Instant-messaging Network. In *Proceedings of the 17th International Conference companion on World Wide Web*, 2008. 105, 107

[83] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic Evolution of Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, 2008. 36

[84] Robert Levine. *A Geography of Time: The Temporal Misadventures of a Social Psychologist or How Every Culture Keeps Time Just a Little Bit Differently.* University Press, 2006. 16, 19, 20, 85, 86, 89, 94

[85] David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks. In *Proceedings of the 2002 ACM Conference on Information and Knowledge Management (CIKM 2003)*, 2003. 29

[86] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The Tweets They are a-changin': Evolution of Twitter Users and Behavior. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014. 13, 55

[87] Linyuan Lu and Tao Zhou. Link Prediction in Complex Networks: A Survey. *Physica A*, 2011. 29

[88] Linyuan Lu, Ci-Hang Jin, and Tao Zhou. Effective and Efficient Similarity Index for Link Prediction of Complex Networks. *CoRR*, 2009.

29

[89] Zhunchen Luo, Miles Osborne, Sasa Petrovic, and Ting Wang. Improving Twitter Retrieval by Exploiting Structural Information. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, JToronto*, 2012. 46

[90] Thomas Mandl. Comparing Chinese and German blogs. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT'09)*, 2009. 96

[91] Mari-Carmen Marcos, Ruth García-Gavilanes, Emad Bataineh, and Lara Pasarin. Using Eye Tracking to Identify Cultural Differences in Information Seeking Behavior. In *Oral presentation at the Workshop* Many People Many Eyes *as part of the ACM CHI 2013*, 2013. 154

[92] Alice Marwick and Danah Boyd. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media and Society*, September 2010. 46, 48

[93] Patrick Philippe Meier. *Do "Liberation Technologies" Change the Balance of Power Between Repressive States and Civil Society?* PhD thesis, The Fletcher School of Law and Diplomacy, 2011. 115

[94] Stanley Milgram. The Experience of Living in Cities. *Science*, 1970. 19

[95] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07)*, 2007. 12, 71

[96] Diana Mok, Barry Wellman, and Juan Carrasco. Does Distance Matter in the Age of The Internet? *Urban Studies*, 2010. 105

[97] Miranda Mowbray. The Twittering Machine. In *The 6th International Conference on Web Information Systems and Technologies (WEBIST'10)*, 2010. 98

[98] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it Really About Me?: Message Content in Social Awareness Streams. In *Proceedings of the 13th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'10)*, 2010. 13, 46

[99] Bonnie A. Nardi, Ravi K. Vatrapu, and Torkil Clemmensen. Comparative informatics. *Interactions*, 18, 2011. URL http://dblp.uni-trier.de/db/journals/interactions/interactions18.html#NardiVC11. 127

[100] Andrew Ng. Machine Learning . *Available at https://www.coursera.org*, 2014. 50

[101] J. Nielsen and K. Pernice. *Eyetracking Web Usability*. Voices That Matter. New Riders, 2010. 153

[102] Richard E. Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. Culture and Systems of Thought: Holistic Versus Analytic Cognition. *Psychological Review New York*, 2001. 154, 158

[103] Jung-Min Oh and Nammee Moon. A Cultural Dimensions Model based on Smart Phone Applications. *Journal of Information Processing Systems*, 2011. 88

[104] Jung-Min Oh and NamMee Moon. Towards a cultural user interface generation principles. *Multimedia Tools and Applications*, 2012. 88

[105] Gary M. Olson and Judith S. Olson. Distance matters. *Human-Computer Interaction*, 2000. 107, 125

[106] Jukka-Pekka Onnela, Samuel Arbesman, Albert-László Barabási, and Nicholas A. Christakis. Geographic Constraints on Social Network Groups. *PLoS ONE*, 2011. 12

[107] Zizi Papacharissi. The presentation of self in virtual life: Characteristics of personal home page. *Journalism and Mass Communication Quarterly*, 2002. 46

[108] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review Psychology*, June 2003. 103

[109] Saša Petrović, Miles Osborne, and Victor Lavrenko. RT to win! Predicting message propagation in Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 27

[110] Barbara Poblete, Ruth García-Gavilanes, Marcelo Mendoza, and Alejandro Jaimes. Do All Birds Tweet the Same? Characterizing Twitter Around the World. In *Proceedings of the 2011 ACM Conference on Information and Knowledge Management (CIKM 2011)*, 2011. 60

