UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

MEASUREMENTS AND ANALYSIS OF ONLINE SOCIAL NETWORKS.

Autor: Roberto González Sánchez
Directores: Dr. Rubén Cuevas Rumín
Dr. Carmen Guerrero López

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

Leganés, Junio de 2014

TESIS DOCTORAL

MEASUREMENTS AND ANALYSIS OF ONLINE SOCIAL NETWORKS.

Autor: Roberto González Sánchez
Director: Dr. Rubén Cuevas Rumín
Directora: Dr. Carmen Guerrero López

Firma del tribunal calificador:

Firma:

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, de de

"It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories,
instead of theories to suit facts"
— *Sir Arthur Conan Doyle, The Adventures of Sherlock Holmes:*
A Scandal in Bohemia

Acknowledgments

No podría empezar esta sección sin dar las gracias a mi familia, sin la cual sería imposible que este documento se hubiera si quiera empezado. El apoyo que siempre he sentido por parte de mis padres y hermana ha sido tan importante para poder acabar con éxito este periodo como la paciencia y el cariño de Ángela en los momentos buenos y en los no tan buenos.

Quiero dedicar también unas líneas de agradecimiento a la Universidad Carlos III de Madrid y, en particular, al Departamento de Ingeniería Telemática y al grupo de investigación NETCOM por su financiación y apoyo. Siento que he tenido grandes compañeros y me llevo mejores amigos. Podría nombrar mucha gente en este apartado, pero para no olvidar a nadie voy a nombrar solo a Carmen, que ha sido mi tutora en este periodo y con la que llevo trabajando desde hace más de seis años. Ella fue quien me empujó a empezar el doctorado y su apoyo ha sido fundamental para poder acabarlo.

Un subconjunto de los anteriores merece especial atención ya que han tenido a bien aguantarme también fuera de las horas de oficina. Nada habría sido lo mismo sin Chema, Grego, Juanmi, Ángel, Gordillo, David, Raquel, Isaías, Isaac, Andrés, Marco, Rosa, Dani, Patricia...

I would want to thank all my "American family" for their support during my stay in Eugene. The Persian community accepted me since the first day, even when they could not understand me at all. I want to specifically mention my buddies Mesh, Isa and Behnaz. They made my stay in The States funny. Moreover, two "American" persons deserve a special mention: first, Reza Rejaie, he allows me to go to Oregon and work with his team. His invaluable advices have improved this work but also because he showed me how Persian people play soccer. The second person is my BFF Reza Motamedi. He made me feel at home since the first day when he picked me up at the airport till the day in which he drove me to the airport again.

Por último, y probablemente más importante tengo que agradecer a Rubén su apoyo. No está en ninguna categoría anterior por que ha sido parte diaria del trabajo, ha sabido aguantarme fuera cuando ha hecho falta, me ayudó a ir a EEUU y me guió antes de llegar

y hablo tanto con el que casi es parte de mi familia y desde luego un gran amigo. Sin los demás esta tesis habría sido más dificil. Sin Rubén no habría sido posible.

Resumen

Las redes sociales (OSNs por sus siglas en inglés) se han convertido en una de las aplicaciones más usadas de Internet atrayendo cientos de millones de usuarios cada día. La gran cantidad de información valiosa en las redes sociales (que antes no estaba disponible) ha llevado a la comunidad cientifica a diseñar sofisticadas tecnicas para recoger, procesar, interpretar y usar esos datos en diferentes disciplinas incluyendo sociología, marketing, informática, etc.

Esta tesis presenta una serie de contribuciones en este incipiente área.

Primero, presentamos un completo marco que permite realizar medidas a gran escala de redes sociales. Con este propósito, el documento describe las herramientas y estrategias seguidas para obtener un conjunto de datos representativo. Tambien, añadimos las lecciones aprendidas durante el proceso de obtención de datos. Estas lecciones pueden ayudar al lector en una futura campaña de medidas sobre redes sociales.

