Finding Relevant People in Online Social Networks

Diego Sáez-Trumper

TESI DOCTORAL UPF / 2013

Director de la tesi

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

A mis padres

Acknowledgements

First of all I want to thank my advisor, Ricardo Baeza-Yates, for giving me the opportunity to work with him. I cannot imagine a better supervisor.

I also want to thank Virgílio Almeida who hosted me at UFMG in Brazil. I learned a lot from him and from his amazing research group, specially I want to thank my co-authors Giovanni Comarela and Gabriel Magno. I also want to thank Emanuel Vianna and Gustavo Rauber and in their name to all the students at the CAMPS laboratory. In addition, I want to thank Fabricio Benevenuto, Loic Cerf, Wagner Meira, Gisele Pappa, Adriano Veloso and Nivio Ziviani.

I have been lucky to have a second internship at Cambridge University, and I want to thank Jon Crowcroft and Daniele Quercia for hosting me there. I also want to thank all the people at the NetOS group, it was an amazing experience to do research at The Computer Laboratory.

I also had the pleasure to collaborate with many people around the world. Here, I want to highlight Meeyoung Cha from KAIST, and also thank my other co-authors and people that I have collaborated with: Mounia Lalmas, Alan Mislove, Balachander Krishnamurthy, and Yabing Liu. Additionally, I want to thank Carlos Castillo for his valuable mentoring during this thesis.

From UPF, I want to thank all the people at the Web Research Group, especially to those that I had directly collaborated: Luca Chiarandini, Eduardo Graells, Mari-Carmen Marcos, David Nettleton, Luz Rello, and Michele Trevisol. I also want to thank Miquel Oliver and the NETS group. Special thanks to Lydia Garcia and all the people from the department's secretary for the support they give to me and to all

the PhD students, making our life easier. I also want to thank my friend and college Susan Ferreira, for the hundreds of hours of chat and thousands of liters of coffee shared during these years.

These years in Barcelona have been amazing, and I want to thank all my friends and flat mates, that have been my "second family" during this period. The list is long, but I want to represent them in Daniela Mardones who helped me a lot when I arrived to this wonderful city; and to my friend Luis Cárcamo for his advices. I also want to thank my friend Xavier Codina who helped to understand the authentic Catalonian spirit.

There are no enough words to express my gratitude to Bianca, she has supported me in everything during these years. She has traveled with me around the world, adapting to new situations, places, and people to start again and again. She has taught me valuable life lessons that I would never learned from the books. *Muito obrigado.*

I cannot finish without to thank my sister Mariel, she has helped me in my education since I was a kid, and she continue to do until today.

Finally, I want to thank my parents, this thesis is completely dedicated to them. They have always believed in me, supported me in every imaginable way, and encouraged me to continue studying. I would never got a PhD without their support, they deserve all the honors. Beyond the academic titles and professional success, my only goal in life is to be as good people as they are.

Summary

The objective of this thesis is to develop novel techniques to find relevant people in Online Social Networks (OSN). To that end, we consider different notions of relevance, taking the point of view of the OSN providers (like Facebook) and advertisers, as well as considering the people who are trying to push new ideas and topics on the network. We go beyond people's popularity, showing that the users with a lot of *followers* are not necessarily the most relevant. Specifically, we develop three algorithms that allow to: *(i)* compute the monetary value that each user produces for OSN provider; *(ii)* find users that push new ideas and create trends; and *(iii)* a recommender system that allows advertisers (focusing in local shops, like restaurants or pubs) to find potential customers. Furthermore, we also provide useful insights about users' behavior according to their relevance and popularity, showing - among other things - that most active users are usually more relevant than the popular ones. Moreover, we show that usually very popular users arrive late to the new trends, and that there are less popular, but very active users that generate value and push new ideas in the network.

Resum

L'objectiu d'aquesta tesi és desenvolupar noves tècniques per trobar persones rellevants en les Xarxes Socials a Internet. Així doncs, considerem diferents nocions de rellevància, tenint en compte el punt de vista dels proveïdors del servei (com Facebook) i dels anunciants, però també de persones que intenten proposar noves idees i temes a la xarxa. La nostra investigació va més enllà de la popularitat de les persones, mostra que els usuaris amb molts *seguidors* no són necessàriament els més rellevants. Específicament, desenvolupem tres algorismes que permeten: *(i)* calcular el valor (monetari) que cada usuari produeix per al proveïdor del servei; *(ii)* trobar usuaris que proposen noves idees i creen tendències; i *(iii)* un sistema de recomanació que permet als anunciants (centrant-nos en botigues locals, com ara un restaurant o un pub) trobar clients potencials. Addicionalment, lliurem informació útil sobre el comportament dels usuaris segons la seva rellevància i popularitat, mostrant, entre altres coses, que els usuaris més actius solen ser més rellevants que els populars. A més a més, mostrem que normalment els usuaris molt populars arriben tard a les noves tendències, mentre que usuaris de menor popularitat, però molt actius, generen valor i fomenten noves idees a la xarxa .

Resumen

El objetivo de esta tesis es desarrollar nuevas técnicas para encontrar personas relevantes en las Redes Sociales en Internet. Para ello, consideramos diferentes nociones de relevancia, tomando el punto de vista de los proveedores del servicio (como Facebook) y de los anunciantes, pero también de las personas que intentan proponer nuevas ideas y temas en la red. Nuestra investigación va más allá de la popularidad de las personas, mostrando que los usuarios con muchos *seguidores* no son necesariamente los más relevantes. Espeficamente, desarollamos tres algoritmos que permiten: *(i)* calcular el valor (monetario) que cada usuario produce para el proveedor del servicio; *(ii)* encontrar usuarios que proponen nuevas ideas y crean tendencias; y *(iii)* un sistema de recomendación que permite a los anunciantes (centrándonos en tiendas locales, tales como un restaurant o un pub) encontrar potenciales clientes. Adicionalmente, proporcionamos información útil sobre el comportamiento de los usuarios según su relevancia y popularidad, mostrando - entre otras cosas - que los usuarios más activos suelen ser más relevantes que los populares. Más aún, mostramos que normalmente los usuarios muy populares llegan tarde a las nuevas tendencias, y que existen usuarios menos populares, pero muy activos que generan valor y fomentan nuevas ideas en la red.

Resumo

O objetivo desta tese é desenvolver novas técnicas para encontrar pessoas relevantes nas Redes Sociais na Internet. Para este fim, levamos em consideração diversos conceitos de relevância, nos colocando sob o ponto de vista dos prestadores de serviço (como Facebook) e dos anunciantes, assim como de pessoas que tentam difundir novas ideias na rede. Além da popularidade das pessoas, a nossa pesquisa revela que os usuários com muitos seguidores não são necessariamente os mais relevantes. Desenvolvemos três algoritmos que nos permitem: *(i)* calcular o valor (monetário) que cada usuário rende para o prestador de serviços; (ii) encontrar usuários que propõe novas ideias e criam tendências; e *(iii)* um sistema de recomendação que permite aos anunciantes (focando em lojas locais, como um bar ou restaurante) encontrar clientes em potencial. Além disso, proporcionamos informação útil sobre o comportamento dos usuários de acordo com a sua relevância e popularidade, mostrando - entre outras cosas - que os usuários mais ativos costumam ser mais relevantes que os mais populares. Além de que, mostramos que normalmente os usuários muito populares chegam tarde às novas tendências, e que existem usuários menos populares porém muito ativos que geram valor e fomentam novas ideias na rede.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the main differences between traditional Web analysis and Online Social Networks (OSNs) studies, is that in the first case the information is organized around content, whereas in the second case information is organized around people [1]. While search engines have done a good job finding relevant content across billions of pages, nowadays we don't have an equivalent tool for finding relevant people in OSNs. Even though an impressive amount of research has been done in this direction, there are still a lot of gaps to cover. Although the first intuition could be to search for popular people, previous research has shown that a users' in-degree (*e.g.* number of friends or followers) is important but not enough to represent importance. Another approach is to study the content of the messages exchanged among users, trying to identify topical experts. Nonetheless, the computational cost of such approaches and the differences in languages are their main limitations.

In this thesis we take a content-agnostic approach, focusing on frequency, type, and time properties of people's actions, mixing their static characteristics (social graph) and their activities (dynamic graphs). Our goal is to understand how people's popularity can be used (or not) to find relevant people in different contexts, and also discover new features (mainly focusing on user dynamics) that are useful to find given types of people. Specifically, we study the following three questions:

i) Which users produce higher value for the OSN provider?

ii) Which people push and propagate new ideas in OSNs?,

iii) How to find relevant people (*i.e.* potential customers) for local advertisers in a given geographical space (*e.g.* a city)?

We show that in all these cases, popularity is not enough to find relevant people. Moreover, we found a low propensity to adopt new ideas in popular users, while "smaller" but more active people are the ones setting new trends and also generating monetary value for the OSN provider. When geographical constrains are included, we show that is also necessary to add domain knowledge in human mobility to find relevant people in that scenario.

The main contributions of this thesis are:

1. **An analysis of people popularity according to their amount of activity**. Comparing two OSNs (Facebook and Twitter) with an e-mail network, we find a high correlation between outgoing (perform an action) and incoming (receive an action) activity for the OSNs, but not for the e-mail network. The strong correlation persists in Facebook's users even for very active people. We find similar results in Twitter, suggesting that this relation could be a distinctive property of Online Social Networks and Social Media.

2. **Develop a novel methodology to compute the monetary value that each person generates for the OSN provider**. A users value can be divided into direct impressions (advertising opportunities that a person provides by browsing OSN site pages) and indirect impressions (advertising opportunities that a person provides by enticing others to browse OSN site pages). Indirect impressions can cascade, where a person's actions ultimately cause other people to visit the OSN through a chain of many other users. Our experiments show that popular people are more active and valuable. However, the correlation is much higher for the activity than for the number of friends, highlighting the importance of the amount of people activity.

3. **Trendsetters ranking**. This algorithm allows to find people that adopt and spread new ideas influencing other people before these ideas become popular. We show that popular people tend to arrive late for new trends, while users in the top of our ranking tend to be early adopters that also influence their social contacts to adopt the new trend.

4. **A recommender system to find potential new clients for local shops (venues) in mobile-social media platforms (such as FourSquare or Facebook Places)**. We show that in this context, a user's activity (in this case the amount of venues visited) is useful to find relevant users, but also requires domain knowledge in human mobility. By combining these elements we are able to produce reasonably accurate and scalable recommendations, matching advertisers with potential costumers. We show the advantages and limitations of each approach (activity based and also considering human mobility knowledge) for different type of venues.

This thesis can be divided in three parts. First, contribution (1) is used to develop contribution (2), and together they study the relation among people popularity, activities and value. Second, contribution (3) introduces the importance of considering temporal information, and propose a methodology that adds that information in a weighted graph. Finally, contribution (4) includes a recommender system approach to find relevant people taking in account geographical patterns. Although these three parts can be considered independently, across all of them we find the limitations of static graph approaches to find relevant users. We show consistently across three different OSNs (Facebook, Twitters and FourSquare) that is possible to find relevant people by adding dynamic information (from activities), without performing a costly content analysis.

This work has produced the following publications:

- **Diego Saez-Trumper**, Yabing Liu, Ricardo Baeza-Yates, Balachander Krishnamurthy and Alan Mislove. *Beyond CPM and CPC: Determining the Value of Users on OSNs*. Submitted for Review. (Chapter 4).

- Gabriel Magno, Giovanni Comarela, **Diego Saez-Trumper**, Meeyoung Cha and Virgilio Almeida. *New Kid on the Block: Exploring the Google+ Social Graph.* In Proceedings of the 12th ACM SIGCOMM/USENIX Internet Measurement Conference. Boston, U.S.A. November, 2012. (Chapter 2).

- **Diego Saez-Trumper**, Daniele Quercia and Jon Crowcroft. *Ads and the City: Considering Geographic Distance Goes a Long Way.* In Proceedings of the 6th ACM Conference on Recommender systems. Dublin, Ireland. September, 2012. (Chapter 6).

- **Diego Saez-Trumper**, Giovanni Comarela, Virgilio Almeida, Ricardo Baeza-Yates, Fabricio Benevenuto. *Finding Trendsetters in Information Networks.* In Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Beijing, China. August, 2012. (Chapter 5).

- **Diego Saez-Trumper**, David Nettleton and Ricardo Baeza-Yates. *High Correlation between Incoming and Outgoing Activity: A Distinctive Property of Online Social Networks?.* In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain. July, 2011. (Chapter 3).

This thesis is divided in seven chapters, each of them based on a published or submitted article. These chapters are organized as follows: Chapter 2 gives a general background and overview of the state of the art. Detailed related work is included and discussed within the rest of the chapters. Chapter 3 introduces a preliminary study about the relation between people's actions and their popularity. Chapter 4 introduces a model to compute the value that each user creates for the OSNs provider. In Chapter 5, we develop an algorithm that allows us to find trendsetters in information networks. In Chapter 6, we take a recommender system approach to finding potential (relevant) customers for local shops such as restaurants and clubs. Finally, in Chapter 7 we present the final remarks, conclusions, and outline some future work.

Chapter 2

Background

This chapter gives a general background in OSNs studies and it is divided into three sections. First, we define some general concepts and terminology. Second, we present an overview about research in social influence. Third, we describe recent findings about OSNs' structure. This section includes a current state of art and our own results. Finally, we describe the data sets used in this thesis.

2.1 Preliminary Concepts

In this section we define some concepts and terms that we will use in this thesis. Even though colloquial terms like "Friends" or "Followers" could be easy to understand, we consider necessary to have a clear definition to avoid ambiguities.

Online Social Networks (OSNs): There are different types of online social networking websites. Facebook is the most popular OSN, thus it could be considered as the "canonical OSN". On the other hand, Twitter (and also Google+) could be considered as social, but also as information networks. Previous research differentiates these two types of networks (Social and Information ones), considering that the former are those with high presence of reciprocal relations (if A is

following B, B is also following A), while the latter has less reciprocal edges.

There are also *specialized OSNs* that focus in specific communities or features. For example CouchSurfing[1] is an OSNs used to find hosting for travelers, and Foursquare is a geo-social network where users share their whereabouts.

In this thesis we use a comprehensive definition of OSN that includes all the OSNs described earlier. We consider that a OSN is any online platform that connects people[2] trough explicit edges, allowing users, to exchange information among them. Later, in each chapter, we detail the specificity of each OSN studied.

Friends and Followers: As a convention we refer to *friends* when the relation between two users is symmetric (like in Facebook, where to create a "Friendship" both users need to explicitly accept it), and as *follower* when the relation can be established unilaterally (like in Twitter).

Geo-social Networks: This is a specific type of OSN, where users share their whereabouts with friends by using mobile social-networking applications. Foursquare, Gowalla, and Facebook Places are some examples. In Chapter 6 we introduce a methodology to find relevant people using the features of geo-social -networks.

Hashtag: It is a form of meta-data tag. Hashtags became a convention in Twitter to tag a post, helping other users to search information about an specific topic. For example, during the *PRISM scandal*[3], people were tagging their post with hashtags like #PRISM or #Snowden. Therefore, other people interested in the issue could find information about that topic using those hashtags. Recently, Facebook also adopted this convention, allowing people to use hashtags to search across public posts.

[1] http://www.couchsurfing.org/

[2] OSNs connect mainly individuals but OSNs accounts can also belong to groups, companies, or even *bots*. However, their main purpose is to connect people.

[3] http://en.wikipedia.org/wiki/PRISM

2.2 Social Influence

In this section we present an overview of the state of art of research about social influence.

2.2.1 Studies across Different Disciplines

The study of social networks to find relevant people has been addressed by different disciplines. In the 1950s the social psychologist Solomon Asch published a well known study about the group's influence in individuals decisions [3]. Years later, in 1968, the marketing researcher Frank Bass proposed a model for the adoption of new technologies on the market [10]. He considered that there are two types of new adopters: the innovators and the imitators. Bass models the relation between these two types with a differential equation. Other two important influence models based on individual thresholds were proposed at the end of 1970s by the economist Thomas Schelling [80] and by the sociologist Mark Granovetter [32]. Nowadays, the study of influence and information diffusion is a hot topic in the computer science community. Next, we review and summarize the different approaches related to our work.

2.2.2 Probabilistic Models

The relevance of viral marketing has inspired studies looking for influential users that can maximize the information propagation in social networks. This problem has been faced by probabilistic models [23], and as a discrete optimization problem [41]. A complementary work in the same direction has been done by Chen *et al.* [19]. Most recently, a probabilistic approach to find topical authorities in microbloging sites has been proposed by Pal and Counts [64].

2.2.3 Group Influence

Backstrom *et al.* [6] shows that Livejournal users[4] and also DBLP authors tends to join new groups when their *friends* joins too. In the same way, Romero [73] has studied the adoption of *hashtags* in Twitter about different topics related with the number of *friends* that have used these tags previously. They show that users wait that their friends use some hashtag before them, and depending on how controversial is the topic, people wait for more friends to follow before they jump in. However, those works focus in the direct influence of the group (friends or friend of friends) but does not take in account how the indirect influence spreads over the graph.

2.2.4 Early Adopters

The concept of Early Adopters was studied by Bakshy *et al.* [9], analyzing data from a popular online virtual world (Second Life[5]), discovering that the trendsetters are usually users with few friends, *i.e.* nodes with low degree, and moreover, they are users that are not too evolved in the game (they play less hours than the average). These are key points for our work, because shows that users that can be considered outsiders in a trivial analysis gain importance when the time factor (to be an early adopter) is considered. In this thesis we take the concept of early adopter, but we propose a way to differentiate those that create cascade behavior. That means that we are interested in influential early adopters (Chapter 5).

2.2.5 Temporal Factor

The importance of the temporal factor in influence studies was remarked by Anagnostopoulos *et al.* [2]. Studying the different sources of correlation among users actions, the authors proposed three possible explanations to a group of users performing the same action: (i)

[4]LiveJournal is a popular social network of bloggers (http://livejournal.com)
[5]http://secondlife.com

environmental factors, (ii) homophily, and (iii) influence. An example of an environmental factor is the fact that a group of social media users living in the same city can post about the same event, because that event is taking place in their city. Homophily is similar, but corresponds to intrinsic characteristics of people. That means, that two users can post about the same topic because they have the same interests, for example a football game or a TV show. The authors point out that we can only talk about social influence when there is a time causality associated to the actions among users.

Another consideration about the time factor in information networks is described by Kossinets *et al.* in [45], showing that considering only the topology is not enough to understand how the information spreads over an information network, because some edges could be *slower* than others and that in many cases the information can go faster through a multi-hop pathway that uses *faster* edges. Other important studies have been conducted taking in account the temporal factor to find the backbone of cascades produced on the Web [29], but they do not propose a ranking function neither model topics, and they only establish a temporal relation among nodes, looking to the most common path in cascades process, creating a influence pathway.

Goyal *et al.* [31] describe leaders as users that take actions that will be imitated by their friends later. They discover, among other things, that there are users that are tribe leaders, meaning that they are imitated in different actions by the same group of friends (the tribe). In this case to be the first is a signal of leadership. On the other hand, in [91] being the last is considered a signal of expertise. This work studies the relation into a Java Forum among users making questions and giving answers for this programming language. These relations are model as a graph and ranking algorithms - such as PageRank [63] and HITS [43] - are used to find relevant users. Comparing their results with human evaluation they conclude that those algorithms allows to find expert users. HITS also was used to model influence and passivity in social media [74].

2.2.6 Memeshapes

In [90], Yang and Leskovec, propose the KSC-algorithm to cluster temporal series by their shape, applying it to *Internet memes*[6] and Twitter hashtags. They found different kinds of shapes and they explain them by the nature of the sources (such as bloggers, mainstream media, etc). This work is a fundamental input for this thesis, as we have used this algorithm to cluster the trends in our data set (Chapter 5). However, the goal of our work is different as we are not interested in who is talking about a topic when it is popular, but just before it became popular. Differences about how we apply the KSC-algorithm are explained in Section 5.5.

2.2.7 Pagerank based Algorithms

The idea to use Pagerank to evaluate user's influence in Social Media was already developed by Weng *et al.* [89]. There, the authors used a topic-sensitive Pagerank extension [36] proposing a *TwitterRank* to evaluate Twitter users. In fact, this work is not focused on influence but in homophily, measuring the similarity among users, avoiding the temporal factor and considering the amount of information that each user posts (*i.e.* number of tweets) in a given topic to assign importance that each user gives to that topic. Something similar has been done by Liu *et al.* [53].

Other rankings on Twitter have been studied by Cha *et al.* [17], showing that the number of followers (node in-degree) is not necessarily an indicator of influence, naming this fact as: "The Million Follower Fallacy". Kwak *et al.* [46] also ranked Twitter users, showing the differences between a node degree based ranking versus the results of Pagerank. However, those works have been using the social graph without consider the temporal dynamics of communication.

[6] In this case Yang and Leskovec define *memes* as "shorted quotes textual phrases" [90].

2.3 OSNs Structure

In order to design efficient algorithms to find relevant people, it is necessary to understand the main properties of OSNs' graphs. In this section we review the most prominent studies on OSNs' graph characterization. These studies focus mainly in the most well-established OSNs platforms like Facebook and Twitter. In addition, we compare this results with Google+ data[7], that is consider one of the fastest growing networks ever [55]. This comparison across different OSNs gives a general overview in the main similarities and differences among these OSNs.

2.3.1 Small World

Mislove *et al.* [58] studied graph theoretic properties of social networks, based on the friendship network of Orkut, Flickr, LiveJournal, and YouTube. They confirmed the power-law, small-world, and scale-free properties of these social network services. Kwak *et al.* found similar topological characteristics in Twitter [46]. We found similar properties in Google+ as shown in Table 2.1. This results are similar with the seminal studies in social networks done by Milgram [56] and more recently by Watts and Strogatz [88].

Network	Nodes	Edges	%[8]	PL	RR	D	ID	OD
Google+	35M	575M	56%	5.9	32%	19	16.4	16.4
Facebook	721M	62G	100%	4.7	100%	41	190.2	190.2
Twitter	41.7M	106M	100%	4.1	̃100%	18	28.19	29.34
Orkut	3M	223M	11%	4.3	11%	9	-	-

Table 2.1: Comparison of topological characteristics of four different OSNs. PL: Path Length; RR: Reciprocity; ID: In-degree; OD: Out-Degree [46; 55; 58; 4].

[7]Google+ data set is available at http://gplus.camps.dcc.ufmg.br

[8]This value correspond to the percentage of nodes considered to compute the statistics. For Google+, Twitter, and Orkut this value is an estimation considering

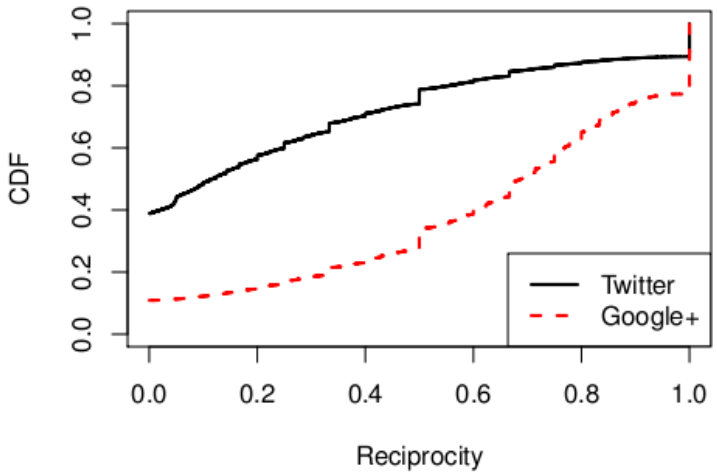


Figure 2.1: Distribution of Reciprocal links in Twitter and Google+.