[111] Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesa na. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 2007. 66

[112] Katharina Reinecke, Minh Khoa Nguyen, Abraham Bernstein, Michael Naf, and Krzysztof Z. Gajos. Doodle Around the World: Online Scheduling Behavior Reflects Cultural Differences in Time Perception and Group Decision-Making. In *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'13)*, February 23-27 2013. 22, 88, 91

[113] Juan J. Rodríguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rota-

tion Forest: A New Classifer Ensemble Method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(10), 2006. 42

[114] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: geo-social metrics for online social networks. In *The 3rd Workshop on Online social networks (WOSN'10)*, 2010. 109

[115] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 107

[116] Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in Folksonomies: social link prediction from shared metadata. In *Proceedings of the 2010 International Conference on Web Search and Data Mining*, 2010. 33

[117] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*, 40, 2011. 33

[118] Sailing Silang. Snapshot of Indonesia Social Media Users, 2011. 97

[119] Barry Smith, Maurice Coyle, and Peter Briggs. *Recommender Systems Handbook*, chapter Communities, Collaboration, and Recommender Systems in Personalized Web Search. Springer, 2010. 1

[120] Chris Smith, Daniele Quercia, and Licia Capra. Anti-gravity Underground? In *2nd Workshop on Pervasive Urban Applications (PURBA), in conjunction with Pervasive*, June 2012. 112

[121] Chris Smith, Daniele Quercia, and Licia Capra. Finger on the pulse: identifying deprivation using transit flow analysis. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013. 108

[122] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 2010. 91

[123] Bogdan State, Patrick Park, Ingmar Weber, Yelena Mejova, and Michael Macy. The Mesh of Civilizations and International Email Flows. *CoRR*, 2013. 105, 106, 108, 109, 127

[124] Piers Steel and Vasyl Taras. Culture as a consequence: A multi-level multivariate meta-analysis of the effects of individual and country characteristics on work-related cultural values. *Journal of International Management*, September 2010. 23

[125] Moritz Sudhof. Politics, Twitter, and information discovery: Using

content and link structures to cluster users based on issue framing. *stanford.edu*, 2012. 40

[126] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *The 2nd IEEE International Conference on Social Computing*, 2010. 27

[127] Yuri Takhteyev. *Coding Places: Software Practice in a South American City*. The MIT Press, 2012. 107, 124, 125

[128] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 2012. 105, 107, 109, 118

[129] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended Accounts in Retrospect: an Analysis of Twitter Spam. In *ACM Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, New York, NY, USA, 2011. 30

[130] Tapanee Tirapat and Tiranee Achalakul. Usability Assessment for Hyperlink Methods. *JCIT*, 2007. 155

[131] Amy Todd. From Polychronicity to Multitasking: The Warping of Time Across Disciplinary Boundaries. *Anthropology of Work Review*, 2009. 22

[132] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 1969. 107

[133] The Pennsylvania State University. STAT 501 - Regression Methods, 2013. URL https://onlinecourses.science.psu.edu/stat501/. 121

[134] Mukundan. Venkataraman, K. P. Subbalakshmi, and Rajarathnam Chandramouli. Measuring and Quantifying the Silent Majority on the Internet. In *The 35th IEEE Sarnoff Symposium*, 2012. 108

[135] Elena Vitkauskaite. Cultural Adaptation Issues in Social Networking Sites. *Economics and Management: 2011.16 ISSN 1822-6515*, 2011. 1

[136] Alexander Vuylsteke, Zhong Wen, Bart Baesens, and Jonas Poelmans. Consumers online information search: a cross-cultural study between China and Western Europe. In *Proceedings of the Academic And Business Research Institute Conference*, 2009. 155

[137] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22nd International Conference companion on World Wide Web*, 2013. 128

[138] Xianwen Wang, Shenmeng Xu, Lian Peng, Zhi Wang, Chuanli Wang, Chunbo Zhang, and Xianbing Wang. Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics*, 2012. 61

[139] Duncan J. Watts. The "New" Science of Networks. *Annu. Rev. Sociol.*, 30:243–270, 2004. 2

[140] R. Wilkinson and K. Pickett. *The Spirit Level: Why Equality is Better for Everyone*. Penguin Books, 2010. 98, 101, 115