Segundo, usando el conjunto de datos obtenido con las herramientas descritas, esta tesis aborda dos aspectos fundamentales que son críticos para entender el ecosistema de las redes sociales. Por un lado, caracterizamos el nacimiento y crecimiento de redes sociales. En particular, llevamos a cabo un análisis en profundidad de una red social de segunda generación como Google+ (una red social lanzada por Google en 2011) y comparamos su crecimiento con otras redes sociales de primera generación como Twitter. Por otro lado caracterizamos la propagación de la información en redes sociales de diferentes maneras. Primero, usamos Twitter para llevar a cabo un analisis geográfico de la propagación de la información. Tambien analizamos la propagación de la información en Google+. En particular, analizamos los árboles de propagación de información y los bosques de propagación de información que incluyen la información sobre la propagación de una misma pieza de contenido a traves de diferentes árboles. A nuestro saber, este es el primer estudio que aborda esta cuestión.

Por último, analizamos la carga soportada por una red social como Twitter.

La investigación realizada nos lleva a los siguientes 4 resultados principales: (i) Es de esperar que las redes sociales de segunda generación crezcan mucho más rápido que las

correspondientes de primera generación, sin embargo, estas tiene muchas dificultades para mantener los usuarios involucrados en el sistema. Este es el caso de G+ que está creciendo al impresionante ritmo de 350K nuevos usuarios registrados por dia. Sin embargo una gran fracción (83%) de ellos no ha llegado nunca a ser activos y los que presentan actividad presentan en general una actividad menos que los usuarios de Facebook o Twitter. (*ii*) La información se propaga más rápico pero siguiendo caminos más cortos en Twitter que en G+. Esto es una consecuencia de la manera en la que la información es mostrada en cada sistema: sistema secuenciales como en Twitter fuerzan que la información sea consumida al instante mientras que sistemas selectivos como el usado en G+ o Facebook, donde la información que se muestra depende las preferencias de los usuarios y el volumen de interacción con otros usuarios ayuda a prolongar la vida del contenido en la red social. (*iii*) Nuestro analisis de la propagacion geográfica de la información en Twitter revela que los usuarios suelen enviar *tweets* desde una única localización geográfica. Además, el nivel de geolocalización asociada a las relaciones sociales varía entre países y encontramos algunos países, como Brasil, donde es más que la información información se mantenga local que en otros como Australia. (*iv*) Nuestro análisis de la carga de Twitter indica que el proceso de llegada de *tweets* sigue un modelo gausiano con un marcado patrón día-noche.

En definitiva, el trabajo presentado en este tesis permite aumentar nuestro conocimiento sobre el ecosistema de las redes sociales en direcciones esenciales como pueden ser la formación y crecimiento de redes sociales o la propagación de información en estos sistemas. Los resultados reportados ayudarán a desarrollar nuevos servicios sobre las redes sociales.

Abstract

Online Social Networks (OSNs) have become the most used Internet applications attracting hundreds of millions active users every day. The large amount of valuable information in OSNs (not even before available) has attracted the research community to design sophisticated techniques to collect, process, interpret and apply these data into a large range of disciplines including Sociology, Marketing, Computer Science, etc.

This thesis presents a series of contributions into this incipient area.

First, we present a comprehensive framework to perform large scale measurements in OSNs. To this end, the tools and strategies followed to capture representative datasets are described. Furthermore, we present the lessons learned during the crawling process in order to help the reader in a future measurement campaign.

Second, using the previous datasets, this thesis address two fundamental aspects that are critical in order to have a clear understanding of the Social Media ecosystem. One the one hand, we characterize the birth and grow of OSNs. In particular, we perform a deep study for a second generation OSN such as Google+ (a OSN released by Google in 2011) and compare its growth with other first generation OSNs such as Twitter. On the other hand, we characterize the information propagation in OSNs in several manners. First, we use Twitter to perform a geographical analysis of the information propagation. Furthermore, we carefully analyze the propagation information in Google+. In particular, we analyze the information propagation trees and the information propagation forests that analyze the propagation information of a piece of content through multiple trees. To the best of our knowledge any previous study has addressed this issue.

Finally, the last contribution of this thesis focuses on the analysis of the load received by an OSN system such as Twitter.