2.3.2 Six Degrees of Separation?

Recently, Ugander *et al.* [84; 5] used the complete Facebook data set to study the social graph of Facebook. They show - among other things - that the degree of separation in that platform is 4.7; similarly, in Twitter this value is 4.1, while in Google+ it is 5.9 [55]. This difference may be explained by the fact that Google+ is a new platform at it should get denser in the future, as studied by [49] for different networks. However, all of them are lower than the popular idea of six degrees of separation (wrongly attributed to Milgram [44]).

2.3.3 Reciprocity

In Section 2.1 we mentioned reciprocity has been used to determine whether a network could be considered a Social or an Information Network. Following this methodology, Kwak *et al.* concludes that Twitter is an information network. Differently Google+ is more social, with the 32% of reciprocal relations, compared with the 22.1% in Twitter.

the total number of users declared for each OSN. For Facebook the study was done using the complete graph provided by this company [4].

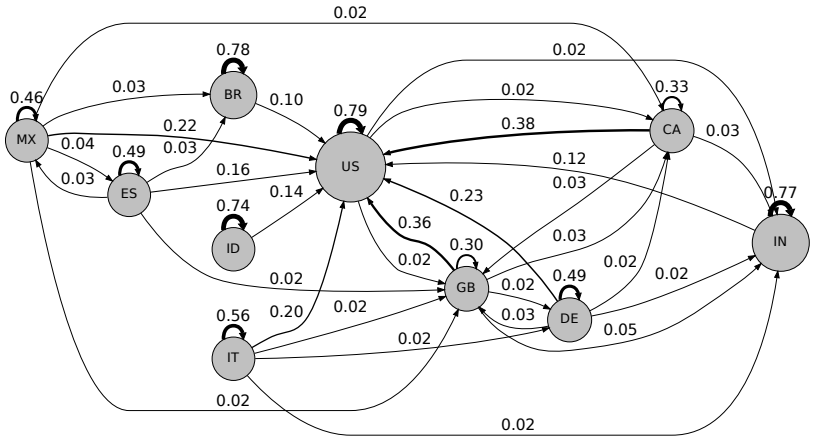


Figure 2.2: Link distribution across the top countries in Google+. Nodes are the top 10 countries in terms of number users, and weights represent fraction of edges over a given country. Country codes represent the following. US: United States; IN: India; BR: Brazil; GB: United Kingdom; CA: Canada; DE: Germany; ID: Indonesia; MX: Mexico; IT: Italy; and ES: Spain.

Figure 2.1 shows the distribution of the reciprocal relations in both networks. Note that in Facebook the relations are always undirected, therefore the reciprocity is always 100%.

2.3.4 Geo-Location

When it comes to research on geo-location of users in online social networks, Liben-Nowell *et al.* [50] analyzed the geographical location of LiveJournal users and found a strong correlation between friendship and geographic proximity. This work confirms that most social links in the blog network are correlated with physical distance and only 33% of the friendships are independent of geography.

Recently, Scellato *et al.* [78] showed that there is a strong relationship between geographical distance and the probability of being friends in social networks. They discuss the implications of geo-location for social networking sites. Rodrigues *et al.* [72] investigate the word-of-mouth

Chapter(s)	Platform	Year	Availability	Reference
2	Google+	2012	Available	[55]
3	Enron	2005	Available	[82]
3	Twitter	2010	Partially Available	[76]
3,4	Facebook	2009	Available	[85]
5	Twitter	2009	Partially Available	[17]
6	FourSquare	2011	Available	[20]

Table 2.2: Summary of the data sets used in this thesis.

based content discovery by analyzing URLs in Twitter. They showed that propagation and physical proximity have correlation.

Poblete *et al.* [66] studied a large amount of data gathered from Twitter and showed the various usages of the system across different countries. We found something similar in Google+ (see Figure 2.2), where certain countries like Brazil, India, and Indonesia appear far more inward looking when forming social links, than those outward looking countries like United Kingdom and Canada. This means that based on the geographical location of where a user lives, her expectation towards finding a stronger local community in the network is different. We believe this kind of social network analysis allows us to study the collective and deviant behavior of particular demographics, which are increasingly considered important and useful, both, in research and practice.

2.4 Data Sets

Here we summarize the data sets used in this thesis and also discuss privacy issues.

2.4.1 Summary

In this thesis we have used data from 4 different OSNs. Table 2.2 summarizes the data sets. Data sets from Google+ and Twitter (2009) are

big, containing tens of millions of users and billions of edges. Moreover, the Twitter (2009) data set has almost all the information (users, tweets and edges) from that platform until 2009. Enron and Twitter (2010) have tens of thousands of users. The data sets from Facebook and FourSquare are smaller, but thoroughly cover specific geographical areas.

Almost all the information that we have used is available for research purposes. In case of Twitter (2009) only the (anonymized) graph is available but not the messages (that we have used in our research), this due privacy reasons. A detailed description of each data set is given in each chapter.

2.4.2 Privacy Concerns and Data Availability

When we talk about OSNs analysis, privacy concerns are a sensitive issue. Therefore, it is important to remark that all the data sets used in this thesis was crawled from the Internet. Part of the data was gathered as part of this PhD work, and other data were obtained from available data sets from the research community. We have no access to "private data", thus this thesis was built only with "public data". However, the boundaries of privacy could be fuzzy. We consider public all the information that can be accessed by any person (or bot) trough a browser or a public API. We consider private all the information that requires special agreements with the OSNs providers or any other company, and we did not used such kind of data.

Despite that these definitions are quite strong, it does not mean that OSNs users are necessarily aware of the data that they publicly share on the Web. Moreover, it is also not clear if most of the users cares about their own privacy [47]. However, data availability it is a sensitive issue in OSNs analysis. Nowadays, the high use of OSNs trough mobile applications, it allow the OSNs providers not only to get all the information that is related with OSN uses (social graphs, messages, users IP address, etc.), but also all the information that they can get from the mobile phone, such as users exact location (GPS), phone contacts, calls, and SMSs among other things. These data is not public available for obvious reasons, but even the data that is directly related with the

OSNs uses it is not accessible for research purposes. This, generates a huge gap between OSNs providers and researchers, and also among companies and this is a big difference with traditional web studies. While all the research to find relevant pages on the Web has been done in an open environment, research to find relevant people is being done on what Sergey Brin has called a "Walled Garden".[9] This is a clear and strong limitation for all OSNs researchers and consequently for this thesis.

[9]http://www.guardian.co.uk/technology/2012/apr/15/web-freedom-threat-google-brin

Chapter 3

Correlation between Incoming and Outgoing Activity

3.1 Introduction

In section 2.2 we showed that there are many different definitions of what it is a relevant or influential user. Therefore, in this chapter we focus in a specific measure of relevance: the user's incoming activity. Incoming activity can be defined as actions that a given user receives from other users. While outgoing activity is the opposite. For example, in an e-mail network, incoming activity is defined as received e-mail, while an outgoing activity it is to send an e-mail. In this study we start comparing two different network, the already mentioned e-mail network, and Facebook wall posts network. A wall post is one way of communicating in Facebook where a *friend* can write a public[1] message to another person. When user U posts a message in user V's wall, we consider that user U has a wall post done and user V has a wall post received. We represent these two networks as weighted directed graphs, where the weight of the edges represent the amount of actions (messages) between a pair of users (see Figure 3.1). Hence, we are able compare both networks.

[1] Users can set different levels of visibility depending on their privacy settings.

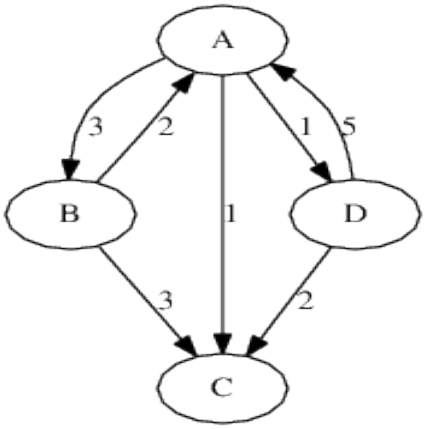


Figure 3.1: Activity example: user A has 4 friends, with an outgoing activity of 5, incoming activity of 7, and total activity equal to 12. User C has incoming activity but no outgoing activity, and therefore we do not consider user C as an active user.

Using this model, we study the relation between incoming and outgoing actions and compare it with people's popularity (number of friends). Our results confirm previous work which showed that the number of friends is not highly correlated with incoming activity (see Section 2.2). However, we find that in the Facebook network, a user's outgoing activity is highly correlated with his/her incoming activity. In contrast, in the e-mails network the correlation is lower. This result suggest that one important factor what is needed to become a relevant user is to generate outgoing activity.

We also extended this analysis to a Twitter sample. As mentioned before, previous work showed that there is low correlation between the amount of followers and user's *mentions*[2] [17]. In this case we consider the number of tweets as an outgoing activity, and define two different kinds of incoming activity: (i) the number of mentions and (ii) the number of followers. Our results show that users with more tweets usually has more mentions and followers than users with a low level of outgoing activity.

[2]In Twitter a "mention" it is when a someone refers explicitly to another user, by writing username.

3.2 Related Work

Differences between OSNs' interaction graphs (that quantify actions among users) and social graphs (that only consider static relations without consider the interactions) have been pointed by Wilson *et. al.* Both graphs (interaction and static) are conceptually different, the first is a weighted graph, while the second is unweighted. Moreover, in real OSNs when the interactions are considered and the inactive edges (weight equal to 0) are removed the graph topology changes significantly, modifying - among other things - nodes' centrality and degree. This explain why previous research shows that the users' popularity (*i.e.* in-degree in the social graph) can not be used as a (strong) signal of influence [17; 46]. While popularity is measured in static graph, influence (beyond the many different definitions of influence depicted on Section 2.2) implies flow information, and therefore interactions among users, thus it can not be measured as static property. For example, Weng *et. al* [89] have used the amount of actions in specific topic as signal of expertise on that topic. In our case, we use a more general definition of activity, that allows to compare different kind of networks (in this case e-mails vs OSNs) and actions (e-mails, tweets and wall posts).

3.3 Data Sets

We have used three different data sets, summarizing their characteristics in Table 3.1, as follows:

3.3.1 Enron's Emails

Enron's e-mails it is a well known data set. From the version that we obtained [82] we processed information from 250,483 users. We consider each unique e-mail address as a user. Because we are interested in active users we filtered considering only users that have at least one e-mail sent and one e-mail received. Applying this filter we obtained 11,254 active users.

3.3.2 Facebook Wall Posts

The data set used corresponds to the New Orleans Facebook's Regional Network[3] [85] containing information about 58,016 different users, who have been anonymized. Specifically we have two lists, one containing user-to-user links (*F*riendships), and a second list with user-to-user *w*all posts (where A,B means user A posting in B's wall), and a timestamp. The information covers a time-span from September 2006 to January 2009. From these lists we can obtain the number of friends for each user and his/her outgoing and incoming activity. Because we are interested in the users' interaction, we only consider "active" users, that means that they have at least one wall post done and one wall post received. Applying this filter we obtained 34,277 active users.

3.3.3 Twitter

In 2010, the Twitter users has an correlative ID [11]. And their information could be accessed by the Twitter API[4]. In August of 2010, we randomly selected 250,000 numbers between 0 an 150,000,000 (the estimated number of accounts at that time) being able to download approximately 55% of these ids, corresponding to 136,662 users. The information about users contains, among other things, their number of tweets, followers, friends (also called followees), profile age (date of profile creation), and, if there exists, a URL associated to the profile. We also downloaded the last 200 tweets from each user. To avoid inactive accounts, we removed all users with less than 5 followers and less than 5 post. After that filtering, we have have 61,096 active eusers.

It is important to remark that our data set is consistent with similar and more detailed studies of Twitter [46], which have demonstrated power law distributions for user's followers and friends. The time of the profile creation (profile's age) covers June 2006 to July 2010, the number of followers ranges from 0 to over 600,000 and the number of tweets from 0 to over 300,000 per user.

[3]Regional Networks have been deprecated by Facebook since August 2009
[4]https://dev.twitter.com

Data set	Users	Active Users	Activity
Enron	250,483	11,254	1,277,214 emails
Facebook	58,016	34,277	836,576 wall posts
Twitter	136,662	61,096	54,764,095 tweets

Table 3.1: Summary of data sets used in this chapter.

3.4 Experiments

Our experiments try to find the relation between incoming activity and outgoing activity. In the case of the Facebook and Enron data sets, we we compare the same kind of outgoing and incoming activity, *i.e.* posts done versus posts received and mails sent versus mails received, respectively. Then we apply a simple correlation parameter. In the case of Twitter, the nature of the outgoing (a tweet) and the incoming activity (followers or mentions) is different. In this case, we group users by their level of activity and then we compare them.

3.4.1 Facebook and Enron

The results in Table 3.2 confirm previous work in terms of that there are not strong correlation between the number of friends and incoming activity (0.43). However, the correlation between posts done and posts received ($\rho_{i,o}$) appears very strong (0.91) in the Facebook data set. However, one can think that this correlation is a property of any communication network. To test this intuition, we compute the same correlation between outgoing and incoming activity for a e-mails network (Enron). Interestingly, we found that in Enron data set this correlation is lower (0.51) than in Facebook.

But given that the distribution of user activities is skewed (see Figure 3.2), with most of users having low activity, is interesting to know whether this correlation holds true when we remove the less active users, *i.e.* we want to know if this correlation is also valid for the most active users (the minority), or if it is just dominated but users with low activity (the majority). To study this problem, we recompute $\rho_{i,o}$ as

	# Friends	Out. Act.	In. Act
Number of Friends	1	0.47	0.43
Outgoing Act. (Posts Done)		1	0.91
Incoming Act. (Posts Received)			1

Table 3.2: Facebook's correlation matrix.

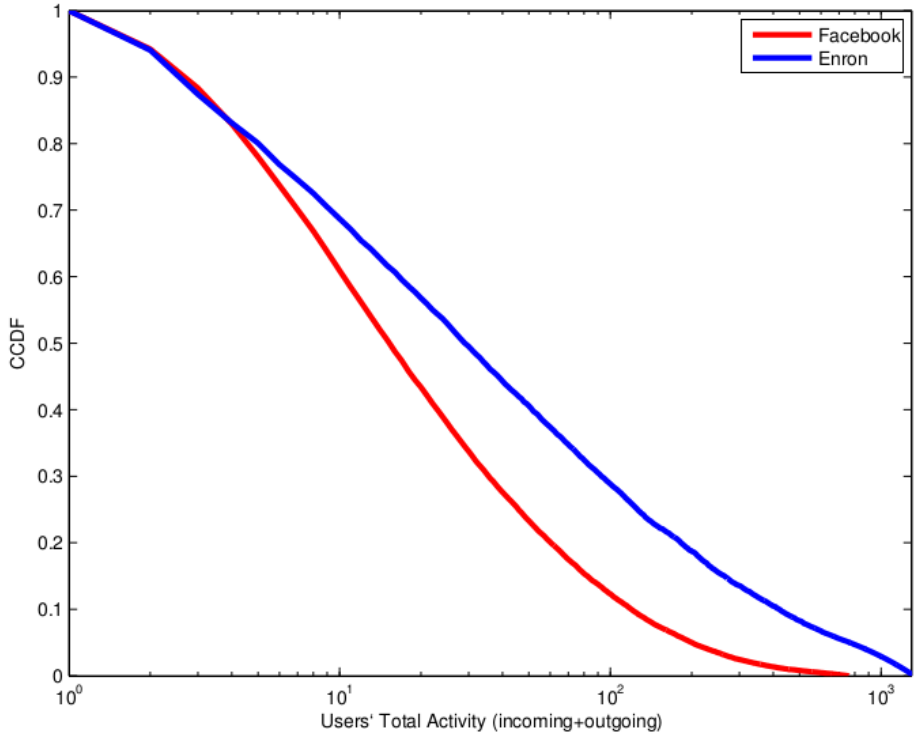


Figure 3.2: Complementary Cumulative Distribution Function (CCDF) of users' total activity (incoming+outgoing).

function of user's total activity (incoming plus outgoing). Specifically, we decreasingly sorted users by their amount of total actions , and recompute correlation $\rho_{i,o}$ cumulatively adding the next (less active) user; *i.e.* we start computing $\rho_{i,o}$ only for the most active users, and progressively adding less active users. We repeat the same procedure for the Enron's data set. The results are reported in Figure 3.3, show-

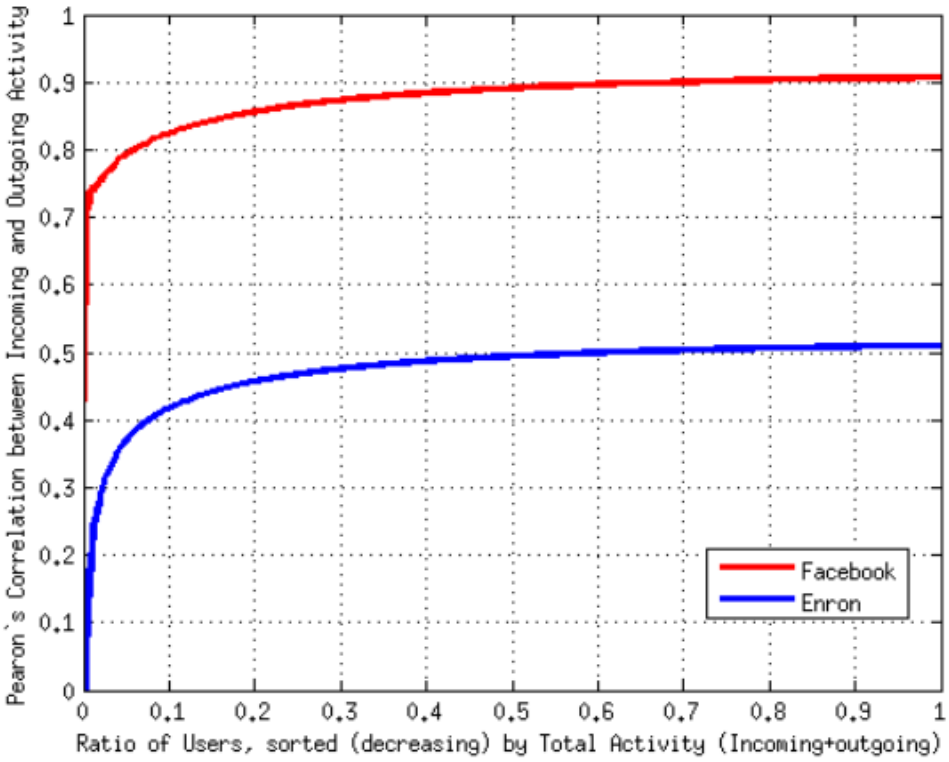


Figure 3.3: Correlation between users' outgoing and incoming activity depending on the amount of total actions. Users are decreasingly order by their amount of total activity (incoming+outgoing).

ing that the correlation is always higher in Facebook than in Enron. In Facebook the correlation is over 0.7 even if we only consider the most active users. In the case of Enron, the correlation is low for active users. Hence, the high correlation between outgoing and incoming activity is not only higher, but also more consistent for Facebook than for Enron.

3.4.2 Twitter

Previously, we found that the strong correlation was higher in a OSN (Facebook), than in a e-mails network (Enron). Here we want to ex-

tend our study to another platform. Given that we can't directly translate the incoming activity (there were not incoming posts in Twitter), in this section we consider two different kinds of incoming activity: first we have used the number of mentions that a user obtains. User A gets one mention when any user X post a message with the prefix @A. This was a standard in Twitter, and intuitively is similar to the e-mail reply we used in previous section. Next, we repeated our study considering the number of followers as an incoming activity.

We have used followers as incoming activity for three reasons: firstly, as we mentioned before, in the literature there are authors who have already studied different kinds of incoming activity such as retweets or mentions; secondly because this information (the number of followers and updates) is publicly available and can be obtained directly from the Twitter, and thirdly because we consider that obtaining followers could be considered as an important goal for Twitter users, so it is interesting to test if producing outgoing activity is a good strategy to that aim.

3.4.3 Number of Posts and Mentions

We started studying the relation between post an mentions. The number of mentions for each user is not contained in the information given by the Twitter API. Therefore, was necessary to use the Twitter Search API[5] to estimate this number. To face this challenge, we divided our sample by outgoing activity level (number of posts) in five categories (where category 1 are users with very low activity and category 5 are the very active users) and created an stratified sub-sample, selecting randomly 2,000 users for each category. Next, we estimated the number of mentions for these 10,000 users. We founded that around 80% of users does not have any mention, and only 5% users have over 15 mentions. Moreover, in the three categories with lowest outgoing activity (that is, users with 300 posts or less) only 10% of the users have at least one mention. Users in the most active groups (category 4 and 5), over 300 and 1200 posts, have 22% and 47% of users with at least one mention. However, a small fraction of users have at least 15 mentions,

[5]https://dev.twitter.com/docs/api/1/get/search

Outgoing Activity	Over 1 reply	Over 15 mentions
Very Low	0%	0%
Low	2%	0%
Medium	7%	1%
High	19%	9%
Very High	48%	17%

Table 3.3: Twitter: Percentage of users with mentions (incoming activity).

and they are concentrated in the category of most active users (see Table 3.3). The small number of mentions does not allow to make conclusions, but we can see that there are a relation between outgoing activity and Twitter mentions.

3.4.4 Number of Posts and Followers

Next, we want to study the relation between number of posts (outgoing activity) and number of followers (incoming activity). To that end, we have divided our data set in 5 bins using a equal frequency discretization for each feature. In this way, each user can be defined as an instance with two features, thus: U($outgoing_bin$,$incoming_bin$). Considering that we have 5 possible values for each feature, it is possible to have 25 different types of users. As an example, a user type T could be characterized by a low outgoing activity and a very high incoming activity. Figure 3.4 shows that when the level of outgoing activity increases, incoming activity also increases. As a consequence, users with a high number of followers have the biggest number of posts.

3.5 Discussion

We have found a strong correlation between outgoing and incoming activity in the Facebook data set. Our analysis of Twitter confirms this relation. This correlation is lower for Enron. However, we have studied three different networks that have no uniform communications

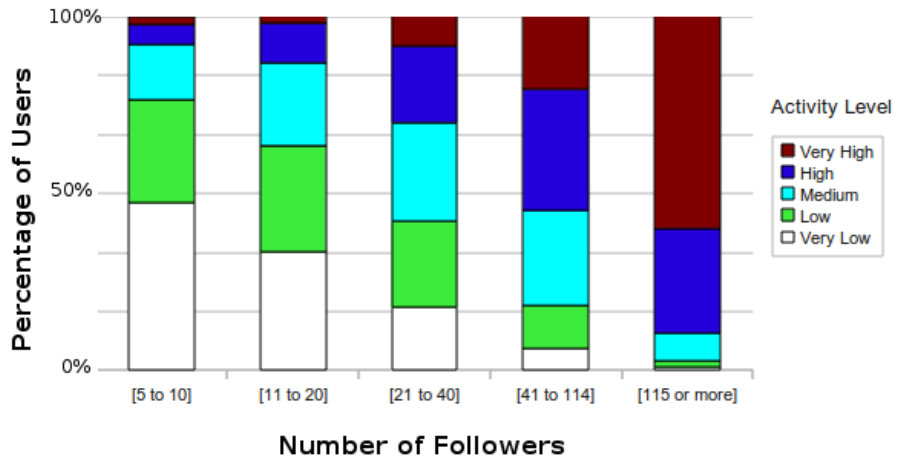


Figure 3.4: Twitter: Relation of followers and number of posts. On the x-axis the users are grouped by their number of followers, and on the y-axis we show the different levels of post activity.

patterns: Facebook wall posts has a one-to-one pattern, e-mails and Twitter could be considered as 1 to n pattern, and all of them shows this outgoing/incoming strong correlation. Moreover, the correlation persists for Facebook's users with a very high outgoing activity, and is also present for users with the same characteristics in Twitter, but is lower for Enron network, suggesting that this relation could be a distinctive property of Online Social Networks and Social Media. Later, in Chapter 4, we discuss how this property can be used to compute user's monetary value.