[141] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition, 2011. 1

[142] Warren Woods. *Tick Tock!...Who Broke the Clock?* Innovations International, 2003. 22, 86

[143] S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts. Who says what to whom on Twitter. In *Proceedings of the 20tt International Conference companion on World Wide Web*, 2011. 45, 52

[144] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. The Pace of Pedestrian Flows in Cities. *Environment and Behavior*, 1989. 19

[145] Keiji Yanai, Keita Yaegashi, and Bingyu Qiu. Detecting Cultural Differences using Consumer-Generated Geotagged Photos. In *2nd International Workshop on Location and the Web (LOCWEB) in CHI*, 2009. 61

[146] Ping Zhang and Gisela von Dran. User Expectations and Rankings of Quality Factors in Different Web Site Domains. *International Journal of Electronic Commerce*, 2001. 155

[147] Tao Zhou, Linyuan Lu, and Yi-Cheng Zhang. Predicting Missing Links Via Local Information. In *European Physical Journal B*, volume 71(4) of *Special Issue: The Physics Approach to Risk: Agent-Based Models and Networks.*, pages 623–630, 2009. 29

[148] George Kingsley Zipf. The P 1 P 2/D hypothesis: on the intercity movement of persons. *American sociological review*, 11, 1946. 112

Appendix A

Using Eye Tracking to Identify Cultural Differences in Information Seeking Behavior

A.1 Introduction

The Internet has revolutionized the way people live, work, study, shop, communicate and do business. Search engines are considered the main entrance to the web because it allows to find pages of interest according to queries. Nevertheless, little attention has been given on the presentation of the search results pages based on cultural differences. Nowadays, the interface looks the same for all users regardless of their location but previous research [74] argue that there is a correlation between culturally determined thinking patterns and search behavior.

Researchers have been particularly interested on the visitor's visual behavior with search engines using eye tracking machines. An eye tracking machine is a device used to monitor and record users' visual search patterns on a screen content. It has also been approved as a reliable tool used extensively in several usability studies [101]. The reason is that eye tracking machines unobtrusively follows a readers' eye movements and gives the most likely locations of where a person has looked and stop reading. The equipment allows to plot a *heat map* that highlights the areas where readers looked the

most. This information is useful to determine where the data such as text, images and adds should be placed.

The primary goal of this research is to investigate how people with different cultural background differ in their interaction style and visual behavior on search engine results pages (SERP), more specifically between groups from the Arabian Peninsula vs. Western Europe. The researchers conducted a controlled eye-tracking experiment to explore and evaluate the visual behavior of U.A.E and Spaniard users when scanning through the first page of the search results in Google. The research aim is to examine if culture influences the behavior of these two groups in the way they evaluate the list of search results to choose a link. It is expect to find some differences in reading patterns, number of search results considered, browsing time on SERP until a result is clicked (dwell time) and success rate for correct answers to the questions. In future work, the researchers will analyze how special elements like ads, multimedia results and rich elements will attract and impact user's visual attention.

The appendix is organized as follows: first we present related work, then we discuss the research questions, methodology as well as the eye tracking experiment design. The subsequent sections present the results, data analysis and discussions. The last section contains the conclusion which includes a summary, implications and importance of the study, limitations and suggestions for future research.

This Appendix is based on [91].

A.2 Related Work

It is not a clear how Hofstede's cultural dimensions (explained in Chapter 3) help to understand the differences of how people use the web; however, several methods have been adapted to specific situations. For example, Chau *et al.* [25] argued that individualism and collectivism are particularly relevant in studying the use of services built around Web 2.0, including OSN. Likewise, Ford *et al.* [37] discussed how to accommodate one of the five cultural dimensions in user interfaces to increase usability.

On the other hand, recent studies have concluded that Western learners tend to have more analytical cognitive learning style whereas East Asians tend to have more holistic or contextual learning style [102]. Hoffstede cultural model is used in the study to explain the behavioral cognitive and perceptual differences between the two culturally diverse user groups.

In a cross-cultural study conducted by [136] on consumers' information search behavior, they found significant differences between Chinese and Western Europeans in their online search behavior with respect to frequency, goal, types of information sought, websites selected as well as users' usage patterns. Nevertheless, this study is based on *interviews and questionnaires* that can cause biases on the answers.