The conducted research lead to the following main four findings: (i) Second Generation OSNs are expected to grow much faster that the correspondent First Generation OSNs, however they struggle to get users actively engage in the system. This is the case of G+ that is growing at a impressive rate of 350K new users registered per day. However a large fraction (83%) of its users have never been active, and those that present activity

are typically significantly less engaged in the system than users in Facebook or Twitter. (*ii*) The information propagates faster but following shorter paths in Twitter than in G+. This is a consequence of the way in which information is shown in each system. Secuential-based systems such as Twitter force short-term conversations among their users whereas Selective-based systems such as those used in G+ or Facebook chooses which content to show to each user based on his preferences, volume of interactions with other users, etc. This helps to prolong the lifespan of conversations in the OSN.(*iii*) Our analysis of the geographical propagation of information in Twitter reveals that users tend to send tweets from a sole geographical location. Furthermore, the level of locality associated to the social relationships varies across countries and thus for some countries like Brazil it is more likely that the information remains local than for other countries such as Australia. (*iv*) Our analysis of the load of Twitter system indicates that the arrival process of *tweets* follows a model similar to a Gaussian with a noticeable day-night pattern.

In short the work presented in this thesis allows advancing our knowledge of the Social Media ecosystem in essential directions such as the formation and growth of OSNs or the propagation of information in these systems. The important reported findings will help to develop new services on top of OSNs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Online Social Networks (OSNs) are one of the most popular Internet services nowadays. In these services the users become part of a community where they can establish relationships with other people. More than 140M people use social networks only in the US [1]. Almost two-thirds of the social network users visit a social network website at least once per day and half of them make this connection through their smartphone. Moreover, despite the huge amount of users already engaged, the use of OSNs is still growing.

These services are usually composed by websites and/or applications that allow the users to generate content and interact with other users. In these websites we can generally find a profile page with some information and content generated by each user as well as a way to connect and interact with other users. Most of the existing online social networks are designed to capture existing social networks outside the Internet. We can observe how Facebook [2] reflects the friendship network of the users while Linkdein [3] reflects the professional one. Nevertheless, other online social networks help the users connecting with strangers with the same interest, for example last.fm [4], or even allow them to connect with celebrities like Twitter [5].

The Online Social Network market is very dynamic and has changed a lot in the last years. Nowadays Facebook and Twitter clearly dominate the Social Media market while other smaller social networks still have a significant presence in particular countries like Sina Weibo [6] or Renren [7] in China or Tuenti [8] in Spain. Moreover, a Second Generation of social networks has recently appeared modifying this Social Media maket. Inside this group we can find Google+ (or G+ for short) that is supported by a major Internet player like Google.

To understand the dynamism of this complex Social Media market as well as its application in different fields such as marketing or network infrastructure enhancement different fundamental aspects must be analyzed. In this thesis we focus on two of these

aspects. On the one hand, we analyze what is the expected impact that second generation OSNs may have in this market to better understand its expected evolution. On the other hand, we carefully analyze the information propagation in major social networks considering its geographical and basic dissemination properties.

In order to achieve the previous goals it is fundamental to obtain meaningful data. In the case of OSNs that are formed by hundreds of millions users, billions of relationships and millions of contents shared daily we need to collect large-scale datasets to have a representative sample of the system properties

In the first part of this thesis we propose a measurement framework for social networks. To this end, we design highly efficient data collection and processing techniques that permit us to analyze large amounts of data in order to derive meaningful conclusions. In particular we focus in the design of crawlers for Twitter and Google+. The difficulties found are presented as well as the way in which they have been overpassed. In this part the datasets are presented and finally some lessons learned are discussed in order to help the reader in the development of a future measurement campaign of an Online Social Network.

Using the previous dataset in the Chapter 4 we address our first goal that it is to understand the birth and growth of Second Generation social network. In particular we study the birth and growth of Google+, the social network launched by Google in June 2011. Our results demonstrate that Second Generation social networks grow faster than the First Generation ones as Facebook or Twitter because the users already know this systems work and are confident to join them, nevertheless, these new OSNs can have problem to keep the users engaged due to the predominant position in the market of other popular OSNs. This is the case of G+ that is growing at an amazing rate but does not achieve to make most of these users active in the system.

Moreover, we also use some of the dataset to study some key properties of the information propagation in OSNs. We present, to the best of our knowledge, the first large scale study that characterize the full propagation of a piece of content in social networks. While some other works have been focused in the analysis of the propagation trees/cascades this is the first study that analyze the propagation forest formed by all the propagation trees sharing the same content. Furthermore, we use Twitter to understand the geographical properties of the information propagation. In particular we study the social links among users, the location from where the users use the system and the geographical distribution of the information.

Finally, we characterize the load supported by social networks. We observe the load supported by Twitter can be characterized as a Gaussian process with a noticeable day-night patter.