It is important to remark that amount of spammers were low in Twitter when this data set was collected [11]. But this has changed in recent years, and nowadays spam is usual [54]. This fact can have an impact on our method, however this can be faced by differentiate users where high activity is concentrated over a short time span - which could indicate spammer behavior- from those users whose activity is periodical and more spread over time.

Chapter 4

Determining the Value of Users on OSNs

4.1 Introduction

In this chapter we go beyond the study of users' activity and use that information as basis for determining the monetary value that each user produces for the OSN.

Nowadays, advertising constitute the economic basis of the Web; many sites provide free services supported by advertising. Typically, advertisers place ads on search engines by specifying keywords of interest; different keywords have different prices with their market price determined via a dynamic auction. The cost to advertisers when using this type of ads is generally expressed in terms of CPC (see Section 4.2.2).

However, the impressive grow of OSNs such as Facebook or Google+, have also seen an advertising market develop. Advertising on OSNs works in a manner similar to advertising on search engines: advertisers specify targeting parameters (i.e., attributes that the advertisers desire users to have in order to be shown their ad) and a CPM/CPC bid price, and the OSN ranks the ads to select the ones to be shown. The ranking is typically based on the bids and the click-through-rate (CTR) of the ad, but other parameters could be considered.

This strong similarity between advertising on the Web and OSNs is surprising given that OSNs are significantly different from a typical Website. On Web-search-based ads such as in Google or Yahoo! networks, the search engine knows relatively little about the user. Instead, the network must track users using cookies and other techniques, extracting more information about users through data mining. On OSN-based ads, users must have an account and be logged in order to even see ads. As part of participating in the OSN, users provide information about themselves in profiles (list of friends, demographics, educational history, hobbies, etc) and through interactions with the site (posting updates, "checking in", installing applications, etc).[1] Moreover, because the OSN is run by a single centralized entity, the OSN observes all user actions on the site (exchanging messages, uploading content, browsing others' profiles, etc.).

But, have all the users the same value? We posit that users on OSNs have sharply different values -in terms of the revenue they generate through ad impressions - to both, the OSN itself and advertisers. For example, influential users, such as those who share lots of content, have many friends, and whose posts get forwarded, are all likely to be more valuable than the average user. Such users bring more value to the OSN itself (as they create more advertising opportunities) and to advertisers as well (as their activities offer more opportunities for the advertiser's message to be spread). However, little work has gone into studying how the value of users in OSNs varies, and determining the extent to which users' value contributions can be extracted, quantified, and presented to advertisers.

Here, we present a framework for reasoning about the value of a user in OSNs like Facebook. We examine the wealth of information that the OSN operator receives about user activity on their site and present a methodology for reasoning about how different user actions correspond to revenue. We argue that a user's value can be divided into direct impressions (advertising opportunities that a user provides by browsing OSN site pages) and indirect impressions (advertising opportunities

[1]It is worth noting that any information may be intentionally falsified by the user; while this may be easy for certain types of information (e.g., profile attributes), it is considerably more difficult for others (e.g., check-ins).

that a user provides by enticing others to browse OSN site pages). Indirect impressions can *cascade*, where a user's actions ultimately cause other users to visit the OSN through a chain of many other users.

It is important to remark that having a proper understanding of each user's value can benefit the OSN provider, the advertisers and also the users. OSNs can enable targeting of "more valuable" users, increasing revenue and making their advertising platform more useful. Advertisers are likely to pay more for such desirable users. Finally, uncovering the value of users will also benefit the users themselves, as they become aware of their relative value.

We explore our framework by leveraging a detailed data set from Facebook covering users in the New Orleans metropolitan area from 2009 [85]. Our data set covers 90,269 users, and contains each user's profile and activity trace (i.e., the activity visible on their "wall"). We show that activity can be used to estimate the number of impressions attributable to users, and that users from our data set have sharply different values.We also show that our model can be extended to represent the user value in monetary terms beyond just the number of advertising impressions.

The remainder of this chapter is organized as follows. Section 4.2 provides background and discusses related work. Section 4.3 summarizes the current advertising model used by Facebook, while Section 4.4 presents a model to estimate the user's value in a OSN. Section 4.5 evaluates the proposed model on real-world Facebook data. Finally, we provide a concluding discussion.

4.2 Related Work

4.2.1 Information Diffusion

A recent study [61] showed that about 70% of the information volume in Twitter can be attributed to network diffusion and only 30% to external influence. While this implies an opportunity for viral marketing campaigns, effective strategies for such campaigns are debatable—

penetrating a chain of small communities [1] may be one route. Deciding where to "seed" advertisement to reach the biggest possible audience remains a challenge. As we discussed in Section 2.2 social contagion is a complex process that has been studied from different approaches such as: finding the set of users that maximize the probability of spreading [24; 41], discovering topical authorities [53; 89] or identifying trendsetters [53; 75]. Other roles in the diffusion process include promoters [12], early adopters, and imitators [9]. All of them have a well defined function, but share a complex relation among their basic elements, that makes modeling and predicting social influence difficult.

Beyond social contagion, cascading behavior is common in OSNs [16; 48], and although the user's popularity does not necessarily create cascades [18], being popular is essential for direct influence [8] (popular users broadcast to a broader audience). Moreover, if we consider that cascades tend to be wider than deeper [71], the size of a user's audience (i.e., node in-degree) is key to estimates their value. While complex contagion is hard to predict, direct broadcasting is easy to compute and hence safer for advertisers. Next, we relate the user's value with the impressions generated and not with the potential for creating deep cascades.

4.2.2 Online Advertising Models

Online advertising has been widely studied [34; 87]. Recent studies are proposing techniques for online advertisers to maximize their revenue [69]. As we mentioned before Cost per Click (CPC) is the standard to express ads prices, and it refers the the cost of a single ad click independent of the number of impressions. Another popular option is the Cost per mille (CPM) that is the cost of 1,000 ad impressions independent of the number of clicks.Furthermore, beyond CPC and CPM [59; 62], there are other proposed pricing models [26; 27] that combine both. Our work is complementary to these, as we focus on differentiating between more- and less-valuable users.

Field	Options (examples)
Age	Min and Max
Location	Country, State, City, Zip code
Gender	Male, Female, All
Activities	Cooking, Dancing, Gaming
Family Status	Baby boomers, Engaged (N years), Parents (child: 0-3yrs)
Sports	Cricket, basketball, baseball

Table 4.1: Example of Facebook targeting parameters.

4.3 Current Advertising Model

To study current advertising models in OSNs, we focus on Facebook, as it is the largest and most mature OSN. Facebook offers targeting parameters—such as location, gender, interests—and advertisers pay either per click or per impression for users who match the advertiser's specified targeting parameters. Although the parameters could be detailed in terms of demography and interest (see Table 4.1) they currently do not relate to the target user's popularity or level of activity within the OSN (e.g., there is generally no mechanism for directly targeting "users who are influential").

As long as the users match the advertiser's targeting parameters, advertisers pay the same amount to show ads to them independent of the number of friends they have. As most advertisements can be shared with their friends (e.g., via the "Like Button"), this approach clearly undervalues some users, as advertisers are paying the same for users who may share ads with thousands of friends or just a handful.

Facebook currently offers a multitude of options for advertisers, with objectives ranging from viewing an ad to "Liking" a page. All these options can be expressed as either Cost per Mille (CPM, or the cost of 1,000 ad impressions) or Cost per Action (the cost of an advertiser-selected user action regardless of the number of impressions). The most common Action chosen is clicking on an advertisement; this is commonly referred to as Cost per Click (CPC). The CPM and CPC prices are set through an auction mechanism, where each advertiser

bids the maximum that she is willing to pay for impressions or clicks. Facebook selects the best ads to present to the user and while ad selection algorithms are secret, the OSNs presumably use the highest bids in the case of CPM, or the highest expected revenues in the case of CPC (similar to the popular approach used in sponsored search advertisement auctions [39]). Thus, targeting parameters that are popular with advertisers are expected to have higher winning auction prices.

4.4 User Value Framework

As discussed above, the value of a user in an OSN is directly proportional to the number of advertising impressions and clicks that the user generates by their actions. The actions may be visible (such as uploading content or commenting on a friend's content) or invisible (such as visiting a friend's profile or browsing a friend's photos without commenting). While invisible actions generate only direct impressions, visible actions can also generate indirect impressions by triggering (visible or invisible) actions by friends. For example, a user commenting on a friend's photo may trigger other users to return to the OSN generating additional impressions. We argue that different user actions are likely to generate different numbers of impressions. The "place" where the action has been done (e.g., in the user's profile, friends' profiles, or on group/community pages) can result in generating different numbers of impressions. Also, the user's characteristics (e.g., the number of friends and demographic information) is also related to his/her value.

We propose a framework that considers all these factors and uses them to compute a user's value (in terms of the advertising revenue of the OSN that can be attributed to the user). First, we analyze why different actions produce different numbers of indirect impressions and how external observers or OSNs providers can measure this. Next, we show how users' characteristics and the places where they perform their actions affects their value. Finally, we propose a comprehensive methodology for computing users' values that can be applied to any OSN.

4.4.1 The Value of Actions

Each page on the OSN that the user visits gives an opportunity to the OSN provider to show advertisements. Whether the user is visiting the profile of a friend, uploading a photo, or watching a video, with each request for a new page, new advertisements can be shown. We call the advertisements shown directly to the user *direct impressions*. However, when a user perform actions that have effects visible to others in the OSN, the user has the potential to also generate *indirect impressions* (e.g., if a user uploads a photo, and some of the user's friends browse the photo, the OSN can show ads to the friends as well). Thus, when more people browse the OSN as a result of one user's action, more impressions can be attributed to the action. Consider an OSN where the most common action is to browse photos of friends, but articles posted by friends are rarely read. In such an OSN, when a user uploads a photo, the user is generating more indirect impressions than by posting an article. The value of an action is thus related to the new actions triggered.

The primary challenge in measuring the value of actions is that many of the impressions cannot be directly observed. Only the OSN knows when they occur and few OSNs provide visibility into how often other users browse content. OSNs alone can determine precise value of each action. An external observer, however, can estimate invisible actions, for example, by considering visible actions as a proxy for invisible actions. If photos consistently generate more visible feedback (such as comments, likes, or retweets) than articles, it is reasonable to conclude that photos generate more invisible actions than articles. Another option is to extrapolate this information from previous studies that have access to (private) invisible actions and show that most user activities on OSNs consist of visiting friend profiles and photos. We take this approach and use two studies [13; 81] that examined user's actions on popular OSNs. Table 4.2 shows the activity distribution of a large number of users in three different OSNs in these two studies. Although the actual distribution of user activities may vary with the OSN, we show in Section 4.5 that small variations in action value assignment do not dramatically impact the final user value.

Finally, there are some actions that we cannot capture and hence are

Facebook		Orkut		Hi5	
Category	Share	Category	Share	Category	Share
Home	35 %	Profile+Friends	41%	Photos	45%
Profile	16 %	Photos	31%	Profile	20%
Photos	16 %	Scrapbook	20%	Home	13%
Friends	4.7 %	Other	3%	Friends	13%
Groups	3 %	Communities	1%	Groups	1%

Table 4.2: Comparison of popular user activities across three OSN sites [13; 81].

not included in our model. For example more than 45% of the activity in Facebook in a large study [81] was messaging and applications. Applications are growing rapidly and is an important factor in revenue for the OSN and for app writers. In our data set these actions affect the value of a user but are unavailable to us.

Knowing whether a specific direct or indirect impression is more valuable is hard as this depends on multiple parameters (including the mood of the user and the actual ads shown). We thus take a neutral position with respect to their relative value. However, as most actions on OSNs are reactive (as seen in Table 4.2), the majority of impressions are indirect. A user uploads a photo and some fraction of her friends may comment on it. Thus, indirect impressions overall contribute more value than direct impressions.

4.4.2 Users Characteristics and Interactions

We next consider the information that external observers can collect to help estimate the value of users. Most OSN services make basic personal information provided by each user (gender, age, location, interests etc.) public by default. It is also generally possible to obtain some information about the social graph, such as the number of friends and their identities. While the basic information is useful for targeting (e.g., an advertiser is targeting 30-year-old men in Barcelona), the latter is useful to estimate the indirect impressions. For example, when a

user with thousands of friends posts an update, the user is broadcasting information to a wider audience than a user with only a few friends. As a user's friends tend to be similar in demographics and tastes to the user [1], it is conceivable that most of the indirect impressions would be shown to a similar target group.[2]

Beyond the number of friends, the location where the user posts can help determine the potential audience. Given that many OSNs allows users to reach friends of friends (e.g., on Facebook, posting on a friend's wall allows the friend's friends to also see that post), the location of the action plays a role in determining the value of the action. Thus, when an action includes interaction with other users, it is necessary to take into account the characteristics from all the users involved to compute this action's value with part of the value assigned to each user.

4.4.3 Measuring User Value

We define the following terms to measure users value covering the user herself, her possible activities, and activities on her profile:

User characteristics (u_c): This term measures individual user characteristics and is composed of two elements: the targeting parameters t and the number of friends d (i.e., the user's degree). We can tailor t to a given target group; if the advertiser is seeking older demographics, t could be defined as being proportional to the age. If the target is related to a geographical location, t would be inversely proportional to (e.g., the logarithm of) the distance. All such targeting parameters can be combined depending on the advertiser requirements. Precision and granularity of targeting will depend on user's demographic information available on each OSN (for example gender, countries GDP, etc.). The second parameter d reflects the amplification of an action as a result of the direct audience reached by each user. To be conservative, we define d as the logarithm of the node degree. Another alternative would be

[2]The implications of influence/cascades were discussed in Section 4.2.

to use a small fixed fraction of the number of friends. Hence

$$u_c \propto t \cdot d \propto t \cdot \log\left(\#friends + 1\right) \tag{4.1}$$

Notice that this only captures the first hop on an activity cascade. The next hop will be captured by the activities of the users influenced by this user.

User activity in her own profile ($u_{a_{self}}$): This is a weighted sum of u actions (such as #photos uploaded, #articles, etc.) done by u in her own profile or home page.

$$u_{a_{self}} \propto \sum w_i \#action_i \tag{4.2}$$

where $w_i \propto$ action value. Most probably, most activities in a user's profile correspond to direct impressions.

Friends activity in a user's profile ($u_{a_{friends}}$): For all the users v that are friends with u, we measure their activity in u's profile. Given that each time v performs an action in u's profile this information is sent both to u and v friends by default (this can be changed through privacy settings but is rarely done), we weight all these actions by v's individual characteristics v_c:

$$u_{a_{friends}} \propto \sum_{v \in |u|} v_c \sum w_i \#action_i \tag{4.3}$$

As discussed earlier, most actions are reactive and thus most of them will correspond to indirect impressions.

User activity in their friend's profiles ($u_{a_{visitor}}$): When u carried out an action in her friends' profiles:

$$u_{a_{visitor}} \propto \sum_{v \in |u|} v_c \sum w_i \#action_i \tag{4.4}$$

As in the previous case, most of these activities will likely correspond to indirect impressions.

If we are targeting all users in the same way (that is, $t = 1$), then the final formula for the u_{value} is a function of her activity and her friends:

$$u_{value} \propto \left(u_{a_{self}} + u_{a_{friends}} + u_{a_{visitor}}\right) u_c \tag{4.5}$$

The first weight captures mainly direct impressions while the other two capture mainly indirect impressions.

The definition could be extended to include different weights for friends based on tie strength (closer friends are more like to see our actions and generate more impressions), privacy settings, and "circles" (groups of users see only partial information about our actions depending on our privacy configuration). Groups or community activity could also be included. However, an easy way to add group activity, is to consider the group as a friend, where all the group members would be friends of friends. Then, group activities can be included in the terms $u_{a_{friends}}$ and $u_{a_{visitor}}$.

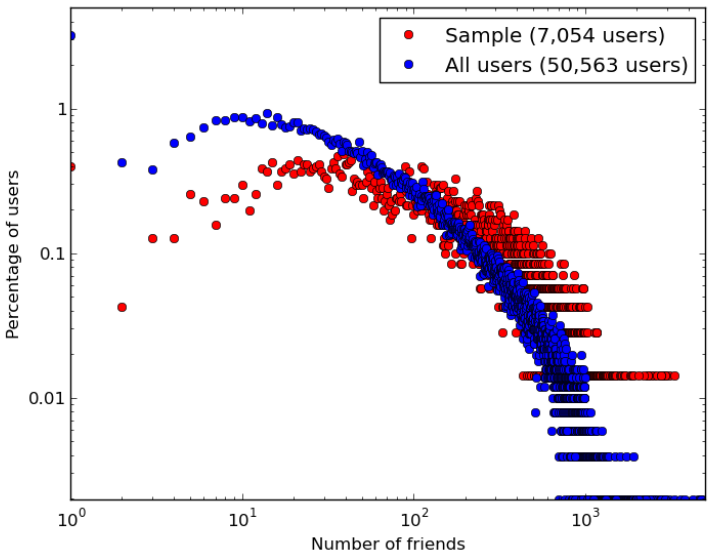
4.5 Evaluation

Having defined a framework for reasoning about the value of users to the OSN we now apply our model on a real OSN data set. The goal is to understand how value is distributed among users and how different strategies to measure invisible actions affect these results.

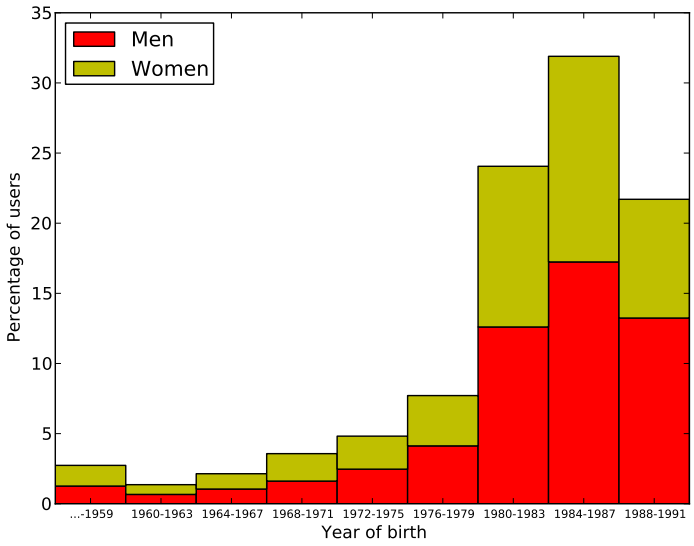
4.5.1 Data Set Description

We use a 2009 data set collected from Facebook covering users in New Orleans. We only consider the 50,564 users with public profiles out of the 90,269 users. As we are interested in classifying users by their interests and demographics, we use only those who share their age and gender, have at least one "interest" (examples of user-provided interests are shown in Figure 4.3(b)), and have at least one "post" on their wall. Most of the users divulge gender and age but only 23,950 users have at least one interest and an even smaller number of 7,054 users have any posts.

Figure 4.1(a) shows that the number of friends follows a power-law distribution in the main part; it also shows the distribution for the full data set to show that the filters do not bias the sample significantly. The sample is gender balanced (54% males and 46% females) and most

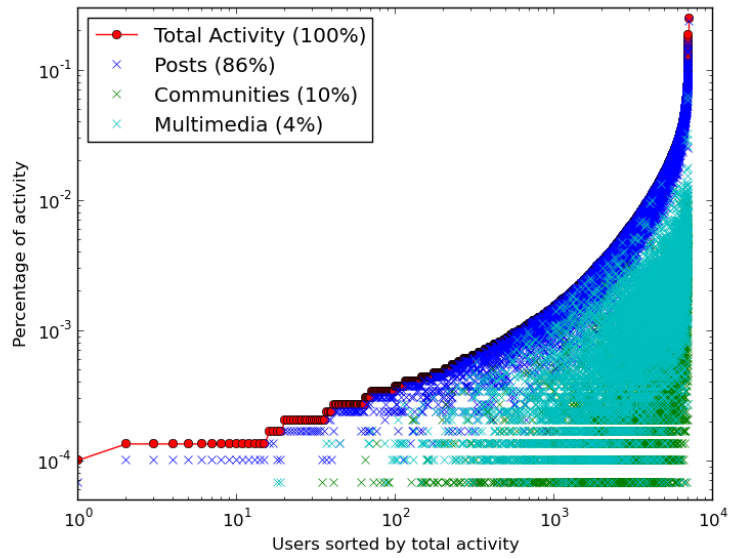


(a) Distribution of number of friends

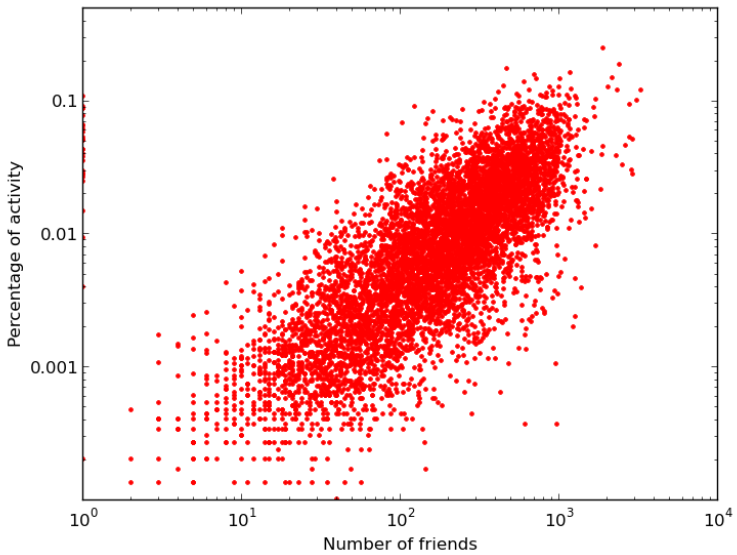


(b) Gender and age distribution

Figure 4.1: Facebook New Orleans data set: users characteristics.



(a) Distribution of activities.



(b)Number of friends vs. Activity

Figure 4.2: Facebook New Orleans data set: users' activities.

Class	Posts	Multimedia
Multimedia	0.61	1
Communities	0.50	0.37

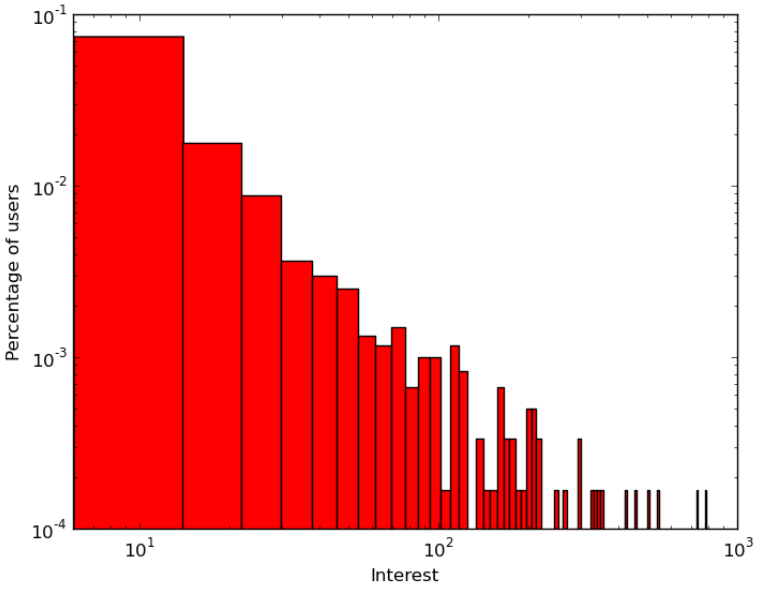
Table 4.3: Correlation between different types of actions.

users were born in the 80's, as shown in Figure 4.1(b). Figure 4.2(a) shows the total activity distribution as well as the distribution in each class. Notice that this corresponds to direct impressions. Users spend more time browsing photos than posts but these will be indirect impressions. As interests are text fields we can extract keywords. The distribution of interests is also Zipfian, as seen in Figure 4.3(a). To deal with sparsity of interests we only consider those that appear more than 5 times, obtaining 753 different keywords.