Several research studies examined consistency of user expectations for major web and user interface elements (such as navigation tools, hyperlinks and colors, logs, search box and others) placement on websites [28; 130; 146]. Other researchers took this common issue a bit further and compare expectations with different users' groups as well. For example, Auinger *et al.* on his work related to user expectations on the web [5] investigated the validity of four web elements design principles using eye tracking data for European vs. Anglo-American users. The findings from their study suggest there are cultural differences regarding the web elements design.

We are not aware of many unobtrusive tests made on visual search behavior on SERPs, specially not studies involving cross-cultural comparisons between middle easters and europeans.

A.3 Methodology

In total, 117 people participated in the test: 60 people in Barcelona (Spain, Western culture) and 57 in Dubai (U.A.E., Arab culture). From these groups, 63% of participants are women and 80% are between 18 and 40 years old. The tests were administered at the respective labs at University Pompeu Fabra in Barcelona and Zayed University in Dubai.

The researchers prepared 12 SERPS with 3 versions of each: first one with ads, second with enriched snippets like images, maps, etc., and third with no adds or enriched snippets. The test intended to cover all kind of elements that search engines usually include in SERPs. Our SERPs were divided in 3 general topics including architecture, ports and tourism. The search results to queries were also made related to each topic.

In the test, participants were asked to answer 12 questions (4 for each topic). For each query, participants were presented with a question and its corresponding SERP. We controlled that all participants have the same SERPs for each specific query. Second, participants were instructed to click on the result they thought was the most appropriate for the query. Finally, they were asked to choose an answer to the query from a list of 4 options: *a wrong*

(a) Spaniard participants (b) U.A.E participants

Figure A.1: Heat map showing vertical reading patterns of Spaniard (a) and Arab (b) participants.

answer, *a right answer*, *I don't know* and *none of the above*. The answer to the query is visually embedded in the results presented in the SERP.

The eye-tracking equipment used in the study were Tobii 1750 in Spain and Tobii T−120 in U.A.E. The Tobii Studio software version 2.3 was used for the data analysis. The metrics obtained per country so far were the following:

1. *Hetmaps for reading patterns*: heat maps indicates the time that the users spent on each result, and a reading pattern can be obtained from them.
2. *Number of results read by the users.*
3. *Time to Fist Click (TFC)*: indicates the dwell time of the users on the list of search results until deciding which result to click. questions.
4. *Success rates*: percentage of correct answers of the total

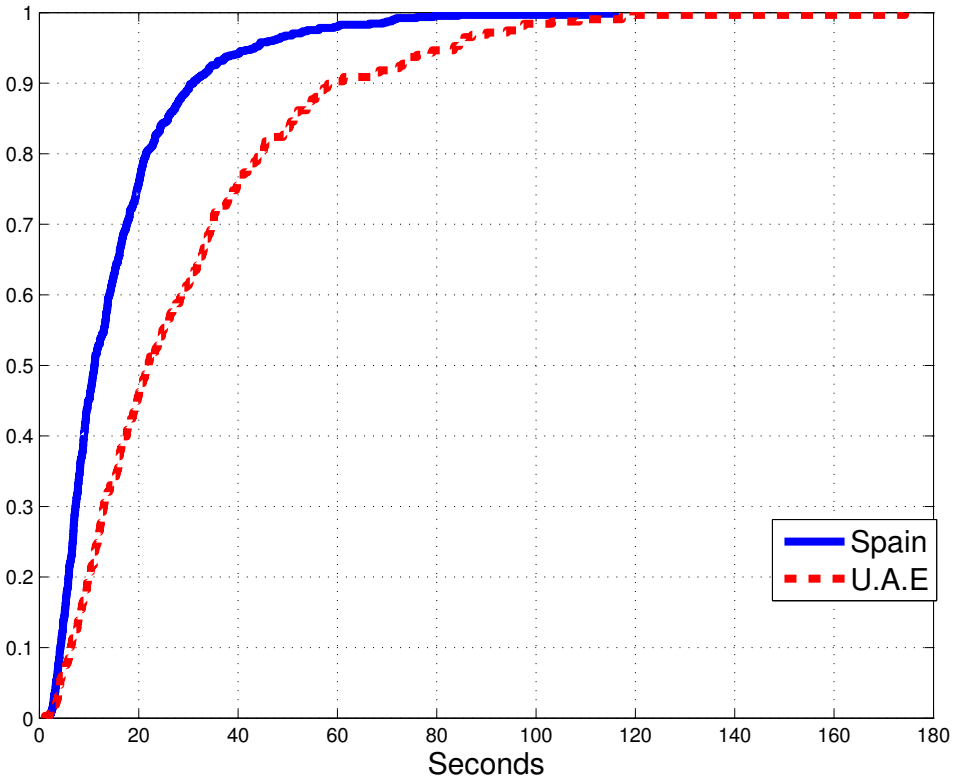


Figure A.2: CDF of dwell time on SERPs of participants from U.A.E and Spain.