A more detailed motivation for these particular analysiss is presented in the next subsections.

1.1 Analyzing the social graph

The graph generated for the social relationships has been widely studied in different disciplines [19–21]. These studies have a key importance in order to understand the human relationships. Moreover, the introduction of the Online Social Networks has given to the researchers a unique opportunity to improve previous studies as well as to use the social data to improve network services (*i.e.*, to improve the content placement in a Content Distribution Network).

There exist previous studies characterizing the social graph of Twitter or Facebook [13, 22] while G+ has not been analyzed in such a deep way even when some researcher have analyzed the system in an early stage during the first months of the system [23, 24]. This lack of knowledge about the new Social Network launched by Google in June 2011 together with the contradictory information about the G+ success encourage us to study the evolution of the G+ graph.

There has been several official reports about the rapid growth of G+ user population (540M active users in Oct 2013) [25] while some observers and users dismissed these claims and called G+ a "ghost town" [26]. This raises the following important question: *"Can a new OSN such as G+ attract a significant number of engaged users and become a relevant player in the social media market?"*. A major Internet company such as Google with many popular services, is perfectly positioned to implicitly or explicitly require (or motivate) its current users to join its OSN. Then, it is interesting to assess to what extent and how Google might have leveraged its position to make users join G+. Nevertheless, any growth in the number of users in an OSN is really meaningful only if the new users adequately connect to the rest of the network (*i.e.* become connected) and become active by using some of the offered services by the OSN on a regular basis.

In this document, we present a comprehensive measurement-based characterization of connectivity among G+ users and their evolution during the first two years after its release in order to shed an insightful light on all the above questions. One of our contributions is our measurement methodology to efficiently capture complete snapshots of G+'s largest connected component (LCC) and several large sets of randomly selected users. To our knowledge, this is one of the largest and more diverse collection of datasets used to characterize an OSN. We describe our datasets in Section 3.2 along with our measurement methodology and validation techniques.

The collected datasets are analyzed in Section 4.1. Using our LLC snapshots, we characterize the evolution of LCC size during the first two years of its operation. Fur-

thermore, we leverage the randomly selected users to characterize the relative size of the main components (*i.e.* LCC, small partitions, and singletons) of G+ network and the evolutions of their relative size over time along with the fraction of active users and users with publicly visible attributes in each component. Our results show that while the size of LCC has increased at an impressive rate over the first two years of system operation, its relative size has consistently decreased such that the LCC users currently make up only 27% of the network and the rest of the users are mostly singletons.

The large and growing fraction of singletons appears to be caused by Google's integrated registration process that implicitly creates a G+ account for any new Google account regardless of the user's interest. Furthermore, we discover that LCC users generate most of the public posts and provide a larger number of attributes in their profile. Since LCC users form the most important component of G+ network, we focus the rest of our analysis on LCC.

In Section 4.1.3, we focus on the percentage of users making individual attributes in their profile publicly available. We also show that users are generally more willing to make their professional attributes publicly available but the fraction of such users has continuously decreased.

Moreover, we explore the evolution of connectivity features of LCC and show that many of its features have initially evolved but have stabilized in recent months despite the continued significant growth in its population. Interestingly, many connectivity features of the G+ network have a striking similarity with the same features in Twitter but are very different from Facebook. More specifically, the fraction of reciprocated edges among LCC users is small (and mostly associated with low degree and non-active users) and the LCC network has become increasingly less clustered.

1.2 Analyzing the user activity

While the graph properties help us to understand the user interactions as well as the nature of the network it is worthy to remark the most important of an OSN is the user activity. The analysis of the activity can help us to correctly design the system as well as to find opportunities to properly create services using the social data (*i.e.*, for marketing purposes).

Even when the number of registered users is growing a lot, it does not necessary means the activity of social network is growing. We also note that today's Internet users are much savvier about using OSN services and connecting to other users than users a decade ago when Facebook and Twitter became popular. This raises another related question: *"how does the activity of G+ users evolve over time as users have become significantly more*

experienced about using OSNs?" and "whether these evolution patterns exhibit different characteristics compared to earlier major OSNs?". These evolution patterns could also offer an insight on whether users willingly join G+ or are added to the system by Google.

In order to correctly answer the previous question we have collected every public activity for every user in the LCC in July 2013 during the first two years of the system as well as their associated number of reactions. Further details of the collected dataset and the collection methodology are presented in Section 3.2.4.