4.5.2 Choosing Weights

Facebook's users can share different types of posts (status updates, posts, urls, etc), upload multimedia content (photos, videos), and perform actions within communities (join a group, event, fan page, etc.). We group all these actions in three categories: posts, multimedia, and communities. We need a way to define the weights w_i described in Section 4.4 for each of these groups. This is difficult because the invisible actions (e.g. watching a video without leaving any comments) are unknown to external observers.

However, previous studies (see Table 4.2) have shown that visiting friend's profiles (corresponding to our "posts" category) is the most frequent activity, followed by browsing photos (multimedia), while group or community actions are infrequent. Although this ordering is consistent among different OSNs, the percentage of time spent in each category varies (e.g people spent 45% browsing photos in Hi5 vs only 16% on Facebook). Thus, we need to study the sensitivity of u_{value} to different w_i values. To understand how these weights affect the results, we can check to see if the correlation among these three groups of actions is high, indicating that the weights are less important. If



(a) Distribution of interests

Topic	Interest	Users	Topic	Interest	Users
Leisure	music	22%	Hobbies	cooking	5%
	reading	11%		photography	4%
	movies	10%		art	4%
	shopping	7%		dancing	4%
	traveling	6%		writing	4%
	friends	3%		cars	3%
	sleeping	2%		fishing	3%
Sports	sports	7%		singing	3%
	football	4%	Misc	politics	2%
	basketball	2%		money	2%

(b) Top 20 users' interests.

Figure 4.3: Facebook New Orleans data set: users' interests.

Strategy	Facebook	Orkut	Hi5	Uniform
Orkut	0.97	1	-	-
Hi5	0.97	0.98	1	-
Uniform	0.89	0.91	0.90	1
#Friends	0.47	0.48	0.48	0.50

Table 4.4: Kendall-τ correlation for the 4 strategies to assign action's weights and the Friends ranking.

user's activities are equally distributed among categories, weights are less important, because all the $\#action_i$ will be proportional.

We found a strong correlation between posts and multimedia actions but a lower correlation between multimedia and communities (Table 4.3). Users uploading significant multimedia content are thus also creating many "wall posts" but do not necessarily engage in group activities.

Using Tables 4.2 and 4.3, we define four different strategies to assign weights:

- **Facebook:** Normalizing by the three categories of actions and grouping "home","profile", and "friends" activities in the posts category, we assign 0.75 for posts, 0.21 for multimedia, and 0.04 for communities.
- **Orkut:** Considering user behaviour we assign 0.54 for posts, 0.41 for multimedia, and 0.05 for communities.
- **Hi5:** With strong correlation between posts and multimedia, and low community activity, we assign 0.5 for posts, 0.49 for multimedia, and 0.01 for communities.
- **Uniform:** We assign the same weight for the three classes.

Next, we created a user ranking, sorted by their value (using the four strategies above), and computed the Kendall-τ correlation among them. Kendall-τ is a standard measure to compute ranking similarity and its value ranges between 1 (equal ranking) to -1 (inverse ranking).

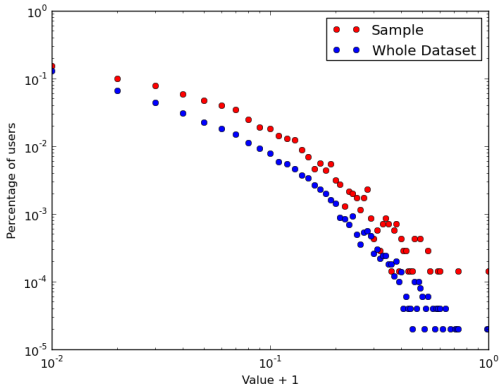
As baseline, we use a ranking based on the number of friends—the top user is the one with most friends, while the last user is the one with fewest friends. We find that the results are similar (see Table 4.4) for all the four strategies, but different from the Friends ranking, meaning that the u_{value} (Equation 4.5) does not depend only on the number friends. However, u_{value} is not too sensitive to the weights. Hence, in the next subsection, for simplicity, we use the Uniform strategy (all classes of actions are weighed equally). Although our value function is not too sensitive to the weights chosen, the values still depend on the weights and vary with the OSN under study.

4.5.3 Value Distributions

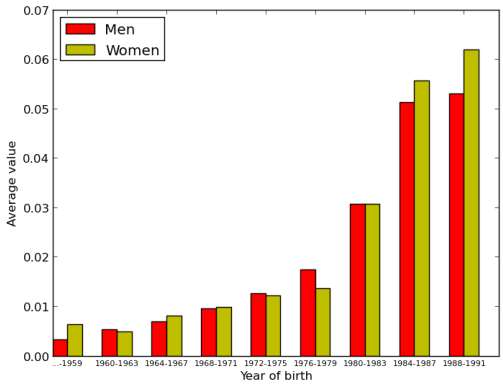
Next, we want to study how value is distributed among users and how they are related to user attributes (age, gender, and interests). Our hypothesis is that the number of impressions (i.e. value) generated by each user varies over a wide range. We expect a small fraction of users to create a lot of impressions and many users generating only a handful.

To test our hypothesis, we first compute the value for each user in our data set. Next, we normalize these values from 0 (no value) to 1 (most valuable). With this scale we have a way to compare users' value. Not too surprisingly we found that the users' value distribution is Zipfian, as shown in Figure 4.4(a), confirming our hypothesis that a small subset generates most of the impressions.

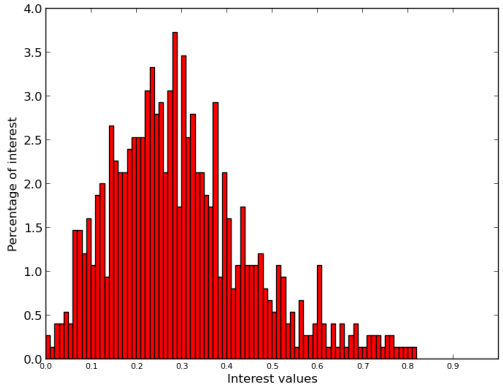
Next, we have to identify the high value users. In our experiment we are considering all the users in our New Orleans data set as the target group. Thus, t in equation 4.1 does not depend on age, gender, or interests and so u_c only depends on the number of friends. We want to compute a generic value that allows us to compare the impressions generated by different demographic groups: do women generate more impressions than men? We find that women are more valuable than men (see Figure 4.4(b)) and young people generate more impressions than mature users. Given that the difference between the least valuable group (males born after 1959) and the most valuable group (women born between 1989-1991) is less than 10%, and there is a huge standard



(a) Users' value distribution.

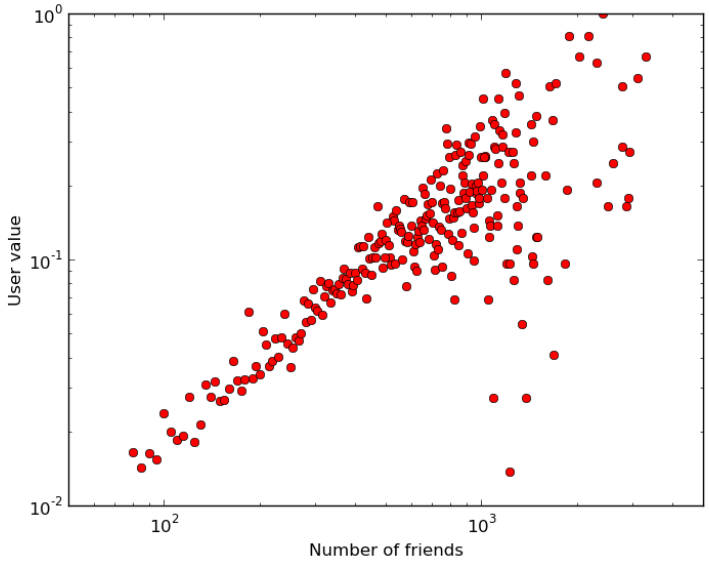


(b) Avg. value per gender and age.

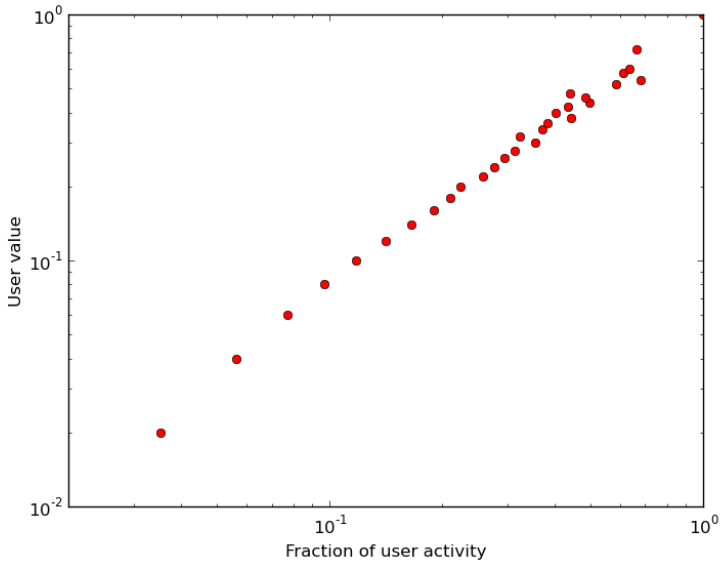


(c) Interest value distribution.

Figure 4.4: Value distributions.



(a)Number of friends vs. user value.



(b)User's value vs. number of friends and activity.

Figure 4.5: Users value vs. number of friends and activity.

Topic	Interest	Value
Leisure & Misc.	taking naps	1.00
	beerpong	0.96
	fireworks	0.92
	doodling	0.89
	graduating	0.86
	peoplewatching	0.83
	reality tv	0.81
	smiling	0.78
	Diet Coke	0.75
	success	0.73
	summer	0.73
	spooning	0.70
	making people laugh	0.69
	adventures	0.69
Hobbies	violin	0.83
	dinosaurs	0.77
	art history	0.78
	tulips	0.75
	hip-hop	0.71
Sports	The New Orleans Saints	0.81

Table 4.5: Top-20 most valuable interests. Values are normalized from 0 (less valuable) to 1 (most valuable).

deviation in each group, we can assume that high and low value users are spread in different demographic groups.

To study how user's interests are related to their value, we compute the average value per interest. The distribution of interests value is similar to a normal distribution (see Figure 4.4(c)) and the most valuable interests are heterogeneous, ranging from hobbies (such as hip-hop or tulips) to leisure activities ("taking naps", "beerpong") (see Table 4.5).

A previous study [76] suggests that there is a strong correlation between number of friends and user activity; users that post/upload more information have more friends. Our experiments show that popular users are more active and valuable (see Figure 4.5), which is partly a

consequence of using friends and activity to compute the user's value. However the correlation is much higher for the activity than for the number of friends—activity produces impressions while the number of friends is only a potential amplifier of the activity.

4.6 Discussion

OSNs provide "free" access to users in return for revenue generated by advertising impressions shown to them. We have explored how different classes of actions result in different advertising impression counts and corresponding revenue. Implicitly, our goal has been to demonstrate through our model that users on OSNs have an intrinsic value that varies with the extent of their participation on OSNs. We identify the actions that are key to generating direct and indirect impressions. We study the feasibility of our model using a data set consisting of several thousand users where we know the set of relevant actions carried out by the users. Our model is extensible, applicable to other OSNs beyond the one studied here, and adaptable to alternate revenue mapping.

The results of our study are intriguing: a small subset of actions that can be carried out on OSNs are responsible for most of the advertising impressions and a small fraction of users are key to the overall advertising revenue. Identifying the classes of users can benefit OSNs who might be motivated to provide more services for such users and advertisers who can target the more valuable users directly. More importantly, knowing the value helps users as they have a better idea of how their actions are actually valued on OSNs. We can imagine an economic *m*odus vivendi where there is an explicit trade between user's actions and profile information and the resulting service from the OSN.

Chapter 5

Finding Trendsetters in Information Networks

5.1 Introduction

In previous chapters we have pointed that people activity can be used as measure of relevance and value. In this chapter, we study a more complex scenario considering how these actions affects the flow of new ideas or trends over an information network.

Online social networks have been pointed out as places where users influence and are influenced by others, and have become ideal channels for spreading news or innovative ideas [25]. Online social networks have emerged as a popular medium where users discuss about everything, including noteworthy events, giving opinions and expressing sentiments concerning facts and ideas of daily life. Additionally, they pose opportunities for sharing information of local interest. For instance, local businesses actively reach out to their customers by announcing promotions and asking users to propagate them.

Recently, the concept of *Follower Hubs* and *Innovative Hubs* has been borrowed from the economics literature, to describe how a new idea or product propagates over such networks [1]. Follower Hubs are nodes

with high in-degree, hence they can deliver content to a larger audience than a normal user. However, previous research [60] shows that Follower Hubs usually have a high threshold for the adoption of new ideas. Here is where the Innovation Hubs are important. Innovation Hubs usually have lower in-degree than Follower Hubs, and also a lower threshold for the adoption of a new idea. Therefore, they have a key role in an information propagation process. In other words, the Followers Hubs are influencers, but the Innovation Hubs are trendsetters. Although these definitions are interesting, in traditional social experiments, it is not easy to identify the different and multiple roles of the participants without restricting the size of the study. Data collected from online information networks allow researchers to carry out detailed studies about dissemination of ideas and information with a large number of participants. Indeed, recent efforts quantified the level of influence of participants on online social networks [17] and proposed techniques to identify those who are likely to spread information to a large audience [41]. While marketing services actively search for potential influencers to promote various items, influencers actively search for innovative ideas and important innovators. In this chapter, we address the problem of identifying trendsetters in information networks.

Trendsetters are people that adopt and spread new ideas (trends, fashions) before these ideas become popular. They are not necessarily well known news outlets, celebrities or politicians, but are the ones whose ideas spread widely and successfully through word-of-mouth. To be an innovator, a person needs to be one of the first people to pick up a new or nascent trend, which may be adopted by other members of a social or information network. On the other hand, not all the early adopters are trendsetters because only few of them have the ability of propagating their ideas to their social contacts through word-of-mouth.

To identify trendsetters there are two important aspects that we need to take into account. The first one is the area or topic of the innovator, as people have different levels of expertise on various subjects. For example, marketing services actively search for potential influential people in a specific domain or area to promote certain products or services. Influential people include “cool” teenagers, local leaders, and popular public figures. Thus, it is important to specify topics and themes that define the context where trendsetters will be identified.

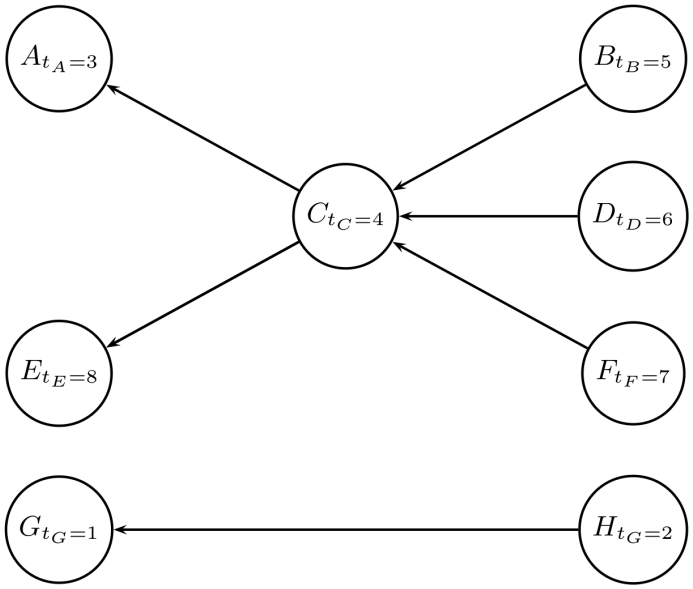


Figure 5.1: Illustrative example of timing importance: Without considering time information, nodes A and E are symmetric, regardless of whether A adopted the trend first. The edges represents social connections between nodes and the arrows goes opposite to the information flow.

Second, it is important to consider time information associated with the posting of innovative ideas. Traditional ranking algorithms on social networks, such as the standard Pagerank algorithm [63] do not consider time information concerned to ideas that become popular. Instead, they consider only aggregate usage statistics and a static network topology. Lets see at the example in Figure 5.1, where node X, $t_X = n$ represents that X adopted the trend h in time n. Thus, node G was the first one to adopt h, while node E was the last one to adopt the same trend. Note that, although node G is an innovator, it's information was passed to H but not to the rest of the network. Thus, node G cannot be considered a trendsetter. On the other hand, if we compute the standard Pagerank algorithm using this graph and ignore the time when trend h was adopted, node C would be considered the top ranked node although it has just incoming links from nodes A and E and simply spread it to a larger audience. However, if we pay at-

tention to time, we will see that C adopted the trend before E, and therefore we cannot consider that C received information from E. We can also observe that nodes A and E have the same rank according to Pagerank, despite that A adopted the trend before E. In this example, the top trendsetter is node A because it was the first one to adopt this trend being followed - directly or indirectly - by many other participants of the network, such as nodes C, D, B, and F.

This chapter presents a novel approach to identify trendsetters in information networks. Differently with previous work on ranking influential users in social and information networks, we introduce timing information on the social graph to be able to identify persons that spark the process of disseminating ideas that become popular in the network. We propose a robust way to model the dissemination of innovation, representing a topic as a collection of trends, that can be applied in several scenarios. We define a topic-sensitive weighted innovation graph that provides key information to understand who adopted a certain topic that triggered attention of others in the network. We then introduce a Pagerank inspired time-sensitive algorithm to find trendsetters. Next, we tested our algorithm using a robust data set containing the complete snapshot of the Twitter network and all tweets from 2006 to the mid-2009. The result shows that the proposed algorithm is able to measure the direct and indirect influence adding also the early adoption as a key feature to be influential. This characteristic is useful to differentiate between trendsetters and other nodes that despite having a large in-degree, adopt the trends only after they became popular.

The rest of the chapter is organized as follows. Section 5.2 reviews related work. Section 5.3 presents a formal definition of the trendsetters ranking. Section 5.5 describes the experimental evaluation and the results obtained. Finally, Section 5.6 discuss the results of this chapter.

5.2 Related Work

A detailed review of related work has been described in Section 2.2. Different from other works using a ranking based on the network topol-

ogy [17; 91; 89], our approach also considers the early adoption as a key to be a trendsetter. To the best of our knowledge, this is the first algorithm that considers Pagerank and temporal factors together to find influential people in information networks. Additionally, our algorithm presents a flexible way to model topics that is adaptable to different scenarios. Also, the influence as a function of time can be easily adjusted with a single parameter. Hence, we believe that our approach is innovative and complementary with existing approaches.

5.3 Ranking Trendsetters

This section presents our algorithm to rank trendsetters in an information network according to some topic of interest. We start with basic definitions related to the concept of a *topic*, network graphs, and the interactions among nodes over time. We represent a network as a directed graph $G(N, E)$, where N is the set of nodes and E the set of edges. Each edge is an ordered pair (u, v), $u, v \in N$, representing a relation between u and v. Furthermore, we define $In_G(v) = \{w|\ (w, v) \in E\}$, $Out_G(v) = \{w|\ (v, w) \in E\}$, the incoming and outgoing neighbor sets respectively, and $|S|$ is the cardinality of a set S.

As we are interested in ranking nodes according to a specific topic, we look only at nodes that are related with the topic. The two next definitions formalize what a topic is and how to select the nodes.

Definition 5.1. *We define a topic as a collection of trends related to a specific theme. We denote this collection by* $\{h_1, \ldots, h_{n_k}\}$. *Each one of the* n_k *trends could be a word, a phrase, a meme, a tag, an URL, or any other kind of label that can be associated with a node.*

Definition 5.2. *We denote* $G_k(N_k, E_k)$ *as the induced graph of* $G(N, E)$ *over the topic* k. *The set* N_k *is obtained by considering all nodes of* N *that used at least one trend of* k *and* E_k *represent all edges* (u, v) *such that, if* $(u, v) \in E$ *and* $u, v \in N_k S$, *then* $(u, v) \in E_k$.

As mentioned in the introduction, the timing information is the key to determine social influence. We include this information in the temporal attributes of nodes and edges of $G_k(N_k, E_k)$ in the following definition.

Definition 5.3. *Let $t_i(v)$ be the time when node $v \in N_k$ adopts the trend $h_i \in k$ ($t_i(v) = 0$, if v does not adopt h_i). We define two vectors, $s_1(v)$ (for all $v \in N_k$) and $s_2(u, v)$ (for all $(u, v) \in E_k$), each one with n_k components given respectively by:*

$$s_1(v)_i = \begin{cases} 1, & \text{if } t_i(v) > 0, \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

and

$$s_2(u, v)_i = \begin{cases} e^{-\frac{\Delta}{\alpha}}, & \text{if } t_i(v) > 0 \text{ and } t_i(v) < t_i(u), \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

for $i = 1, \ldots, n_k$, where $\Delta = t_i(u) - t_i(v)$ and $\alpha > 0$.

Vector $s_1(v)$ informs if node v adopted (or not) each trend of k, while $s_2(u, v)$ shows if u adopted these trends after v and weights the relation as a function of the period of time between $t_i(u)$ and $t_i(v)$. For a fixed α, if $\Delta \to 0^+$ then $e^{-\frac{\Delta}{\alpha}} \to 1$ and if $\Delta \to +\infty$ then $e^{-\frac{\Delta}{\alpha}} \to 0$. These limits mean that if the node u adopts a trend just after v then $s_1(v)_i$ is very close to $s_2(u, v)_i$, and, on the other hand, if u adopts the trend after a long time, we have that $s_1(v)_i$ and $s_2(u, v)_i$ are very different. The exponential time decay to compute influence has been proposed in previous work related to temporal factors in the web graph [7].

The α parameter allows to control the time window that will be considered to compute $s_2(u, v)$. This settable parameter is useful because it can be adapted to different scenarios. Depending on the nature of the problem, we want to consider that one node is strongly influencing another if the second one imitates the first one in few seconds, and in other cases we want to use a longer span of time. Moreover, for the same problem we could be interested in studying the influence of a short span of time, or a long term influence.

So, when many components of $s_1(v)$ and $s_2(u, v)$ are similar and different from 0, we assume that v has a strong influence over u according to a topic k. Based on the previous definitions we can now define influence:

Definition 5.4. *Let $G_k(N_k, E_k)$ be an induced graph of a network $G(N, E)$ over a topic k with n_k trends. For each $(u, v) \in E_k$ we define the influence of v over u by:*

$$I_k^*(u, v) = \left(\frac{s_1(v) \cdot s_2(u, v)}{||s_1(v)|| \times ||s_2(u, v)||} \right) \times \left(\frac{L(s_2(u, v))}{n_k} \right), \tag{5.3}$$

where the operator $\cdot$ refers to the scalar product, $||x||$ to the Euclidian norm of any vector x, and $L(s_2(u, v))$ to the number of components of $s_2(u, v)$ that are different from 0. If $||s_2(u, v)|| = 0$, we define $I^*(u, v) = 0$. It is important to notice that, by definition, $||s_1(v)|| \neq 0$ for all $v \in N_k$.

Equation 5.3 is the main outcome of our previous discussion. The first part is given by the cosine similarity between $s_1(v)$ and $s_2(u, v)$, which is close to 1 if u adopted the same trends than v in a reasonable lag of time, and close to 0, otherwise. The second term is the fraction of trends of k that u adopted after v. We use this to indicate that if u adopted more trends influenced by v than by other node z, then the influence of v over u is greater than the influence of z(see Figure 5.2).