A.4 Results

The 4 metrics were analyzed to show if there were significant differences between both countries:

1. *Reading patterns by country*: while a scanning vertical pattern can be observed in Spain (see figure A.1a),), a clear horizontal one is given by U.A.E. participants (see figure A.1b).

2. *Dwell time on SERPs*: Spain users spent much less time on SERPs than U.A.E. participants, who prefer to read more before taking a decision. On average, Spaniards took 39.26 seconds per page and Arabs took 62.99 seconds. We used Mann Whitney approach to verify

that there was actually a statistical difference between the two groups. Figure A.2 shows that 50% of participants in Spain took less than 10 seconds per page before choosing an answer in contrast to 20 seconds for the other group.

3. *Number of scanned results by country*: accordingly to the previous results, Spain participants read fewer results than U.A.E. users. In particular Spaniards read mostly top ranked results while Arab users considered bottom results as well before clicking (figure A.3a and A.3b).

4. *Success rate*: Surprisingly, we shows that more successful answers were found in Spain tests (50% chose the correct answers in Spain, vs. 40% in U.A.E) but the percentage of wrong answers is similar (43% chose the wrong answers in Spain, 39% in U.A.E). The difference is due to the none of the above option (Spain, 2% vs. U.A.E. 12%) and I don't know option (Spain 3% vs. U.A.E. 9%). This implies that Arab people from our test preferred to choose one of these options (I don't know or none of the above) more than Spaniards who preferred to risk for a right answer more often. Table A.1 shows the success rates for U.A.E and Spain.

A.5 Discussion

Significant differences were found in the 4 aspects covered by the study: U.A.E. participants stayed on the SERPs for longer, they read more results and viewed each snippet in more detailed way than Spaniards. In Spain, people tended to scan the SERP, reading less text on each snippet, and choose a result among the first top ranked ones without hardly seeing those in bottom positions. Further work is necessary to determine the actual reasons for these differences. Based on the current results, we explain the results by the following factors:

1. Cultural aspects:

    a) The results seem to be coincided with [102] theories about holistic cognition of Eastern cultures versus the analytic style in Western cultures. They compared East Asian and American, while we observed people from an Arab culture and a Mediterranean one, so Nisbett's work is not necessarily applied to this study.

(a) Spaniard participants read few results (b) U.A.E participants read many results

Figure A.3: Heat map showing the amount of scanned results.

b) The presence of the moderator may have intimidated people at Zayed University. According to [61], Arab countries are ranked between the 12-14 place in power distance while Spain is ranked between 45-46 (lower power distance). A typical behavior of a large power distance countries is the high respect to teachers even outside the classroom. It is believed that this motivated Dubai users to spend more time searching for the answers in SERP.

2. Language skills: Spaniards viewed SERPs in their native language, while U.A.E. users saw them in English. Although U.A.E. participants have a good level of English skills due to the fact that English is the language of instruction at Zayed University, being a non-native speakers could have caused a lack of self-confidence at the user side which resulted in a low performance doing the tasks.

For now, we have preliminary results that should be validated and compared with future experiments considering:

	Spain	U.A.E
Right	50%	40%
Wrong	43%	39%
None	2%	12%
Don't know	3%	9%

Table A.1: Success rate between Spaniards and Emiratis.

- Run an Arabic version of the test in U.A.E. and English test in Spain
- Include more countries and cultures to the study
- Perform a detailed analysis filtered by age and gender, comparing organic results to ads with different types of results (multimedia, site links, social recommendation, etc)
- Add more questions with new topics
- Allow users to type their own queries and see the clicked results

This is the first study on how cultural background can affect the users' visual and cognitive behavior on information seeking in search engines environment. We consider this as an interesting research topic for both Human-Computer Interaction and Information Retrieval communities.