We then turn our attention to the publicly visible activity of LCC users and its evolution during the entire lifetime of G+ in Section 4.1.2. We discover that the aggregate number of posts by LCC users and their reactions (namely comments, plusones or reshares) from other users have been steadily growing over time. Furthermore, a very small fraction of LCC users generate posts and the post from an even smaller fraction of these users receive most of the reactions from other users, *i.e.*, user actions and reactions are concentrated around a very small fraction of LCC users. The average number of daily active users is growing around 670 users per day and only 17% of LCC users have ever become active. The comparison of user activity among G+, Twitter and Facebook reveals that G+ users are significantly less active than other two OSNs. More specifically, the number of G+ users who have ever become active during the first two years after the release of the system is 2.3 and 8.6 times smaller than that in MySpace and Twitter, respectively.

1.3 Analyzing the information propagation

Information propagation is an inherent property of human beings that are continuously retransmitting and sharing the information they receive with other human beings. The process of propagating the information has evolved over the history from the creation of the first human language, passing through the invention of writing, up to more recent propagation mechanism based on technology innovations such as mass media communication (e.g., radio, TV, etc). Researchers in different areas have been always interested on answering questions like *how, when or how fast the information is propagated or which persons and elements have a key role in the propagation of the information.* For example, we can find relatively old studies digging into this intriguing issue in fields like traditional media communication [27] or social science [28]. Furthermore, the irruption of the Internet have brought the modern society to the so called *Information Era* in which human beings have access to a huge volume of information as it never happened before in the History. This trend has been multiplied by the recent irruption of OSNs that have rapidly become one of the most used information propagation media for hundreds of millions of people. Therefore, the described context has defined the understanding of the information

propagation in OSNs as a topic of great relevance for the scientific community.

We can find some research efforts that study the information propagation in some of the most popular OSNs like Twitter (TW) [13], Facebook [29] or Flickr [30–32]. However all these works just use a sample of the information in those OSNs to perform the analysis of the information propagation. In this work we aim at characterizing the propagation of information in Google+ using a complete sample of the public available information in this system.

Moreover, the user's influence has been widely studied [33–40]. There does not exist a unique way to define the user's influence, in this work we define the influence of a users in the content dissemination as *"the amount of content that wouldn't be disseminated if the given users didn't exist"*. Following this definition, we claim a user A who posts an original content in G+ and obtain 10 resharers should be considered more influential than a user B who posts a very popular content (shared hundreds of times for different users) and obtain the same 10 resharers. We make this assumption because removing the user B only a small part of the distribution of this content is removed, while the content from the Online Social Network would be completely removed if the user A would not exist.

In Google+, similarly to other networks such as Facebook, the basic piece of information is the so called *post*. A post can attach different types of content (e.g., a simple text, a video, a photo, etc). The process to propagate information in Google+ occurs as follow: First, the post is initially fed into the system by a user that we refer to as *root user*. From this moment the post is available in the G+ wall of the root user and it is accessible either to all users in G+ (if the root user defines it as a public post) or to a limited number of users selected by the root user (e.g., his work colleagues). Any G+ user with access to the post can reshare it, which makes that post available in that user's wall. Then, the post is exposed to a new set of users, that in turn could decide to also reshare the post. Therefore, each post in G+ generates a propagation tree (or *reshare tree)* that constitutes the basic information propagation structure that we use for our analysis. In addition we can put together all the propagation trees sharing the same content (*i.e.*, the same Youtube video or a link to the New York Times web page) forming propagation forests.

In order to perform our study we developed a sophisticated crawling tool that allowed us to collect all public posts available in the system[1] and related information to each of them like the total number of reshares and the type of post (e.g., text, video, photo, etc). Overall, we collected 540M of posts since the release date of G+ (june 28th 2011) during a period of two years (until July 3rd 2013). Next, we leverage a public feature of G+ named Ripples [41] that provides the reshare tree of each post that has been reshared at least once and the reshare forest associate with each content that has been shared in

[1]It must be noted that we can only retrieve the information related to public posts and public reshares.

G+. In addition, the Ripple of a post/content provides detailed information such as the timestamp or the user-id associated to each reshare. Our final dataset includes almost 30M reshare trees after filtering those activities without reshares and more than 34M reshare forests[2]. We will leverage this data to carefully characterize the main aspects of information propagation in G+. In particular, we divide this work in three parts each of them addressing fundamental questions in the dissemination of information in G+.