One important fact is that u can be influenced to adopt a trend of k by several nodes in $G_k(N_k, E_k)$. So, we normalize $I^*(u, v)$ as follows:

$$I_k(u, v) = \frac{I_k^*(u, v)}{\sum_{w \in Out_{G_k}(u)} I_k^*(u, w)}, \tag{5.4}$$

noticing that if the denominator of Equation 5.4 is zero, we define $I_k(u, v)$ as 0.

The next definition presents how we rank trendsetters according to a Pagerank-like algorithm.

Definition 5.5. *The trendsetters (TS) rank of node v in a network $G_k(N_k, E_k)$, denoted by $TS_k(v)$, is given by:*

$$TS_k(v) = d\ D_k(v) + (1 - d) \sum_{w \in In_{G_k}(v)} TS_k(w) I_k(w, v), \tag{5.5}$$

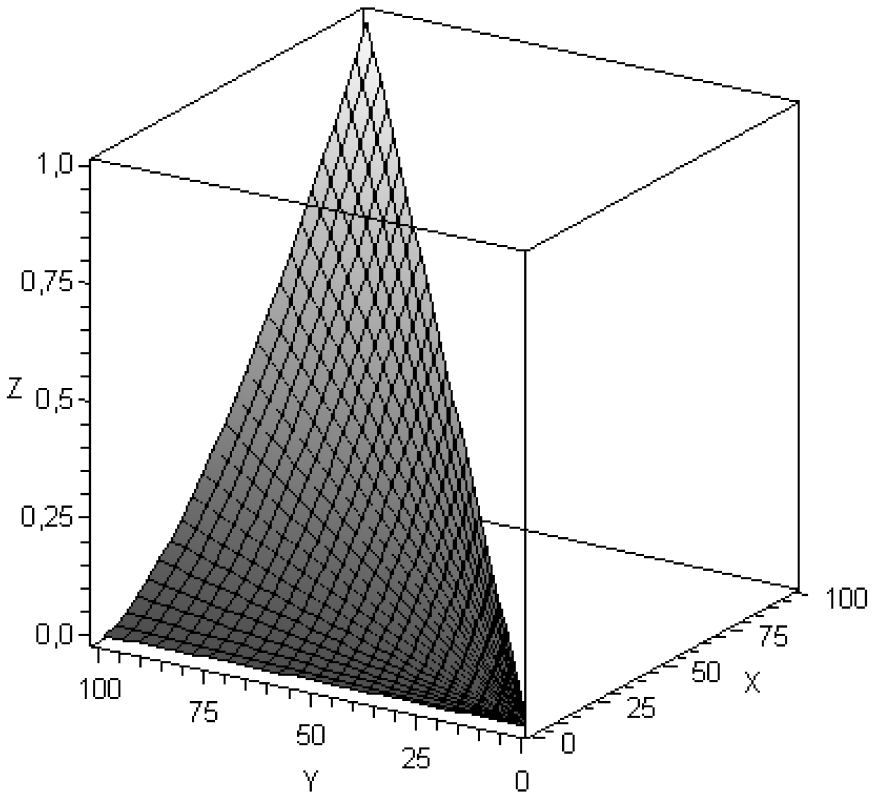


Figure 5.2: $I_k^*(u, v)$ for a topic k with 100 trends. Axis Y represents the number of trends adopted by v, while axis X is the number of trends adopted by u after v, and axis Z is the influence $(I_k^*(u, v))$ of v over u, in the topic k. For simplicity we use in this case $\frac{\Delta}{\alpha} = 0$.

where $0 \leq d \leq 1$ *is the damping factor and* D_k *is a probability distribution over all nodes of* $G_k(N_k, E_k)$*. Here we consider a uniform distribution that is* $D_k(v) = 1/|N_k|$ *for all* $v \in N_k$*, but this distribution could be topic dependent.*

Making an analogy with the random surfer model in the Pagerank algorithm presented in [63] on graph $G_k(N_k, E_k)$ we can analyze Equation 5.5 in the following way: consider that the surfer is in any node of $G_k(N_k, E_k)$, for example u. With probability $1 - d$, the surfer leaves u and goes to other node in $Out_{G_k}(u)$, and with probability d, to any

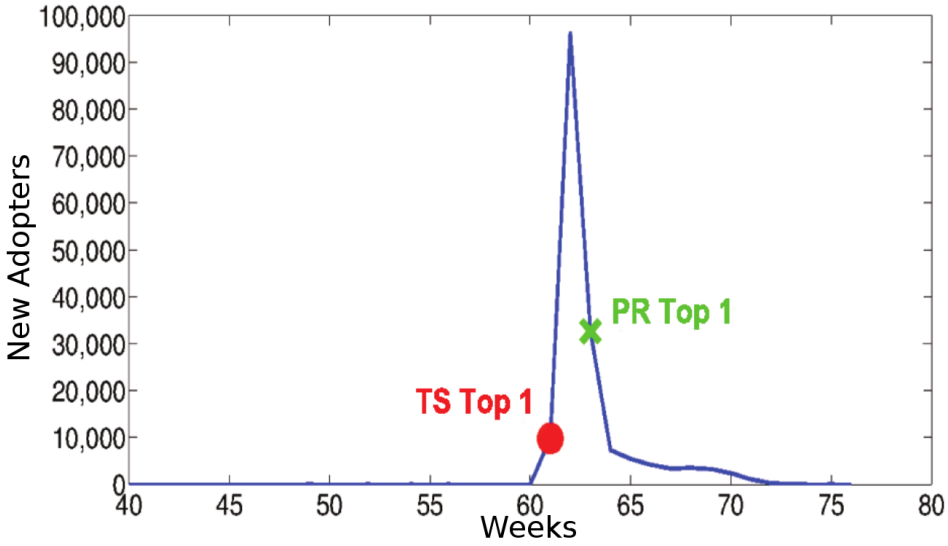


Figure 5.3: Iran Election topic timeline of new adopters: comparison between top TS and top PR users.

node in N_k. In the first case the node $v \in Out_{G_k}(u)$ will be visited with probability $I_k(u, v)$. So, the node that influences more u has a higher probability to be visited. In the second case, any $v \in N_k$ will be visited with probability $D_k(v)$, reflecting the independent adoption of that topic. Hence, in the steady state the surfer will spend more time in the most influential nodes of the network.

5.4 Examples

Let us now see some examples in Twitter. For all of them, we have set $\alpha = 1$ *day*, and the dumping factor $d = 0.2$. For more details about the implications of these parameters, please refer to Section 5.5.6.

The Iran election was an important topic during 2009. The main hashtag used to talk about this event was #iranelection, and other related tags where the #iran and #tehran. So using definition 5.1, the topic Iran Election could be represented by: $k_{\text{Iran Election}} =$

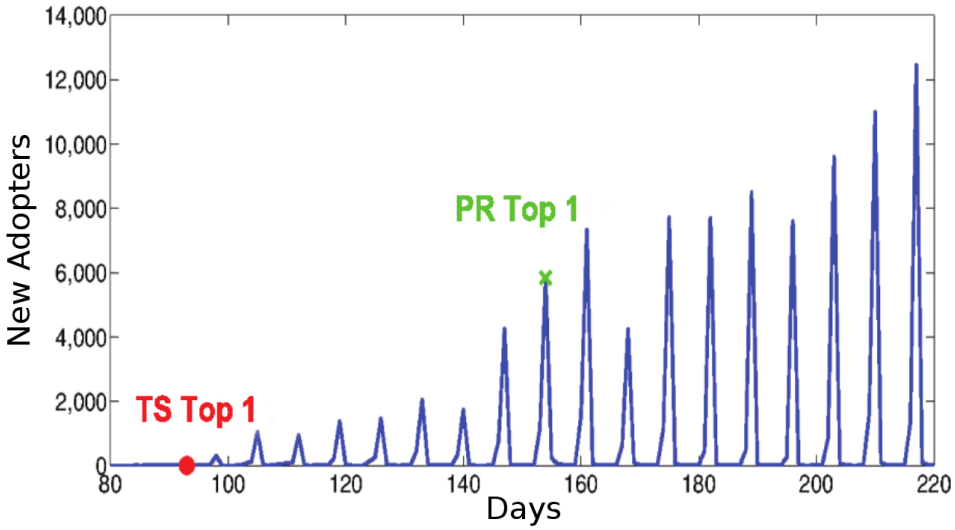


Figure 5.4: #musicmonday topic, timeline of new adopters: comparison between top TS and top PR users.

[#iranelection, #iran, #tehran]. Following the methodology proposed previously, we compute the graph for all the nodes that used at least one of the trends. Next, we compute the PageRank(PR), InDegree Rank(ID), and the TS rank for this graph. PR and ID selects @cnnbr (CNN Breaking News) as top user, while TS selects a user named @Lara, self-described as "Reporting from the Middle East for ABC News and Bloomberg Television." This user twitted with the two most popular hashtags in this topic, adopting them before than they became popular (see Figure 5.3). In other cases related with politics we have also find other activists or reporters "on site" being ranked on the top of TS, while PR and ID selects CNN for all of them.

Another interesting example is the idiom #musicmonday, which is one of the most popular in Twitter. Its name is very descriptive because it is used to share music on Mondays. Considering that we have almost the complete Twitter information from its beginning, we can know who invented this tag: *@rubenharris*. This fact should identify that user as influential in this trend. However, if we analyze this topic using PR, that user is ranked in the position 164,970 among 179,119 users.

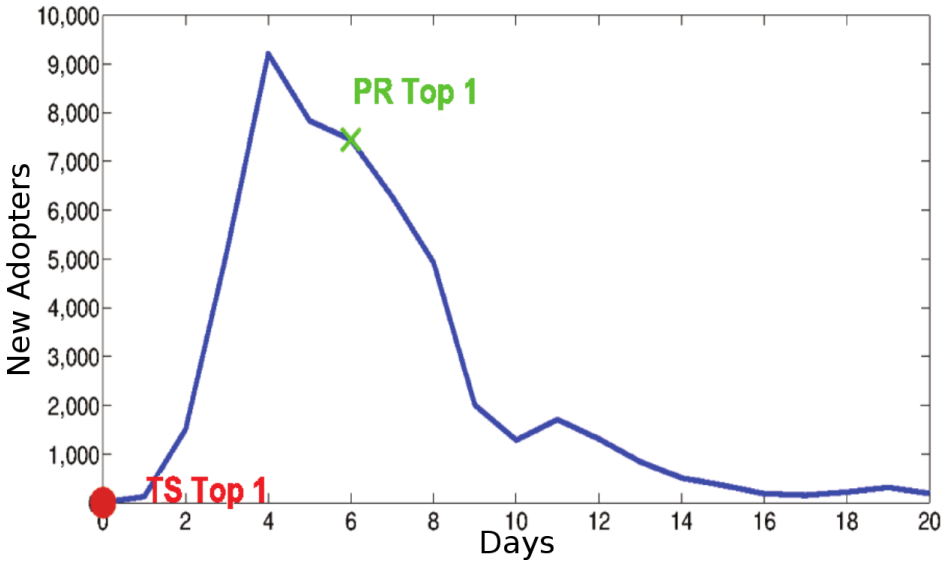


Figure 5.5: Swine Flu topic, timeline of new adopters: Comparison between top TS and top PR users.

But using TS this user appears in the 4th position. Note that, as we explained before, to be an innovator - and thus an early adopter - a necessary but not sufficient characteristic is to be a trendsetter. TS considers that the most trendsetter for #musicmonday is user *@twtfm*, which is the corporate user of the site http://twt.fm, a company that offers a service to share and search music using Twitter. However this user was the 76th to adopt the trend. Now, note that PR and ID selects the same user as the top one: *@perezhilton*. He is a famous professional blogger, and he has been indicated as one of the users with more followers in Twitter [46]. However, if we check *@perezhilton*, he was the 18,718th user to adopt the trend, therefore we cannot consider he as an innovator. However, because he is prestigious, TS ranked it in the top-100. The example of #musicmonday is useful to understand that TS captures both characteristics that we consider important for a trendsetter: the early adoption and the capacity to spread the trend over the graph.

As final example, we consider a complete different topic: the hasgtag

#swineflu was used to alert and give information about this epidemic in 2009. The top user for the traditional Pagerank is Stephen Fry (@stephenfry), a famous blogger/activist. The top for ID is *@mashable*, another well-known professional blogger. For TS the top user is *@CDCemergency*, that is, the official account from the "Centers for Disease Control and Prevention", a public agency of the U.S.A government. In this case the hasthtag inventor, that is the most innovator, was also successful. As we can see in Figure 5.5, the behavior is similar to the case of #iranelection, the top user from PR started to use the hashtag after it became popular but the top user of TS proposed the trend just before started to grow.

The results of these three examples shows that TS capture the behavior of trendsetters.

5.5 Experimental Evaluation

In order to test the TS ranking we have conducted a set of experiments over a huge Twitter data set. We consider the Twitter social graph, where the connections among users are directed. Using the notation described in the previous section the Twitter graph will be $G(N, E)$, where an edge $(u, v) \in E$ means that user u *follows* v. Trends are modeled using *hashtags*, so a topic k is a collection of hashtags, that is $k = [\#tag_1, \ldots, \#tag_{n_k}]$. Next, we create the induced graph G_k considering all the nodes that have posted at least a tweet with one hashtag of k. Over this graph we compute the TS ranking using a time window of one day ($\alpha = 86,400$ seconds) in Equation 5.2 and $d = 0.2$ in Equation 5.5. We have tested other values and they do not change the results significantly.

Our hypothesis is that other measures of influence are not suitable to find trendsetters because they tend to favor nodes that do not propose new trends, but follow those that are already popular. To test this, we grouped the trends by different methodologies: first we group them by categories related with topics such as music, sports, movies, etc., next we grouped the new adopters curve shape. In each case we compare the TS ranking with In-Degree (ID) ranking - where nodes are sorted

by incoming links; and the traditional Pagerank, PR. We also quantify the followers influenced by the top users of each ranking and we compute how similar are the rankings under study.

5.5.1 Data Set

We have used a data set containing almost the total information in Twitter until August 2009. We have over 50 millions users, with all their social connections (*Followers* and *Followees*) and approximately 1.6 billions of Tweets. Note that differently from other works that use a big amount of tweets, we also have the complete social graph, so we do not need to use heuristics to infer it. A detailed description of this data set can be found in [17].

To select the hashtag for the experiment, we use the classification made by Romero *et al.* [73], where each of the 500 most popular hashtags in their dat aset was assigned to a category such as politics, music, or celebrities (see Table 5.1). From those 500 hashtags, only 370 are mentioned among the 2,000 most popular of our data set, with #followfriday being the most popular with 3,051,316 mentions and #jemi the least mentioned, with only 1,810 occurrences. The 130 remaining hashtags do not appear or have a very low level of mentions. This is because each data set was obtained on different dates.

To complete the topic modeling, we looked for other hashtags related with the main one. For the 370 hashtags we searched for others that had a co-occurrence of at least 5% with the main one. This means that each of the 370 topics is modeled by a vector containing the main hashtag and others related to it. For example, the topic modeled with more hasthtags was #realstate, having other 20 related hashtags. Over 74% of the topics were modeled with at least two hashtags. Note that these related hashtags creates the vectors that are described in Definition 5.2. The categories are used only to facilitate the analysis of the results.

Category	#Topics	Example of Hashtags	#Tweets
Celebrity	16	#michaeljackson, #niley	1,036,101
Games	13	#mafiawars, #ps3 #	2,556,437
Idioms	35	#musicmonday,#followfriday	7,882,209
Movies	29	#heroes,#tv	1,769,945
Music	33	#lastfm, #musicmonday	2,785,522
None	153	#quotes, #sale	2,227,971
Political	39	#honduras, #Iranelection,	8,156,786
Sports	27	#soccer, #rugby	1,914,061
Technology	41	#twitter,#android	7,459,471
Total	370	-	41,442,741

Table 5.1: Summary of categories. Note that some hashtags and hence tweets can belong to more than one topic.

5.5.2 Adoption before Peak: Categories

Our first approach to answer this question is to analyze the percent of users of each ranking adopting the trend before the peak of adoption. By peak of adoption we refer to the time when a trend had its bigger number of new adopters. To do that, first we obtained the peak of adoption from each of the 370 topics studied. We denote by P_k the peak of adoption of trend k. Next, we have to compare it with the time of adoption of the top-p users of each ranking, where $T(i)_{k_r}$ represents the time in which user i from ranking r adopted topic k. Therefore, if $P_k - T(i)_{k_r} < 0$, this means that user i adopted the trend before the peak. Finally, we grouped all the topics k by their category, and we computed the percentage of users adopting the trend after the peak for each ranking. The results presented have been calculated with $p = 100$, but values from 5 to 1000 do not present significant variations.

Figure 5.6 shows that in categories such as music, celebrity, and idioms, most of the nodes in the top of ID and PR start talking about these topics after the peak, and only in sports and technology they obtain a good performance. In contrast, in 6 of the 9 categories, more than 50% of the TS top users adopted the trend before the peak.

One motivation to develop the TS ranking was our intuition that nodes

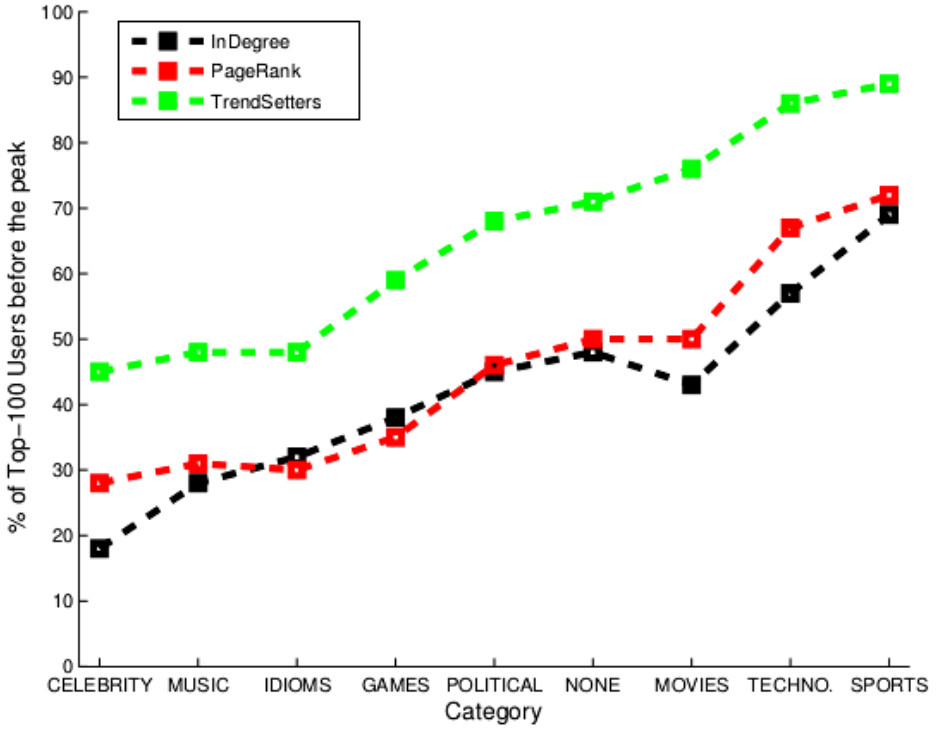


Figure 5.6: Percentage of top-100 users of each ranking that adopts the trend before the peak.

with high in-degree do not propose or push new trends but follow those that are already popular. The performance of ID in the previous experiment tends to confirm this intuition. In order to better understand how the in-degree is related with the adoption time, we repeated the previous experiment, but instead of computing only if $P_k - T(i)_{k_r}$ is > 0, we recorded this time span as well as the user i in-degree. Therefore, for each ranking in each trend we have a list of tuples representing the time span and the in-degree of the top-p nodes. Again we group the trends by their category, and at the end we computed the median of the time span and of the in-degree of each ranking for each category. In Figure 5.7 we plot the time span median in the horizontal axis and the in-degree median in the vertical axis for each ranking. It is clear that top users of TS adopt trends before the peak and they have a smaller in-degree than top users of the other rankings. These results

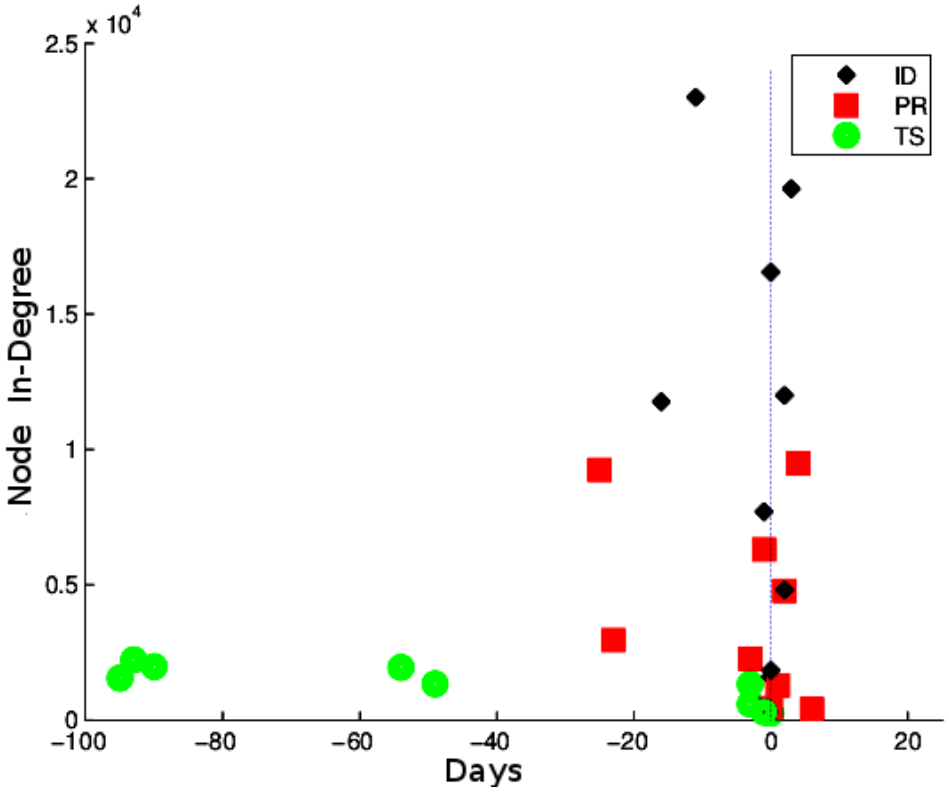


Figure 5.7: Relation among time span and in-degree for the top-100 users of each ranking in all the categories (peak is at time 0).

confirm that nodes with high in-degree tend to be *slower* in adopting a trend than other nodes.

5.5.3 Adoption before Peak: Shapes

The categories by topic are very descriptive, but the nature of the TS ranking suggests that the quality is also related with the shape of the curve of adoption. For this reason we grouped the trends by their curve of new adopters. For this aim we have used the KSC-algorithm [90], that receives as input a set of time series, and gives as an output a classification by shape and the centroids of each cluster, providing a

#Cluster	1	2	3	4
#Topics	91	115	**128**	36

Table 5.2: Number of topics in KSC Clusters.

visual representation. Note that, unlike the previous cited paper, we are interested in the adoption of the trend, that is, the first time that the hashtag is mentioned by a user. Additionally, we are interested in all the popular trends, not only in those with a short duration. Hence, our time discretization is done by days, not hours.

To apply the KSC-algorithm we have created a time series to represent the curve of new adopters from each trend, considering that the time of adoption is the first mention of any of the hashtags in the topic (repeated or later mentions are not taken in account). We have created time series of 128 elements, where each element represents a day. Next, we align the peaks of all the time series at 2/3, that is position 86. This is different than the 1/3 peak centering used in [90]. Our reason to move the peak to the right is because we are more interested in what happens before the peak, rather than later. The last step, was to select the number of clusters K to use. We tried with values from 2 to 12, finding that with more than 4 clusters we found only small variations of the 4 main clusters. Figure 5.8 shows the shapes of the 4 clusters obtained. Next, we repeatead the calculations described in the previous subsection to find the time span and the median of time adoption of the top users of each ranking but now grouped by cluster. Finally, we plot these points over the curves.

Table 5.2 show that most of the topics corresponds to clusters with a clear peak of adoption such as cluster 3. In figure 5.8 we can see that TS appears clearly before the peak in all the clusters. In contrast, the top users of ID and PR only appears before the peak in cluster 2, that is, in the cluster with the less pronounced peak. Specially interesting are the results for cluster 1, where TS appears over a little first peak before the largest one. This suggests that TS is detecting a topic that will be potentially interesting in the future.

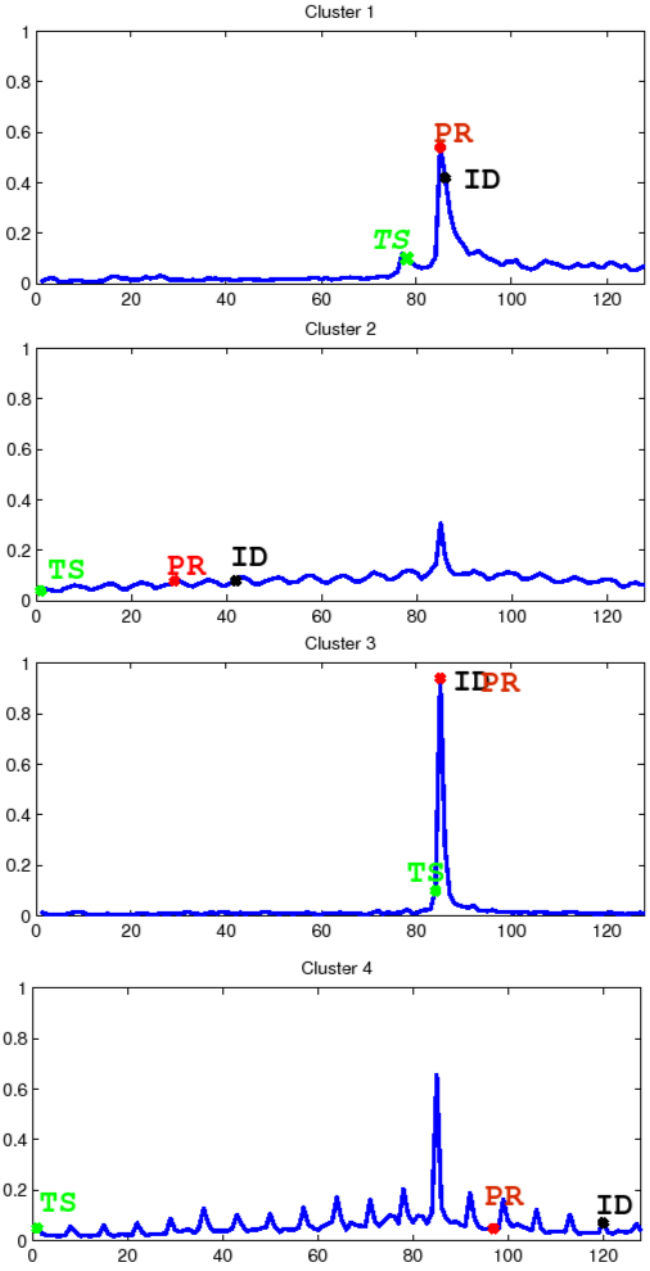


Figure 5.8: KSC clusters. Each ranking is represented with the median of time deviation with respect to the peak in each cluster.

Category	ID (%)	PR(%)	**TS(%)**
Political	0.013	0.084	**0.174**
Music	0.013	0.096	**0.160**
Celebrity	0.015	0.089	**0.148**
Games	0.022	0.058	**0.115**
Sports	0.004	0.054	**0.098**
Idioms	0.001	0.034	**0.088**
None	0.011	0.001	**0.085**
Technology	0.006	0.054	**0.078**
Movies	0.006	0.043	**0.067**

Table 5.3: Influenced Followers (IF) ratio for top-100 users of each ranking.

5.5.4 Influenced Followers Ratio

Now we try to understand how many of the social contacts were influenced by the top users of each ranking, that is, how many of their total followers adopted the trends after them. To evaluate this, we create a simple indicator that we call *Influenced Followers* ratio for a topic k, $IF_k(v)$, defined as the fraction of followers of v that adopted at least one trend of the topic k after v.

Table 5.3 shows that TS top users have a bigger IF ratio than for PR and ID. It is interesting to note that in the category Political, the TS rankings obtain the best performance, and in all of them is always over 0.06. Note that in 7 of the 9 categories TS is one order of magnitude better than ID and almost doubles PR. This confirms that TS users influence more their social contacts than other rankings.

5.5.5 Ranking Similarities

Whereas one of the main features of TS ranking is to capture the early adoption behavior, makes sense to compare it with an Early Adoption EA ranking. That is, a ranking where the top one will be the first adopter and the next position will be assigned by the adoption time.

	EA	PR	ID
PR	0.11	-	-
ID	0.09	0.74	-
TS	0.37	0.56	0.48

Table 5.4: Kendall-τ comparison among rankings.

To compare rankings we use the Kendall Rank Correlation Coefficient τ. This coefficient gives an idea of the agreement between two rankings. It gives a value in the interval $[-1, 1]$, where 1 means total agreement and -1 means that one ranking is the reverse of the other, similarly to standard correlation.

For this experiment we use the trends with more mentions in each category and then we compute the average among trends for the four rankings: EA, TS, ID, and PR.

Table 5.4 shows that PR and ID tend to be similar but completely different from EA. TS is not too similar with any of them but presents a nice balance among them. This results shows that TS has the ability to mix different characteristics of the other rankings. It also shows that not all the early adopters are trendsetters.

5.5.6 Damping Factor and Time Window

This section aims to evaluate the importance of α and damping parameters, introduced in Equation 5.2, in the TS proposed in Equation 5.5. Keeping α fixed we analyzed the importance of the damping factor for the examples given in Section 5.4. We used Kendall-τ to compare the rankings obtained for different values of damping factor between 0.1 and 0.9, for all comparisons the Kendall-τ values were always over 0.9 meaning that does not exist important variations on the ranking.

Next, we repeat the same procedure using different values of α. Figure 5.9 shows that the biggest changes are the result of varying α. To study this, for each trend we computed the TS for 3 values of α, using a fixed value $d = 0.20$, finding the top-1 in each case and we measured the Kendall-τ correlation among these three different rankings.

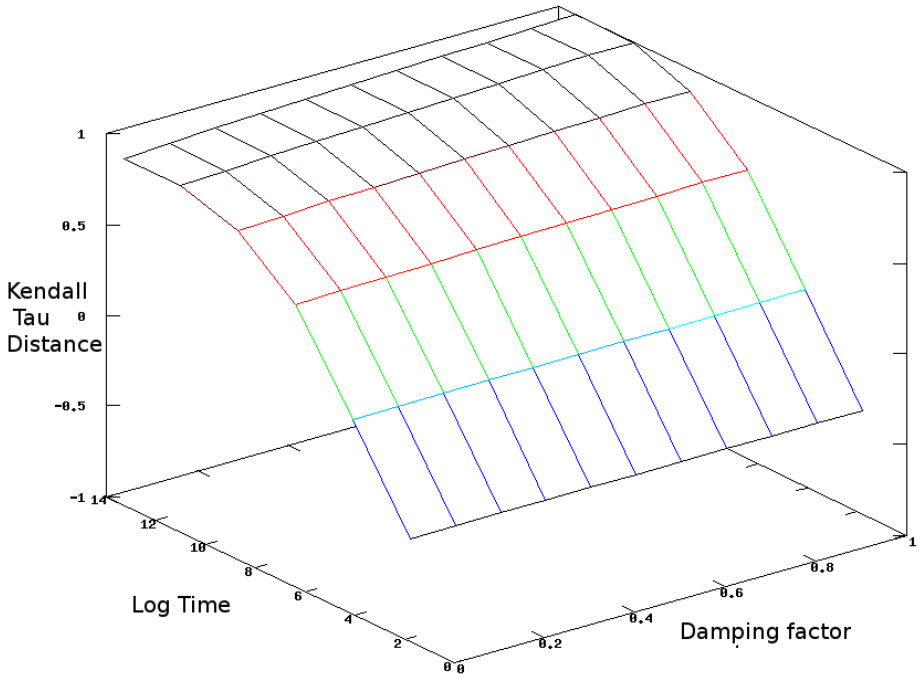


Figure 5.9: Example of variations on TS ranking depending on α (log Time) and Damping factor (d) parameters (see Equation 5.5) for Iran elections. First we compute the TS ranking using $\alpha = 1$ *week* and $d = 0.9$, and next we compare it (using the Kendall-τ value) with the result obtained for different values of α and d

Table 5.5 presents the top user of TS for each situation. Using a short time window ($\alpha = 1$ minute), we can see that users ranked in the top tend to have a high number of followers. For the trend #iranelection the user @BreakingNews has more than 1 million followers and is a news account which posts tweets at high rate. For #musicmonday the top-1 is a band with more than 500 thousand followers, while for #swineflu is stephenfry, a blogger/activist, already mentioned before.

When we increase the value of α the results change in a strong way. For #iranelection we have the user @shahrzadmo as the top-1 user, an activist that we have described in the previous section, with around 2,000 followers. For #musicmonday with $\alpha = 1$ hour, the top-1 is

α	Trend		
	#iranelection	#musicmonday	#swineflu
1 min	BreakingNews	Jonasbrothers	stephenfry
1 hour	shahrzadmo	PerezHilton	CDCemergency
1 day	shahrzadmo	twtfm	CDCemergency

Table 5.5: Top-1 TS for different α values.

$\alpha_1 \times \alpha_2$	Trend		
	#iranelection	#musicmonday	#swineflu
1 min × 1 hour	0.126	0.075	0.158
1 min × 1 day	0.087	-0.139	0.122
1 hour × 1 day	**0.831**	**0.604**	**0.797**

Table 5.6: Kendall-τ of TS for different α values.

@PerezHilton (famous blogger with high number of followers) , and for $\alpha = 1$ day the top-1 is *@twtfm* (an account for sharing music on Twitter with more than 1 million of followers). For #swineflu the top-1 is *@CDCemergency*, a governmental agency account with more than 700 thousand followers, for both $\alpha = 1$ *hour* and $\alpha = 1$ *day* .

These results suggest that users with higher in-degree, such as celebrities or professional bloggers and news tend to have impact in a short span of time. So, their followers decide quickly if they will adopt or not the trend proposed from them. If they don't adopt the trend, probably they forget about this. On the other hand, activists or bloggers tends to have a more close relation with their followers, so they have more time to decide if they will adopt the trend, with slower spread but more impact along the time.

Table 5.6 presents the Kendall correlation among the TS computed for each trend. We can see that the rank with $\alpha = 1$ minute has a low correlation with the others, indicating that changing this parameter from small to large values produce strong changes in the final results. On the other hand, the correlation tends to increase when the α parameter also increases. This is an expected result, once the components of the s_2 vector tend to 1 when α is large.

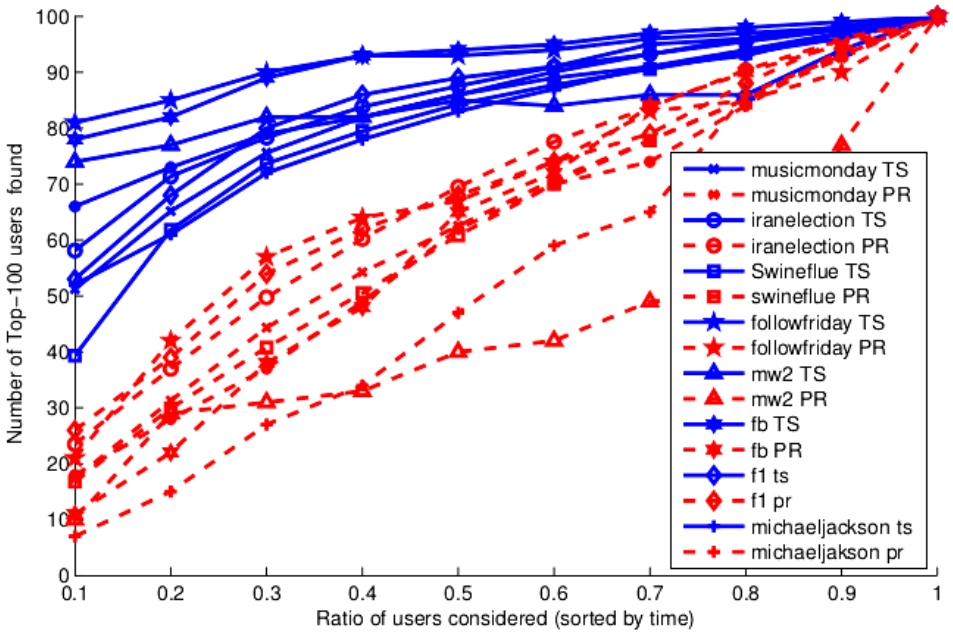


Figure 5.10: Number of top-100 users found using only a fraction of total users sorted by time, comparing PR and TS in eight trends.

5.5.7 Ranking with Partial Information

Previous results show that TS give high scores to the early adopters. Considering this we can conjecture that probably it is not necessary to use the information about all users to find the top ones. To answer this question we conducted the following experiment: first we selected the topic with more users for each category to use it as representative of this category. Next, we ordered the users by adoption time and then we compute the ranking considering only the first 10% of them, then 20%, and so on in increments of 10%. Next, we compared it with the final ranking (i.e. with the 100% of users). For all the trends, we were able to find the top-one user, considering only the initial 10%. Moreover, we found 7 of the top 10 users, and more than 70 of the top-100 users if we consider the 20% of initial users. In contrast, PR could find the top-one with the 10% only in one case, requiring more than 50% of the users in the other cases. Moreover, PR required over

60% of users, to find at least 7 of the top-10 users.

Figure 5.10 shows the total number of top-100 found considering different fractions of users. These results suggest that TS is able to find the most influential users faster than PR. This behavior can be explained considering that TS ranks on the top many early adopters, unlike PR that is more sensitive to the arrival of a node with high in-degree at anytime that they arrive. The time decay used makes TS less sensitive if those nodes arrive late.

5.6 Discussion

Although we have conducted experiments only on Twitter, the problem formulation makes it possible to apply the trendsetters ranking to other information networks.

One important finding is that users with high in-degree do not propose the ideas that became popular, as usually they adopt them when they are already popular. This confirms the importance of developing new techniques such as TS to find the users that create or early adopt these trends. This appear to be critical in topics related with celebrities, music, idioms and politics.

The results presented in Section 5.5.7 highlight two important advantages of TS over other algorithms. First, the possibility to find a big fraction of trendsetters requiring only the first 10% of trend adopters, something very useful in real-time scenarios. Second, the differences in the behavior along the time of TS against PR results could be explained because when nodes with a high in-degree adopt a trend late, the time decay function reduces their impact in the final rank.

Chapter 6

Finding Relevant Users considering Mobility Patterns

6.1 Introduction

In this chapter we go beyond user activity and consider user mobility patterns. The high penetration of mobile devices with GPS capabilities, are allowing to collect information about users' location. Therefore, this gives an additional context for finding relevant people. Specifically, we take a recommender system approach, to find potential customers for local shops.

Social media sites have been recently testing features that return lists of people ("guests") that users might want to consider inviting to their events (e.g., law firm parties, birthday parties, *PR*'s club invitations) [22]. Guests are selected based on relevance to the event and to the other fellow guests.

The problem of predicting relevant "guests" for venues or events has thus started to receive attention on the Web but has not been fully explored on mobile-social media platforms such as Foursquare, as discussed in Section 6.2 on "Related Work". One way of recommending venues to people is to use existing Web-based collaborative filtering

algorithms. In Section 6.3, we show that such algorithms are not effective, mainly because of data sparsity: a venue is visited, on average, by very few users. Therefore, we propose two simple techniques for "recommending guests" that are reasonably accurate and scalable, and whose recommendations are easy to explain. Here, we make two main contributions:

- We put forward three proposals - two Bayesian models and one linear regression - that incorporate domain knowledge from the literature of human mobility and that cope with data sparsity (Section 6.4).
- We evaluate how the models perform against Foursquare data for the whole city of London (Section 6.5). We find that the simplest model - linear regression - returns the most accurate recommendations for all types of venues.

Finally, we discuss some open questions (Section 6.6), including that of when our models do *not* work, and consequently, the limitations of our approach.

6.2 Related Work

The problem of recommending events has been initially tackled on the Web. In this context, researchers have mainly worked on detecting and tracking events [38; 42]. They initially considered how textual content evolves over time and left out network effects. Zhu and Sasha [92] then started to model social interactions and topic evolutions by treating these two elements separately. More recently, Lin *et al.* [51] built a model that considers these two elements simultaneously and showed that it worked upon two very different types of data - Twitter and DBLP. After detecting events, one can then recommend them. That is what Daly and Geyer *et al.* [21]did: they built a system that recommends events in an internal event management service and proposed a new way of recommending events to new users. Before that, Minkov *et al.* [57] had run large user studies in which they evaluated the effectiveness of different strategies for recommending academic talks. They

found that, in a situation of limited data sparsity, collaborative filtering approaches work better than content-based ones. The recommendation process generally relies on user ratings but has also been enriched by social networks at times. A case in point is Golbeck *et al.* [28] who built a recommender system that integrates social networks to offer well-informed movie recommendations.

Hence, past work on recommending events has mostly gone into web platforms, while mobile ones have been investigated only recently. Takeuchi and Sugimoto proposed [83] a system that recommends shops based on past visited locations, and found item-based collaborative filtering to work reasonably well. Ricci and Nguyen [70] proposed a system that recommends nearby restaurants using a critique-based model. More recently, for major mobile social-networking services, Scellato *et al.* [79] studied their geographic properties at scale and suggested that these properties could well inform venue recommendation in large cities. Upon mobile phone data in the metropolitan area of Boston, Quercia *et al.* [68] studied strategies for recommending large-scale events (e.g., concerts, baseball matches) and showed how different types of events require different recommendation strategies.

Shifting attention from recommending events to recommending people, one sees that most of the work has again gone into web platforms. Within an enterprise social network, Guy *et al.* [35] proposed ways to recommend people a user is not likely to know but might be interested in. Few months ago, Facebook launched a new feature called "suggested guests" [22]: this returns a list of people (three at the time) a user might want to consider inviting to their event, and the list is compiled based on relevance to the event and to the people who are attending. Since work on recommending people for events has just started on the Web, it comes as no surprise that little work about it has gone into mobile social-networking platforms.

6.3 Collaborative Targeting: Unfit

To begin with, we state our research problem: Given a venue (e.g., Italian restaurant), select individuals who are likely to visit it.

Category	#Venues	#Users
food	1,293	1,566
nightlife	1,075	1,207
travel	850	1,744
home/work/etc.	411	1,037
shops	362	878
arts&entertainment	348	841
parks&outdoors	184	363
education	49	117
Total	4,572	3,110[1]

Table 6.1: London Foursquare Data. Number of users and venues across venue categories.

This simple problem, if solved, might enable a variety of applications, which include target advertising, commercial property evaluation, and social marketing (as we shall discuss in Section 6.6).

The problem might be formulated in simple "recommender system" terms - that is, it is the problem of how to recommend venues (items) to people (users). One way of solving it is to run a state-of-the-art matrix factorization algorithm on the inverted *venue-by-people* matrix (whose value m_{ij} is 1, if user j checked-in in venue i; 0 otherwise) and obtain, for each venue, a list of people who might like to visit it. We do just that: we use the state-of-the art *Implicit SVD* method introduced by Hu, Koren and Volinsky [37] and implemented it within the Mahout framework. To evaluate its effectiveness, we measure its precision and recall on the following data set.

6.3.1 FourSquare Data Set

Foursquare is a mobile social-networking application that allows registered users to share their presence in a venue (e.g., share their "check-in" in a restaurant) with their social contacts. Users can share their

[1]Note that users can belong to more than one category. The value listed as "Total" reflects the number of different users and not sum of that column.

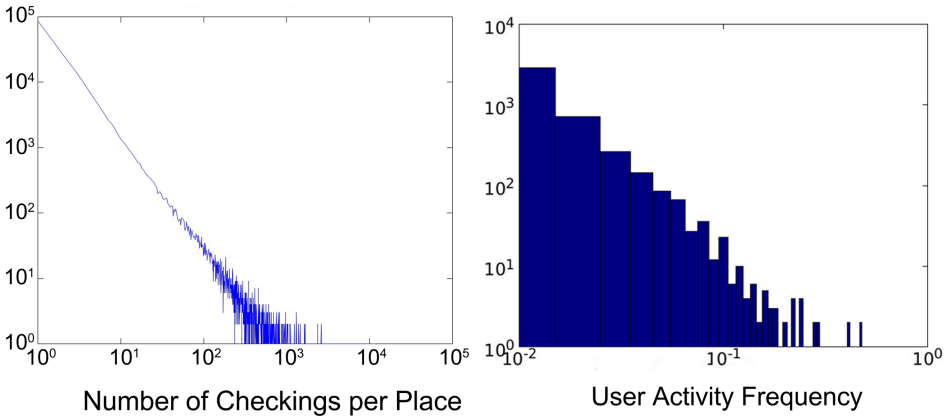
Category	Precision@10	Recall@10
nightlife	0.019	0.018
food	0.013	0.012
travel	0.004	0.005
shops	0.003	0.003
home/work/etc.	0.001	0.001
arts&entertainment	0.000	0.000
parks&outdoors	0.000	0.000
education	0.000	0.000

Table 6.2: Implicit SVD's precision and recall across categories.

check-ins not only on Foursquare but also on Twitter and Facebook. Each venue is associated with a category (e.g., "nightlife", "food") and a sub-category (e.g., "bar", "club", "Italian restaurant"). In 2011, Cheng *et al.* [20] collected 22 million check-ins of 225,098 users. We take the 228,625 check-ins in Greater London, which are generated by 29,044 users across 7,205 venues. To this data in the form of pairs *(user,venue)*, with further crawling, we add each venue's category and subcategory. After considering venues and users that disjointly appear at least twice in our *(user,venue)* pairs, we end up with 3,110 users and 4,572 venues in the city of London. Table 6.1 breaks statistics about users and venues down into the different categories. One can, for example, see that food venues are numerous and attract many users, while educational venues are rare but proportionally attract more users.

6.3.2 Implicit SVD Performance

We arrange this data in a *venue-by-user* matrix and measure the *Implicit SVD*'s precision and recall. For each venue, precision is the probability that a recommended user is relevant ($\frac{relevant \cap recommended}{recommended}$), while recall is the probability that a relevant user will be recommended ($\frac{relevant \cap recommended}{relevant}$). By relevant, we mean users who visited the venue. Also, we consider that the recommendation list for each venue contains the *top-10* recommended users. The results reported in Table 6.2 shows that precision and recall are extremely low - for some



(a) Visitors per Venue (b) User Activity

Figure 6.1: *(a)* Number of Visitors per Venue; *(b)* Frequency Distribution of user activity: this is the user's fraction of visited locations over the total of locations.

categories, they are even zero. These appalling results have a clear explanation - the data is sparse. There are too few people going to the same venue; indeed, the number visitors per venue can be modeled by a power law (Figure 6.1).