In the first part of Section 4.2 we aim at answering the following questions: *how many people propagate a piece of information in G+?, how far a piece of information travels in G+?, how fast a piece of information travels in G+?*. To this end we study the main spatial and temporal properties associated to each one of the 30M resharers trees in our dataset. First, we study the spatial properties of the reshare trees. This is, what is the size and the height of the reshare trees in G+ that permits us to characterize the number of people that propagate a post in G+ and how far posts travel in G+. Second, we analyze two temporal metrics associated to reshare trees in order to characterize how fast information travels in G+. In particular, we refer to these metrics as *root delay* that measures the time difference between the original posting time and the time of each reshare in the tree and the *transition level delay* that captures the time that a post needs to cross a given level in its associated research tree. Furthermore, we compare the results obtained for the analysis of the spatial and temporal metrics with those obtained for TW by two different sources: our own dataset including information of more than 2.3M tweets and the results reported in [13].

In the second part of the section we focus on the reshare forests in order to understand *how many users share the same content independently? Are the most popular content ported independently for different users or they are concentrated around a small number of influential users? Do the users share content always from the same web pages? and, what is the role of the social graph in the content dissemination in G+?*. To this end we characterize the main properties of the forests following the methodology used for the reshare trees. Moreover, we study the role played by the social graph in the content dissemination.

The third part of the work propose new metrics for the user's influence taking into account the content popularity. This part provides a study case to show the importance of the content popularity in the user's influence. The ranking obtained with the new metrics is presented and compared with traditional metrics as the number of followers, the PageRank [42] or the number of comments attracted by the users activities.

This section presents four main contributions that extend the existing work: (*i*) We present the first characterization of information propagation in Google+ (G+) using all

[2]A reshare forest can be composed for more than one isolated nodes that would not form a tree for the reshare trees dataset

the publicly available information. (*ii*) To the best of our knowledge, this thesis presents the most extensive head-to-head comparison of information propagation between two major OSNs such as Google+ and Twitter. (*iii*) We present the characterization of the content dissemination in G+. To the best of our knowledge this is the first study that analyzes the dissemination of all the external contents shared in a social network. (*iv*) We propose three meaningful new metrics that takes into account the content popularity in order to rank the users using their influence in the network.

Finally, the main findings of this on this topic are:

- We confirm that only a minor fraction of the information published in OSNs is propagated. This indicates that most of the information posted in major OSNs is not interesting enough for anyone to share it.

- Although the information is propagated faster in Twitter than in G+, it gets more reshares and travels longer paths in G+. Furthermore, the probability of getting a post reshared is higher in G+ than in TW.

- Popular posts in G+ are characterized by experiencing a long lifespan rather than generating a flash crowd reaction across G+ users as it uses to happen in other systems like P2P networks. Moreover, the lifespan of the content in G+ is much bigger than the lifespan of the posts.

- Most of the popular content in G+ is independently posted by different users rather than been propagated around very influential users.

- The social graph speed up the propagation of the content in the system. Nevertheless there are more important aspects (the external popularity of the content, the G+ communities, the G+ Hot Topic list...) in the content dissemination in G+.

1.4 Analyzing the system load and the locality effect

Finally, in this thesis we analyze some key variables we should know in order to improve a OSN service. For this part of the document we will focus in Twitter since Twitter service has suffered significant loss of availability due to traffic overload or even due to malicious attacks [43–45]. Twitter has made a lot of efforts in the past year to avoid these problems. Towards this, they have doubled the capacity of their internal network and they have improved their traffic balance and monitoring system [46]. Anyway, the quickly growth of traffic demands makes this a short term solution, more so when Twitter has added a service to share pictures which may increase the traffic of their networks. The research community has shown a big interest to improve this architecture and there are some works proposing a distributed architecture for different Online Social Networks

(OSNs) [47, 48] or microblogging systems [49]. In this work we analyze first the load of the whole system and then the locality effect in Twitter.

There is little published work concerning the analysis of the traffic process of Twitter. Such a study may provide useful tools in the dimensioning, monitoring, security and fast detection of unusual events (say sudden hot topics) in Twitter. This work attempts to address this research gap by: (1) providing a measurement-based study of the tweet arrival process at Twitter on a per-hour basis; and (2) assessing the suitability of Gaussian processes to characterize the tweet arrival process.