It thus seems that an alternative mechanism for recommending people is needed. But what sort of mechanism should we use? The widely-used classification algorithm of *SVM* does not work in the presence of data sparsity [67]. Therefore, we need a solution that: 1) is robust to sparse data; and 2) integrates domain knowledge (after all, our goal is to model how people "move" as much as is to model their preferences).

6.4 Domain-aware Recommendation

We take these two requirements and translate them into a solution that unfolds in three steps:

1. Incorporate domain knowledge from the complex system literature in human mobility (Section 6.4.1);

2. Deal with data sparsity by using item-based collaborative filtering to model user preferences (Section 6.4.2);

3. Integrate the previous two steps into two different Bayesian models and one linear regression (Section 6.4.3).

6.4.1 Individual Closeness

For starters, one might go to a venue not only because one likes it but also because is nearby. Thus, leaving out the users' taste from a moment, one can model the probability of an individual visiting a venue as $p(go|close)$ - i.e., the probability of going to a venue given that it is close - and can do so using Bayes' Law:

$$p(go|close) \propto p_{close} \cdot p_{go} \tag{6.1}$$

where p_{close} is the probability of the user being close (being at a certain distance), and p_{go} is the probability of a user going to any venue:

$$p_{go} = \frac{\#\text{venues visited by user } u}{\text{total } \#\text{venues}} \tag{6.2}$$

This latter probability reflects the general activity of a given user, which is a skewed distribution (Figure 6.1(b)), as one would expect: the vast majority of users visit few places, while a tiny fraction of (power) users (0.3%) visited roughly 20% of the London venues (within a category).

Literature: How people move

Scientists have long wondered how to measure something as ephemeral as movement. Early studies suggested that humans wander in a random fashion, similar to a so-called "Levy flight" pattern displayed by foraging animals. In 2006, to track human movements, researchers used more than half a million USA one-dollar bills as a proxy measure and analyzed their movements as they were passed around over five years [15]. They found many short movements and occasional longer ones. Similar patterns were found by Gonzalez *et al.* [30] who

studied the trajectories of 100,000 mobile phone users tracked for six months. These researchers found that people are regular, in that, the vast majority of them move around over a very short distance (from 5 to 10Kms) and make regular trips to the same few destinations such as work and home on a daily basis (70 percent of the time they were found in their two most frequently visited locations); people occasionally make longer trips when they, for example, go on vacation. More recently, Cheng *et al.* [20] analyzed the movement of Foursquare users across venues and found similar patterns: a mixture of short, random movements with occasional long jumps. As such, the vast majority of users had a small radius of exploration - typically less than 16 Kms.

Considering Geographic Closeness

To sum up, upon different types of movement (derived from dollar bills, mobile phones, and mobile social-networking applications), researchers in different disciplines have independently concluded that people rarely stray from familiar areas - they travel to a limited number of nearby locations and, consequently, short-range movements are more frequent than long-range ones (i.e., the frequency distribution of distance is exponentially distributed). This is also the case in our London data: Figure 6.2 plots the probability of one's traveling a certain distance for different venue categories. The distributions (for different categories) are very skewed and all fit the same distribution:

$$p_{close} = \frac{k_1}{d_{ui}^{\alpha}} \tag{6.3}$$

where d_{ui} is the distance between the user's (u's) center of geographic interest - that is, the center of mass or barycenter computed considering the locations where the user has previously checked-in - and the venue i. Interestingly, different venue categories are associated with different α, and for higher α, the less distance matters in one's choice when visiting a venue. Table 6.3 reports the α's for the different categories. The highest α (2.22) is associated with venues in the category "travel": those include train stations and bus stations, and it makes sense that people travel farther when going to places of limited supply (e.g., not all neighbourhoods have a train station). The lowest α's are registered

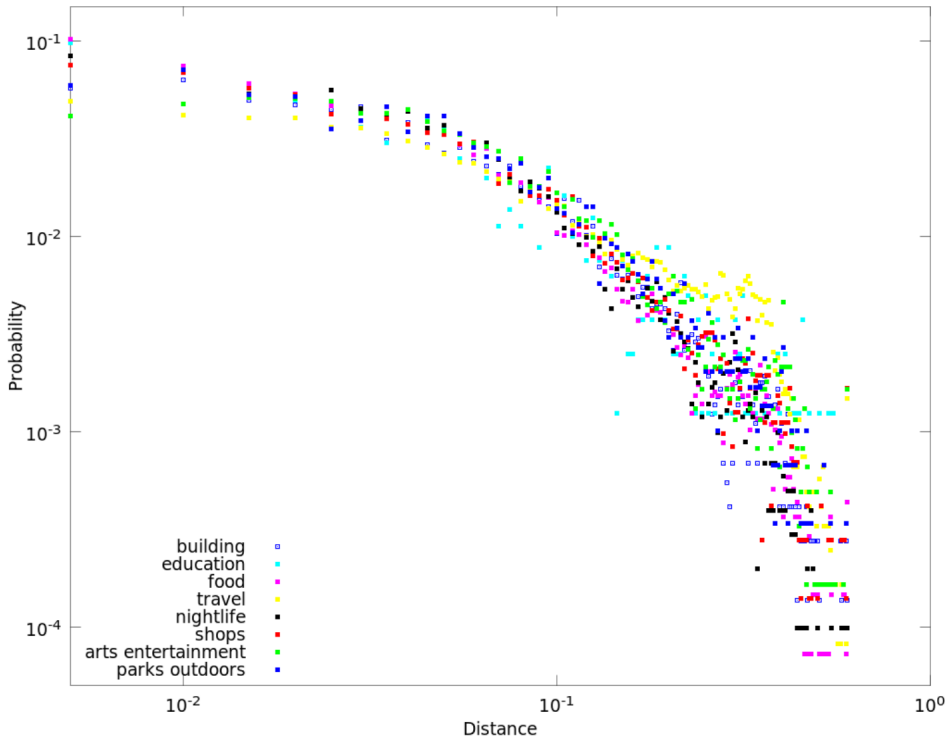


Figure 6.2: Probability of one's traveling a certain distance across different types of venues (best seen in color).

for venues in the categories "nightlife" and "home/work/etc.". That is, one's center of geographic interest revolves around home and work locations, and when going to bars, one goes to nearby ones.

Considering Power Users

Another conclusion from the literature is that not all mobile users are equally mobile. Individuals display significant regularity, yet, when compared to each other, there are few users who travel a lot, while the vast majority have limited travel activity. By framing the problem probabilistically, expression (6.1) is able to account for those special (power) users. It does so with p_{go} in expression (6.2), which reflects

Category	α
nightlife	1.61
home/work/etc.	1.62
shops	1.64
food	1.64
arts&entertainment	1.64
parks&outdoors	1.68
education	1.93
travel	2.22

Table 6.3: Why People Visit Different Types of Venues. Higher α, more one travels farther than usual to reach the venue in that category.

the extent when one is a power user or not. Hence our model takes in account power users.

6.4.2 Likes

The model in expression (6.1) has only considered whether one user is close or not and whether is a power user or not; but the model has not taken into account personal preferences. To fix that, we need to compute $p(like|go)$ - we need to compute the extent to which a user visits venues that are predictable from his/her past visits/likes. However, to do so, we need a way to measure user's *like*s. Since our data is sparse (Section 6.3), we measure likes not based on similarity among users but among venues. That is, we use an *item*-based collaborative filtering [77], which has been found to work well in such situations: "Unlike traditional collaborative filtering, the algorithm also performs well with limited user data, producing high-quality recommendations based on as few as two or three items." [52]. Rather than matching the user to other similar users, item-to-item collaborative filtering matches each of the user's venues with similar venues. A common way of computing the similarity between two venues is to compute the cosine similarity between two binary vectors: each vector reflects a venue, and a vector's i^{th} position reflects whether the i^{th} user visited the venue or not. Upon a so-constructed venue similarity table, the algorithm finds, for

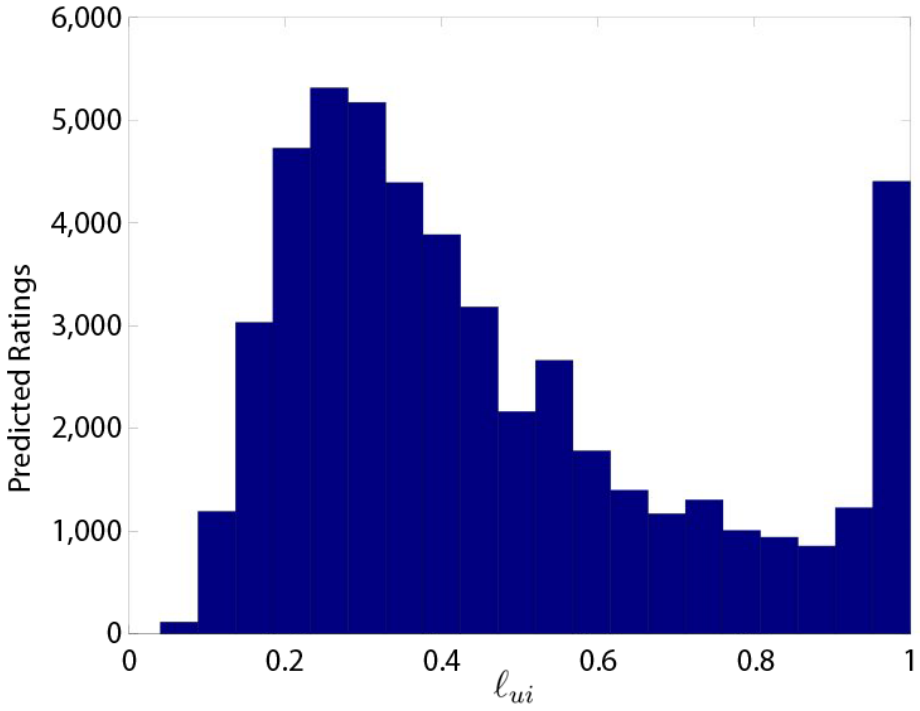


Figure 6.3: Distribution of Predicted Ratings.

each user, the venues similar to the ones previously visited by the user.

We apply the item-based collaborative filtering algorithm on the *user*-by-*venue* matrix and obtain a rating ℓ_{ui} for each user u and venue i. Figure 6.3 shows the distribution of the predicted ratings. Upon these ratings, we compute $p(like = \ell_{ui}|go)$, which is the fraction of venues i visited by u that have predicted ratings ℓ_{ui}:

$$p(like = \ell_{ui}|go) = \frac{\#\text{venues visited by user } u \text{ with rating } \ell_{ui}}{\text{total } \#\text{venues visited by user } u} \quad (6.4)$$

6.4.3 Putting All Together

Having users' whereabouts and preferences at hand, we now need to predict which users are likely to be at a certain venue. We do so using

a Naive Bayesian model, a Bayesian model, and a linear regression.

Naive Bayesian model

One simple way of modeling all the three factors together is to compute $p(go|like, close)$ using Bayes' Law:

$$\begin{aligned} p \quad & (go|like, close) \propto \\ \propto \quad & p_{like} \cdot p_{close} \cdot p_{go} \\ \propto \quad & p_{like} \cdot p_{go} \cdot \frac{k_1}{d_{ui}^{\alpha}} \end{aligned}$$

For each pair *(user, venue)*, we compute p_{close} with expression (6.3) and p_{like} with (6.4); and for each user, we compute p_{go} with (6.2). The importance of venue i for user u is then proportional to the above $p(go|like, close)$, and we call it $rank_{u,i}$.

Bayesian model

The previous model assumes that whereabouts and preferences are independent. This might well be not the case: those addicted to luxury goods will often be found near Bond Street (a major shopping street in the West End of London with many high price fashion shops). Here preference and whereabout go hand in hand. To go beyond independence, we could model jointly the two attributes:

$$p(go|like, close) = \frac{p_{like|go,close} \cdot p_{go|close}}{p_{like|close}}$$

where:

$$p_{like|go,close} = \frac{\#\text{venues visited by user } u \text{ at distance } d_{ui} \text{with rating } \ell_{ui}}{\#\text{venues visited by user } u \text{ at distance } d_{ui}}$$

$$p_{go|close} = \frac{\#\text{venues at distance } d_{ui} \text{ visited by user } u}{\#\text{venues at distance } d_{ui}}$$

$$p_{like|close} = \frac{\#\text{venues at distance } d_{ui} \text{with rating } \ell_{ui}}{\#\text{venues at distance } d_{ui}}$$

Linear Regression

Another approach for combining preferences and whereabouts is to run a linear regression:

$$rank_{u,i} = \alpha + \beta_1 I_{like} + \beta_2 I_{close} + \beta_3 I_{close} \cdot I_{like}$$

where I's are normalized values of whereabouts and preferences: I_{close} is $\frac{1}{\log(d_{ui})}$ (the logarithm because the frequency distribution of distance is very skewed, so we smooth it), and I_{like} is ℓ_{ui}. The product $I_{close} \cdot I_{like}$ controls the interaction effects between whereabouts and preferences.

6.5 Evaluation

The goal of this work is to predict which users are more likely to visit a given venue. To ascertain the effectiveness of our proposed techniques at meeting this goal, we need to select a desirable metric, measure it, and interpret those measurements. We execute these three steps next.

Metric. We need to find a measure that reflects the extent to which the predicted users for a venue are those who actually visited the venue. One such measure is called percentile-ranking [37]. The percentile-ranking $rank_{u,i}$ of user u for venue i ranges from 0% to 100%: it is 0%, if user u is first in venue i's recommendation list; it is 100%, if the user is last. Percentile-ranks have the advantage over absolute ranks of being independent of the number of users. Our quality measure is then the total average percentile-ranking:

$$\overline{rank} = \frac{\sum_{u,i} gone_{u,i} \cdot rank_{u,i}}{\sum_{u,i} gone_{u,i}} \tag{6.5}$$

where $gone_{u,i}$ is a flag that reflects whether user u was in venue i: it is 0, if u was not there; otherwise, it is 1. The lower $\overline{rank}$ for a list, the better the list's quality. For random predictions, the expected value for $rank_{u,j}$ is 50% (averaging infinite placements of users for a

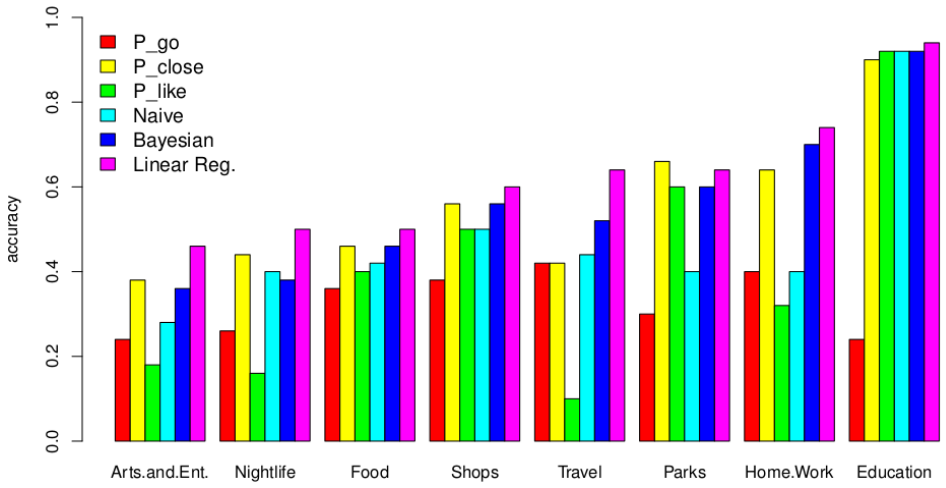


Figure 6.4: Rank Precision. The rank goes from 0 (random baseline predictions) to 1 (relevant user always ranked first in the recommendation list).

venue returns the middle position of the list). Therefore, $\overline{rank} < 50\%$ indicates an algorithm better than random. To ease illustration, we convert percentile ranking into ranking accuracy, which is 1, if the percentile ranking is 0% (best); and it is 0, if the percentile ranking is 50% (random):

$$accuracy = \frac{50\% - \overline{rank}}{50\%} \tag{6.6}$$

Accuracy would be 0 for a random predictor (baseline), and would be 1 for an ideal (oracle) predictor.

6.5.1 Analysis

To measure the ranking accuracy, we run a *10-fold* cross validation. That is, we divide the data set into 10 segments, we take one segment s at a time, consider it to be the testing set, and go through the following steps:

1. For each venue in the *training* set (the venues in all segments other than s), associate it with the users who visited that venue.

2. Train the model using the venues (and corresponding visitors) in the training set.

3. Use the trained model to then infer a ranked list of users who are likely to go to each venue in the *testing* set (the venues in s).

We finally compare the users predicted for each venue to those who actually visited it (those who are in the testing set of the ground truth).

6.5.2 Results

Figure 6.4 reports the ranking precision for the individual components of the Bayesian models (first three bars in each venue category) and for the overall models (Naive in the fourth bar, Bayesian in the fifth, and Linear Regression in the sixth). Starting from the first bar in each category (p_{go}), one sees that recommending power users works better than random (accuracy is always well above zero): more so for shops (0.38) than for arts&entertainment venues (0.24). Considering only nearby places (second bar in each set) returns more accurate rankings - again, more for shops (0.6) than for arts&entertainment venues (0.38). However, if one considers only past user preferences (third bar p_{like}), then accuracy is comparable to that of recommendations based on proximity (second and third bars do not differ much). This suggests that the simple concept of geographic distance is as important as that of the user's taste in all venue categories. It also suggests that, by only knowing where a user usually hangs out (without any information on the user's taste), one can produce reasonable recommendations (ideal for cold start situations). If we then combine these previous elements in a Naive Bayesian model, results do not improve; on the contrary, they are worse than those offered by simple geographic proximity for venues in the categories "food" and "arts&entertainment". That might be because the model treats its components as like they were completely independent. However, on average, the Pearson correlation coefficients ρ between each pairs of components are small: $\rho(p_{go}, p_{like}) = .13$,

$\rho(p_{go}, p_{close}) = 0.05$, and $\rho(p_{like}, p_{close}) = 0.21$. Yet, looking at the fifth bar in each set, one registers improvements with the traditional Bayesian modeled (in which dependencies are model). Another common reason for which Naive Bayesian does not work well in certain situations is that the addition of redundant components and arbitrary discretization of the random variables skews the learning process, and that seems to be the case here. Indeed, the linear regression (last bar) - which just models taste, whereabouts, and interactions between the two - works best in all categories. As one would expect, for categories characterized by less data sparsity and periodic patterns (e.g., education buildings), the models perform extremely well (accuracy above 0.90): the performance tend to be comparable to, if not better than, those registered in Web applications.

6.6 Discussion

6.6.1 Putting Results into Context

We have studied different strategies for recommending "guests" for real-world venues and, not surprisingly, found that results are best not only for venues with considerable historical data (e.g., educational institutions) but also for venues that are visited regularly (e.g., work locations). For other types of venues such as restaurants and bars, geographic closeness plays a very important role. Combining user preferences and geographic closeness has the expected result of offering more accurate recommendations, and that result can be achieved by using very simple models - Bayesian or linear regression. Being simple, these models not only are scalable and cost efficient but also produce recommendations that are easy to explain. The main criticism for the new Facebook "suggested guests" feature has been that it "does not offer... any sort of context" [22]. Our recommendations - which depend on whether one has visited similar locations or whether one often hangs out in certain neighbourhoods - are likely to be easier to explain than those produced by black-box approaches.

For the case of recommending shows on set top boxes, Hu *et al.* [37] had 17K of unique programs (roughly twice our number of venues)

and 32M non-zero ratings (140 times ours). In that context of less sparsity, they managed to achieve a ranking accuracy as good as 0.8 (upon learning from 200 distinct factors). Thus our results with the linear regression (always above 0.5 and above 0.6 for categories such as "shops" and "parks" and "travel") are comparable to those reported in the literature in far more favorable contexts (140 times less sparsity). Also, the percentile rankings are expected to slightly improve in more 'realistic' situations. To see why, consider that our data has been collected within a limited time window; by contrast, if one were to crawl the entire Foursquare history, then the resulting data would be still sparse but less so, and, as such, the prediction results would improve, as we have already registered with the category "educational" venues for which the accuracy was above 0.9.

6.6.2 When It Does not Work

When putting forward new predictive models, one often tends to focus on favorable situations in which predictions are best. Next, we briefly focus on the opposite case - we focus on situations in which predictions are worst. The idea behind this exercise is to find out which aspects future models should consider to increase accuracy. To this end, we run a qualitative study. For each venue i, we compute four predictability and unpredictability measures upon the following quantities: $gone_{ui}$, which reflects whether user u visited venue i; the geographic decay constant α taken from Table 6.3; the predicted rating l_{ui} for user u and venue i; and the distance d_{ui} between u's geographic center of interest and venue i. More specifically, upon these quantities, for each venue i, we compute:

Geo Predictability. The higher it is, the more the venue's visitors are predictable based on distance. It is higher for venues (e.g., bakery shops) whose visitors travel nearby:

$$P^i_{geo} = \frac{\sum_u \frac{1}{\log(d^\alpha_{ui})} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average inverse (log) distance for the venue's visitors.

Geo Unpredictability. The higher it is, the less its visitors are predictable based on distance. It is higher for venues (e.g., airports, high-end restaurants) whose visitors travel farther:

$$U^i_{geo} = \frac{\sum_u \log(d_{ui}) \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average (log) distance for the venue's visitors.

Like Predictability. The higher it is, the more its visitors are predictable based on past preferences (past likes). It is higher for venues whose visitors have common preferences:

$$P^i_{like} = \frac{\sum_u \ell_{ui} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average predicted ratings for the venue's visitors.

Like Unpredictability. The higher it is, the less its visitors are predictable based on past preferences. It is higher for venues whose visitors have diverse preferences:

$$U^i_{like} = \frac{\sum_u \frac{1}{\ell_{ui}} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average inverse predicted ratings for the venue's visitors.

We create four tables that contain the *top*-10 venues ranked by each of those four measures and ask three coders (three Londoners with diverse background - architect, barrister, and medical doctor) to build predictability boxes of the kind in Figure 6.5(a). For them, that translated into ordering venue categories that are predicted (hard to predict) by geographic distance based on the table ranked by P^i_{geo} (by U^i_{geo}), and categories that are predicted (hard to predict) by user preferences based on the table ranked by P^i_{like} (by U^i_{like}). We consider only the answers for which two out of three coders or all three have independently agreed. In Figures 6.5(b) and 6.5(c), word size is proportional to the coders' agreement.

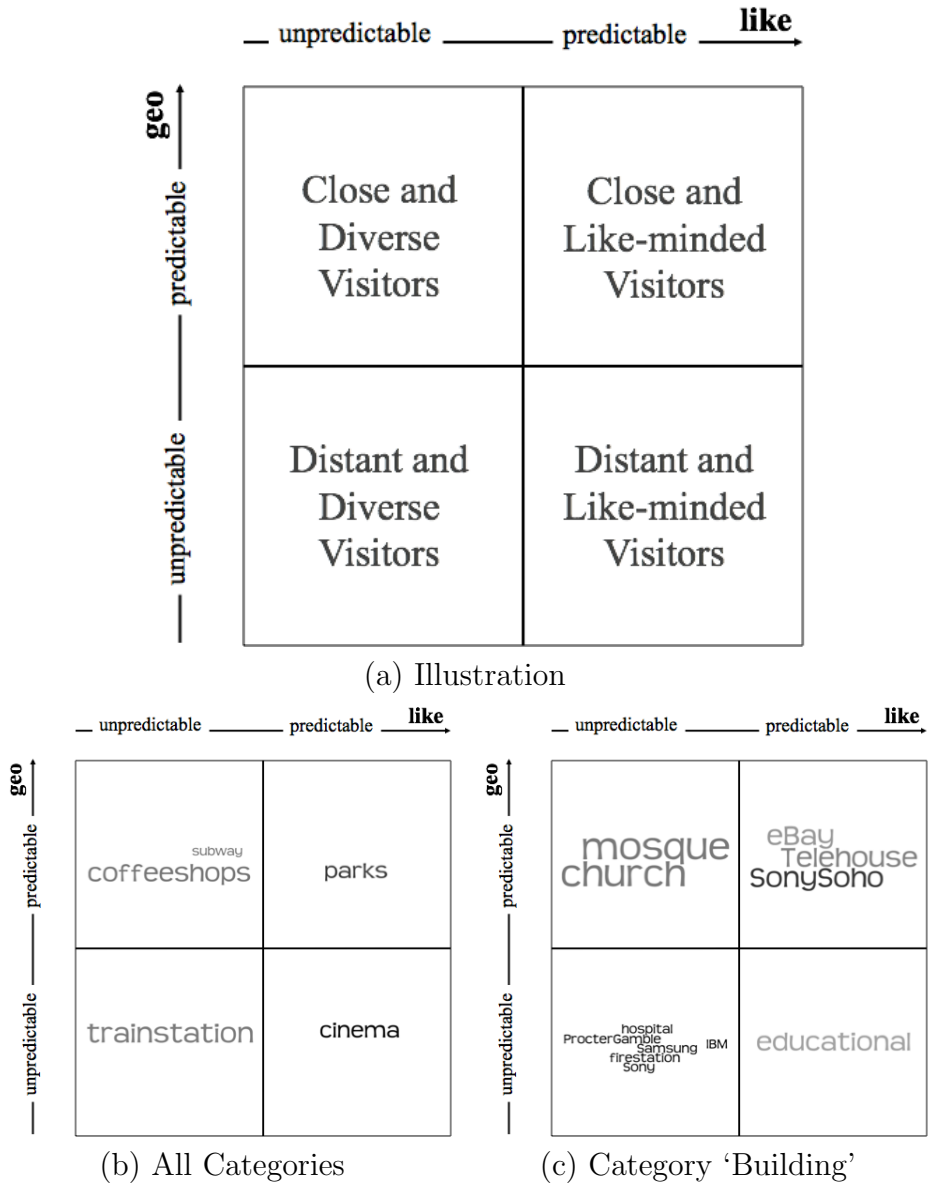
For all venue categories (Figure 6.5(b)), the unpredictable venues (predicted neither by closeness nor by taste) are train stations. That is because train stations are often far from where one hangs out and do not

reflect a specific taste in, say, music, bars, clubs, or food. By contrast, local parks and outdoor activities are predictable either by closeness or by taste, suggesting that people prefer their local parks over bigger parks (they stay close), and that residents of the same area tend to be like-minded (a tendency often called "geographic sorting" [14]). Closeness is more informative for predicting visits to coffee shops (one tends to go to local coffee shops); while user taste is more informative for cinemas in central London areas, where diversified choice of movies motivates visitors to travel farther than usual. For the specific category "buildings" (Figure 6.5(c)), the unpredictable venues (predicted neither by closeness nor by taste) are companies such as IBM, Procter&Gamble, Samsung, whose headquarters are in suburban areas where people with diverse background work but do not hangout, not because of limited supply of amenities. By contrast, the behavior of employees (mostly interaction designers) of Sony, eBay or Telehouse, working in central areas like Soho is predictable either by closeness or by taste. Finally, closeness is more informative for predicting visits to mosques and churches (one tends to go to local religious venues); while the user taste is more informative for visitors of universities (e.g., UCL's, Birkbeck's) facilities in central areas. From these qualitative results, one can extrapolate two key insights:

1. Predictable situations are those in which people: a) stay close because they have what they need at hand; or b) congregate in places where other like-minded people tend to be (e.g., local parks and cinemas).

2. By contrast, unpredictable situations are those in which people: a) travel because they do not have what they want at hand; or b) go to places that attract individuals of very diverse backgrounds (e.g., coffee shops, train stations).

Future work should go into models that are able to simultaneously account for these (at times) conflicting situations.



(a) Illustration

(b) All Categories

(c) Category 'Building'

Figure 6.5: Four-quadrant Predictability Box. Quadrants are defined by the venue's unpredictability and predictability measures, which are based on visitors' geographic closeness (rows) and likes (columns).

6.6.3 Applications

The practical implications of this work go beyond traditional applications of recommender systems:

Commercial Property Evaluation. This is the process of identifying and quantifying the value of commercial properties and is generally carried out by experts who analyze properties similar to the one being valued. A primary factor that affects this assessment is location, yet this factor is generally quantified based on the valuer's expert knowledge of a locality. More recently, well-informed ways of valuing properties have been proposed, and they rely on the *creation and maintenance* of GIS-based property valuation databases. These databases (especially those for commercial properties) might well be enriched by this work - in particular, by knowing how close a venue is to its target audience.

Social Marketing. Social marketing can be defined as a research-driven approach to promote voluntary behavior change in a priority population. A case in point is "Stop the Sores", a social marketing campaign designed to increase syphilis testing in Los Angeles County [65]. Social marketing has its foundation in consumer marketing and consists of three key elements: market research [86], audience segmentation [33], and branding [40]. The second element of segmentation is related to this work and is essential for developing campaign messages that resonate with the target population and helps in identifying the largest or highest-risk subgroup (e.g., swingers, men having sex with men) at minimal cost.

Target Advertising. The first step when promoting new nightclubs, bars and restaurants is often to identify the target market. Thus, knowing the kind of people who are willing to go to, say, certain restaurants or bars (which is what this work is about) translates into low-cost marketing strategies for bars and restaurants that are willing to attract new crowds.

6.6.4 Scalability

The two main parts of this work - which model whereabouts and preferences - are highly scalable:

Whereabout Part. This requires to know a geographic point for each user (where an individual usually hangs out) and one single decay constant α (which is universal in that it equally applies to all users). Learning a point per user and a constant for all of them is extremely scalable. In addition to being scalable, the models are likely to be generalizable, not least because they have been built upon previous general rules of people's wanderings [15; 20; 30], and, being general, they are also likely to work for any instance of mobility (not only for Foursquare users).

Preference Part. This translates into item-based collaborative filtering. The (computationally) expensive part of this algorithm (venue similarity table) can be computed offline, while what needs to be computed online - matching the user's venues with similar ones - scales independently of the total number of venues and total number of users, in that, it only depends on the number of venues each single user has visited (which is generally extremely low).

Chapter 7

Final Remarks

7.1 Summary and Conclusions

In this thesis we have developed new techniques to find relevant people in Online Social Networks (OSNs), providing practical and theoretical contributions. While previous work focused in specifics definitions of users' influence or importance, we have proposed the concept of people's relevance, describing different types of relevant users, finding a set of features that allows to model users' relevance, and developing tools to find them. Here, we summarize the main contributions of this thesis:

- Define a set of tasks to find relevant users considering different goals and contexts, providing a detailed description of such users and proposing novel metrics to describe people's relevance.
- Develop tools to find relevant users, taking the point of view of the OSN providers and advertisers, as well as considering the people that is trying to push new ideas and topics on the network.
- Provide useful insights about users' behavior according to their relevance, popularity, and activity showing - across different platforms such as Facebook, FourSquare, and Twitter - that most

active users are usually more relevant than the popular ones. Moreover, we show that usually very popular users arrive late to the new trends, and that there are less popular, but very active users that generates value and push new ideas in the network.

- Propose novel methodologies to model users' behavior, incorporating geographical and time information, allowing to model complex relations among users beyond the social graph.

Specifically, in Chapter 3 we have presented valuable insights about the relation between incoming and outgoing people's activity. In Chapter 4 we have developed a robust framework to compute the monetary value that users produce for a OSN. In Chapter 5 we have proposed a robust algorithm to rank trendsetters, presenting the problem of the spread of an innovation as a ranking problem considering the time factor, and also presented a sound and extensible way to model topics and influence allowing to run this algorithm in different contexts combining the social graph properties - like node degree and Pagerank - with users activity (*e.g.* messages exchanged by users). Finally, in Chapter 6 we have presented a model to find potential customers for venues (advertisers), considering people personal taste and mobility their mobility patterns.

In detail, the main conclusions of this thesis are:

- **There is a relation among people's popularity, online activity, and the monetary value that they produce (Chapters 3 and 4)**. In Chapter 3 we showed that the amount of activities performed by a user correlates better with the attention that she/he obtain than her/his number of *friends*. In Chapter 4 we showed that the monetary value of users is directly related with their activity. In general, very active users are more relevant than the popular ones.

- **To be relevant is not necessary to be popular (*i.e.* have a lot of friends) (Chapters 4 and 5)**. Popularity and relevancy are related, but are not the same: in Chapter 4 we note that popularity is necessary, but not enough to produce monetary

value. In Chapter 5 we find that popular people are unlike to push new ideas, while trendsetters are usually "smaller" in terms of popularity.

- **To find relevant people it is necessary to use temporal graphs instead of static graphs (Chapter 5)**. Since social graphs (friend/follow relations) do not capture the exchange of messages within the network, they are not sufficient to find relevant people. Therefore, it is necessary to introduce this information into the graph. We showed that weighting the edges using the time information of these messages is a useful technique.

- **Demographic information must be considered to find relevant people (Chapter 6)**. In Chapters 2, 4, and 6 we showed that people's demography have clear implications in their online behavior. Specifically, people's geographical location gives valuable information to understand if they would be relevant in a given context. For example, when a local shop (*e.g.* a restaurant) is looking for relevant people for their business, the mobility patterns of potential customers would be the most valuable information to determine if they are a good target (relevant) or not. This finding is just an example of how contextual information must be considered to find relevant people in a given scenario.

Overall, we show and explain the differences between popularity and relevance, and provide useful tools for finding relevant people in OSNs.

7.2 Applications

The applications of our findings are many. Considering that our social relations are becoming more and more digitalized (and logged), the importance of finding relevant people in social networks will increase in the future and the limits are still unknown.

Just to mention some examples, the applications in fields like marketing and politics are clear:

- The insights given in Chapter 3 would be useful for *community managers* and people in general when designing strategies to gain more attention in OSNs.
- The algorithm to find trendsetters (Chapter 4), would be of interest for advertisers, politicians, and social movements, who may use it to identify important people for seeding viral campaigns.
- The methodology to determine the monetary value of users (Chapter 5) could be useful for OSNs providers to improve their revenue model and also for users that could became aware of monetary value of their accounts, and
- the insights given in Chapter 6 about people mobility, may be of interest for advertisers and the (new) geo-social networks providers.

Moreover, in terms of theoretical implications, the idea of weighting the social graph using time information (presented in Chapter 4) can be extended and applied in new contexts, such as improvement of content discovery applications and recommender systems. The results of our work can also be used to improve the analysis of information diffusion processes. Therefore, we believe that our research can be applied both in industry and academia.

7.3 Future Work

The fast evolution, growth, and change of Online Social Networks is one the biggest challenges for researchers on this field. Just to give an example, when we started this thesis in 2009, Twitter had around 50 millions of users. In 2012, this number reached more than 500 millions of registered users.[1] In a similar period Facebook grew from

[1] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city

150 millions[2] to 1.1 billion of users.[3] This is not only a change in terms of the amount of data that needs to be analyzed, but it is also a behavioural change. The increase of OSN penetration makes that many relations and interactions that used to be offline, nowadays are happening online. The complexity of these actions is growing day by day. Thus, the boundaries of what is exogenous and endogenous in a OSN are changing constantly and the models that we build need to be prepared to consider new inputs when they become available.

In this thesis we were able to incorporate one of these changes, that is, to consider people's location, a feature increasingly available due to the high penetration of smartphones. This is not a minor change because beyond the growth in terms of users, the use of mobile devices was one of the main changes in the use of OSNs during this last four years. Mobile phones, not only made OSNs ubiquitous, but also gave new information about the users. The present and the future of OSNs will be inextricably linked with the usage and evolution of mobile devices. Therefore, future work must consider this new scenario.

However, with the available data there are clear paths for future work open in each chapter. In Chapter 3 we need to study how spammers affects our models and how this problem can be addressed. In Chapter 4, we plan to address the tussle between the perceived value of users and the impact on their privacy as a result of their increased awareness of their value. In Chapter 5, using machine learning techniques we want to compare the trendsetters with other users that appear to be similar, but that do not achieve success, trying to identify the key characteristics of trendsetters. In Chapter 6, we want to study outlier cases, to understand which factors make people decide to travel far away and visit new venues, and how the actions of (relevant) people influences other user's decisions.

Finally, we plan to study the relation among the different types of relevant users described on this thesis, and how they interact among them. To define a taxonomy of relevant people would be the next step for this line of research.

[2] https://blog.facebook.com/blog.php?post=46881667130
[3] http://investor.fb.com/releasedetail.cfm?ReleaseID=761090

Bibliography

[1] P. Adams. *Grouped: How Small Groups of Friends Are the Key to Influence on the Social Web*. Voices That Matter. Pearson Education, 2011. 1, 30, 35, 49

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of International Conference on Knowledge discovery and data mining*, KDD'08, pages 7–15, NY, USA, 2008. ACM. 8

[3] S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow, editor, *Groups, Leadership, and Men*, pages 177–190. Carnegie Press, Pittsburgh, PA, 1951. 7

[4] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 33–42. ACM, 2012. xvii, 11, 12

[5] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 33–42. ACM, 2012. 12

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedingsof International Conference on Knowledge discovery and data mining*, KDD'06, pages 44–54, NY,USA, 2006. ACM. 8

[7] R. Baeza-Yates and E. Davis. Web page ranking using link attributes. In *Proceedingsof International Conference World Wide Web*, WWW'04, pages 328–329, NY, USA, 2004. ACM. 54

[8] E. Bakshy, J. Hofman, W. Mason, and D. Watts. Everyone's an in-

fluencer: quantifying influence on twitter. In *Proceedingsof International Conference on Web search and data mining*, WSDM'11, pages 65–74, NY, USA, 2011. ACM. 30

[9] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proceedingsof Conference on Electronic commerce*, EC '09, pages 325–334, NY, USA, 2009. ACM. 8, 30

[10] F. Bass. A new product growth for model consumer durables. *Management Sciences*, 15(1):215–227, 1969. 7

[11] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010. 20, 26

[12] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Goncalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of International ACM SIGIR*, Boston, MA, USA, July 2009. ACM. 30

[13] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *IMC '09: Proceedings of the 9th ACM SIGCOMM Conference on Internet measurement Conference*, pages 49–62, New York, NY, USA, 2009. ACM. xvii, 33, 34

[14] B. Bishop. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Houghton Mifflin, May 2008. 91

[15] Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439, January 2006. 79, 94

[16] M. Cha, F. Benevenuto, Y.-Y. Ahn, and K. P. Gummadi. Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks*, 56(3):1066–1076, 2012. 30

[17] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedingsof the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA, 2010. 10, 14, 18, 19, 50, 53, 61

[18] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy.

ICWSM, 10:10–17, 2010. 30

[19] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings International Conference on Knowledge discovery and data mining*, KDD '09, pages 199–208, NY, USA, 2009. ACM. 7

[20] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 14, 77, 80, 94

[21] E. M. Daly and W. Geyer. Effective event discovery: using location and social information for scoping event recommendations. In *Proceedings of the* 5^{th} *ACM Conference on Recommender Systems (RecSys)*. ACM, 2011. 74

[22] B. Darwell. Facebook tests 'suggested guests' for events. In *Inside Facebook*, February 2012. 73, 75, 88

[23] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedingsof International Conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, NY, USA, 2001. ACM. 7

[24] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge discovery and data mining*, KDD '01, 2001. 30

[25] *A special report on social networking, The Economist*, Jan 2010. http://tinyurl.com/ylxtsek. Accessed 07/2011. 49

[26] A. Goel and K. Munagala. Hybrid keyword search auctions. In *Proceedings of the 18th International Conference on World wide web*, WWW '09, 2009. 30

[27] S. Goel, S. Lahaie, and S. Vassilvitskii. Contract auctions for sponsored search. In *Internet and Network Economics*, pages 196–207. Springer, 2009. 30

[28] J. Golbeck. Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web*, 3(4):12:1–12:33, September 2009. 75

[29] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 1019–1028. ACM, 2010. 9

[30] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196), June 2008. 79, 94

[31] A. Goyal, F. Bonchi, and L. Lakshmanan. Discovering leaders from community actions. In *Proceedings of International Conference on Information and knowledge management*, CIKM '08, pages 499–508, NY, USA, 2008. ACM. 9

[32] M. Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 15:1420 – 1443, 1978. 7

[33] S. Grier and C. A. Bryant. Social marketing in public health. *Annual Review of Public Health*, 26(1), 2005. 93

[34] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet measurement*, IMC '10, 2010. 30

[35] I. Guy, S. Ur, I. Ronen, A. Perer, and M. Jacovi. Do you want to know?: recommending strangers in the enterprise. In *Proceedings of the ACM Conference on Computer supported Cooperative Work (CSCW)*, 2011. 75

[36] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003. 10

[37] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the* 8^{th} *IEEE International Conference on Data Mining (ICDM)*, 2008. 76, 85, 88

[38] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the* 12^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006. 74

[39] B. J. Jansen, T. Flaherty, R. Baeza-Yates, L. Hunter, B. Kitts, and J. Murphy. The components and impact of sponsored search. *IEEE Computer*, 42(5):98–101, 2009. 32

[40] K. L. Keller. Branding perspectives on social marketing. *Advances in Consumer Research*, 25(1), 1998. 93

[41] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedingsof International Conference on Knowledge discovery and data mining*, KDD'03, pages 137–146, NY, USA, 2003. ACM. 7, 30, 50

[42] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8^{th} ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002. 74

[43] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999. 9

[44] J. Kleinfeld. Six degrees: Urban myth?, November 2006. http://www.psychologytoday.com/articles/200203/six-degrees-urban-myth 12

[45] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *Proceedings International Conference on Knowledge discovery and data mining*, KDD '08, pages 435–443, NY, USA, 2008. ACM. 9

[46] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedingsof International Conference on World wide web*, WWW'10, pages 591–600, NY, USA, 2010. ACM. xvii, 10, 11, 19, 20, 59

[47] I. Leontiadis, C. Efstratiou, M. Picone, and C. Mascolo. Don't kill my ads!: balancing privacy in an ad-supported mobile application market. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, page 2. ACM, 2012. 15

[48] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedingsof International Conference on Knowledge discovery and data mining*, KDD'09, pages 497–506, NY, USA, 2009. ACM. 30

[49] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 177–187, 2005. 12

[50] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005. 13

[51] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the* 16^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010. 74

[52] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003. 82

[53] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedingsof International Conference on Information and knowledge management*, CIKM '10, pages 199–208, NY, USA, 2010. ACM. 10, 30

[54] C. Lumezanu and N. Feamster. Observing common spam in twitter and email. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, IMC '12, pages 461–466, New York, NY, USA, 2012. ACM. 26

[55] G. Magno, G. Comarela, D. Saez-Trumper, M. Cha, and V. Almeida. New kid on the block: Exploring the google+ social graph. In *ACM Internet Measurement Conference 2012 (IMC'12)*, Boston, USA, 2012. xvii, 11, 12, 14

[56] S. Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967. 11

[57] E. Minkov, B. Charrow, J. Ledlie, S. Teller, and T. Jaakkola. Collaborative future event recommendation. In *Proceedings of the* 19^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, 2010. 74

[58] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference*, 2007. xvii, 11

[59] Y. Moon and C. Kwon. Online advertisement service pricing and an option contract. *Electron. Commer. Rec. Appl.*, 10(1):38–48, Jan. 2011. 30

[60] G. Moore. *Crossing the Chasm.* HarperBusiness, revised edition, Sept. 2002. 50

[61] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM*

SIGKDD International Conference on Knowledge discovery and data mining, KDD '12, pages 33–41, New York, NY, USA, 2012. ACM. 29

[62] H. Nazerzadeh, A. Saberi, and R. Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *In WWW 08: Proceeding of the 17th International Conference on World Wide Web*, 2008. 30

[63] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. 9, 51, 56

[64] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *ACM WSDM*, pages 45–54, 2011. 7

[65] A. Plant, J. A. Montoya, H. Rotblatt, P. R. Kerndt, K. L. Mall, L. G. Pappas, C. K. Kent, and D. Klausner. Stop the Sores: The Making and Evaluation of a Successful Social Marketing Campaign. *Health Promotion Practice*, 11(1), 2010. 93

[66] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes. Do All Birds Tweet the Same? Characterizing Twitter Around the World. In *Proceedings of ACM Conference on Information and Knowledge Management*, Glasgow,UK, 2011. 14

[67] D. Quercia, H. Askham, and J. Crowcroft. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. In *Proceedings of the* 4^{th} *ACM International Conference on Web Science (WebSci)*, 2012. 78

[68] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo, and J. Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the* 10^{th} *IEEE International Conference on Data Mining (ICDM)*, 2010. 75

[69] A. Radovanovic and W. D. Heavlin. Risk-aware revenue maximization in display advertising. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, 2012. 30

[70] F. Ricci and Q. N. Nguyen. Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. *IEEE Intelligent Systems*, 22(3):22–29, May 2007. 75

[71] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and

V. Almeida. On word-of-mouth based discovery of the Web. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011. 30

[72] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida. On Word-of-Mouth Web Based Discovery of the Web. In *Proceedings of ACM SIGCOMM Internet Measurement Conference*, 2011. 13

[73] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of International Conference on World wide web*, WWW'11, pages 695–704, NY, USA, 2011. ACM. 8, 61

[74] W. Romero, D.and Galuba, S. Asur, and B. Huberman. Influence and passivity in social media. In *ECML/PKDD (3)*, pages 18–33, 2011. 9

[75] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge discovery and data mining*, KDD '12, pages 1014–1022, New York, NY, USA, 2012. ACM. 30

[76] D. Saez-Trumper, D. Nettleton, and R. Baeza-Yates. High correlation between incoming and outgoing activity: a distinctive property of OSNs? In *Fifth International AAAI Conference on Weblogs and Social Media,*, Barcelona, Spain, 2011. 14, 46

[77] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the* 10^{th} *ACM Conference on World Wide Web (WWW)*, 2001. 82

[78] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: geo-social metrics for online social networks. In *Proceedings of ACM SIGCOMM Workshop on Social Networks*, Berkeley, CA, USA, 2010. USENIX Association. 13

[79] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 75

[80] T. Schelling. *Micromotives and Macrobehavior.* Norton, 1978. 7

[81] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *Proceedings of IMC*, November 2009. xvii, 33, 34

[82] J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 74–81, New York, NY, USA, 2005. ACM. 14, 19

[83] Y. Takeuchi and M. Sugimoto. CityVoyager: An Outdoor Recommendation System Based on User Location History. In *Ubiquitous Intelligence and Computing*, Lecture Notes in Computer Science, 2006. 75

[84] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011. 12

[85] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, Barcelona, Spain, August 2009. 14, 20, 29

[86] D. C. Walsh, R. E. Rudd, B. A. Moeykens, and T. W. Moloney. Social marketing for public health. *Health Affairs*, 12(2), 1993. 93

[87] Y. Wang, D. Burgener, A. Kuzmanovic, and G. Maci-Fernández. Understanding the network and user-targeting properties of web advertising networks. In *Proceedings of the 2011 31st International Conference on Distributed Computing Systems*, ICDCS '11, 2011. 30

[88] D. J. Watts and S. H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998. 11

[89] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedingsof International Conference on Web search and data mining*, WSDM '10, pages 261–270, NY, USA, 2010. ACM. 10, 19, 30, 53

[90] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedingsof International Conference on Web search and data mining*, WSDM'11, pages 177–186, NY, USA, 2011. ACM. 10, 64, 65

[91] J. Zhang, M. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedingsof the 16th International Conference on World Wide Web*, WWW '07, pages 221–230, NY, USA, 2007. ACM. 9, 53

[92] D. Zhou, X. Ji, H. Zha, and C. L. Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the* 15^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, 2006. 74