Concerning data measurements, the Twitter REST API allows to query tweet arrivals with millisecond granularity, but the API limits the capturing process to 350 queries/hour per IP address. To accelerate the data collection process, we have implemented a distributed crawler that performs 8400 queries/hour, yielding a data set of more than one million tweet measurements in total. Such a measurement set comprises the consecutive tweet arrivals at forty-eight times of the day for Monday 16th of January, 2012, GMT+1. These times are: 00.00, 00.30, 1.00, 1.30 and so on until 23.30 (at every hour exactly and its half-past). This measurement set provides a good picture of the tweet arrival process for that day, and we further evaluate whether or not a Gaussian process can be used to model or approximate the tweet arrival process.

Moreover, understanding the Locality phenomenon of large scale systems such as p2p systems [50–53] or OSNs [54] is critical in order to improve the system design and the users performance while reducing the infrastructural and operational costs. There are also some previous work which analyze the Locality effect to improve the design and performance of the data storage system [55].

In this section we study the Locality effect for three different variables that will have influence in a future decentralized design of the Twitter architecture. These variables are:

The *User Locatity* on Twitter. Since Twitter is designed for being used in a comfortable way from a mobile phone, it is important to know whether the users of Twitter stay in few locations or they use the system from a large number of different locations. For this purpose we have collected a real dataset including more than 400K tweets with their coordinates sent by more than 22K unique users. Note that we only consider those users that have posted at least 10 tweets including geolocation information.

The *Follower Locality* in Twitter. This is, we look whether the followers of a given user are geographically concentrated, and if so we identify where. The dataset collected to study this parameter consist of the geographical location of around 1M Twitter users (or friends) and more than 16M followers associated to them.

The *Information Locality* in Twitter. This is, we look for each popular tweet (those with more than 100 retweets) the geographical distribution of their retweets. In this case, we

follow the retweets associated to more than 1.3K popular tweet obtaining the information about more than 130K users who retweet them and provide their localization information.

This work is, to the best of the authors knowledge, a first step to understand the behavior of these three variables in Twitter.

First, we analyze the *User Locality* in Twitter using three different levels of granularity according the number of cities, regions or countries from where a user posts tweets. We find that users cannot be mapped to a unique city because they usually send tweets from 2 to 5 cities, but they will be mapped to a single country. We also study the percentage of tweets sent from outside the principal city, region or country of the each user. We find that more than 70% of the users send more than a half of their tweets from only one city while around the 90% of the users post all their tweets from a unique country.

Moreover, we capture the *Follower Locality* effect with two different metrics: (i) the *link level distance* accounts the distance associated to any *friend→follower* pair, whereas (ii) the *user level distance* captures a representative metric per user such as the median distance to its followers. Therefore, the main difference is that very popular users (with many followers) weight more at *link level*, while all users have the same influence (median distance) at *user level*.

Using the described metrics we perform a country-based analysis. We have selected the country criteria since: first, we observe a high level of intra-country Locality, second, it allows to accurately group users sharing a language and a culture (which obviously influence the users' relationships in Twitter) and, finally, because our previous results showed us that we can map a user to a single country with a low error probability. Specifically, we analyze the 15 countries with a larger number of friends and followers in our dataset. The first result is the predominance of US that is responsible of around half of the friends, followers and links in our dataset. We also analyze for each of the 15 Top countries the locality at the *link level*. For this purpose, we compute the percentage of *friend→follower* links of the Twitter users of a given country that stay local within the country, go to US and go to a different country than US. We found three different profiles. On the one hand, we have countries with a quite high *intra-country* Locality effect such as Brazil that keep most of the connections local. These countries have typically a different official language than English and a strong and old culture. On the other hand, we found countries that suffer from the *external* Locality phenomenon at the *link level*. This is, the major portion of their links goes to US. These are those countries where English is the official (or co-official) language. Finally, we observe a set of countries that equally share their links among those staying local, those going to US and those going to other countries. Afterwards, we perform the Locality analysis at the user level for 4 countries, US and the most important representative of each of the defined profiles. These are Brazil, UK and France respectively. We confirm that the intra-country Locality grows

as follows among these countries Brazil>US>France>UK, which is coherent with the different profiles defined at the *link level*. Specifically, Brazil shows a surprisingly high intra-country Locality, indeed for most of Brazilian users (independently of the popularity) 80 to 90% of the followers are local. US shows a slightly lower intra-country Locality than Brazil. In UK we observe a clear bi-polarity, unpopular users show an important level of intra-country Locality whereas popular users typically experience an external Locality and have most of its followers in the US. Finally, in France we observe a similar bi-polarity with a major bias towards intra-country Locality.

Finally, we try to explain the way in which the information is geographically distributed through Twitter, this is the *Information Locality* effect in Twitter. For that, we analyze the retweets for the most popular tweets finding the surprisingly result that less than a 30% of the tweets has the half of their retweets outside their original country.

1.5 Thesis overview

The reminder of this document is organized as follow. Chapter 2 presents the relevant literature related to this thesis. In Chapter 3 the datasets used are explained, moreover, the different tools developed in order to collect these dataset are presented, explaining, first, how we have obtained the data needed to study the locality effect and the load of Twitter and, second, how we the crawling of the G+ data has been done. Finally, this chapter gives some advices that can help the reader to avoid the rate limit usually imposed for the OSNs APIs and to develop web crawlers in an efficient way. The aforementioned datasets are analyzed in Chapter 4 starting with a complete characterization of the Google+ structure and activity. Then the arrival process of messages to Twitter is analyzed and the information propagation in G+ and Twitter is compared to, finally, study the locality effect in Online Social Networks. To finish this document the conclusion obtained and some future research lines are presented in Chapter 5.

This thesis covers contributions from the following literature:

- R. Gonzalez, R. Cuevas, R. Rejaie, R. Motamedi, A. Cuevas, "Google+ or Google-?: Dissecting the evolution of the new OSN in its first year," in *Proc. of The 22nd International World Wide Web Conference. (WWW 2013), May. 2013.*.
- R. Gonzalez, A. Munoz, R. Cuevas, J. A. Hernandez, "On the Tweet arrival process at Twitter: Analysis and applications," in *Transactions on Emerging Telecommunications Technologies 2014.*
- R. Cuevas, R. Gonzalez, A. Cuevas, C. Guerrero, "Understanding the Locality Effect in Twitter: Measurement and Analysis," in *Springer Personal and Ubiquitous Computing (Special Issue in Cross-Community Mining). Online. Apr. 2013.*.

- R. Gonzalez, R. Cuevas, R. Rejaie, R. Motamedi, A. Cuevas, "Assessing The Evolution of Google+ in its First Two Years," *Transaction on Networking, Under review.*
- R. Gonzalez, R. Farahbakhsh, R. Motamedi, R. Cuevas, A. Cuevas, R. Rejaie, "Characterization of information propagation in Google+ and its Comparison with Twitter," *ACM International Conference on Information and Knowledge Management (CIKM'14). Submitted.*
- R. Motamedi, R. Gonzalez, R. Farahbakhsh, R. Rejaie, A. Cuevas, R. Cuevas, "Characterizing Group-Level User Behavior in Major Online Social Networks," *Conference on Online Social Networks (COSN 2014). Submitted.*
- R. Farahbakhsh, R. Gonzalez, R. Motamedi, A. Cuevas, R. Cuevas, R. Rejaie, "Behavior comparison of professional users across major OSNs," *ACM International Conference on Information and Knowledge Management (CIKM'14). Submitted.*

Additionally to the above, the following papers with related content have been published during the development of this thesis:

- M. Kryczka, R. Cuevas, R. Gonzalez, A. Cuevas, A. Azcorra, "TorrentGuard: stopping scam and malware distribution in the BitTorrent ecosystem," in *Elsevier Computer Networks. Accepted for publication.*
- R. Farahbakhsh, A. Cuevas, R. Cuevas, R. Gonzalez, N. Crespi, "Understanding the evolution of multimedia content in the Internet through BitTorrent glasses," in *IEEE Network. Accepted for publication..*
- I. Martinez-Yelmo, R. Gonzalez, C. Guerrero, "Validation of H-P2PSIP, a scalable solution for interoperability among different overlay networks.," in *Peer-to-Peer Networking and Applications, pp. 119, Springer New York, 2012..*
- R. Motamedi, R. Rejaie, D. Lowd, R. Gonzalez, W. Willinger, "Inferring Coarse Views of Connectivity in Very Large Graphs," *The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'14), Under Review.*
- J. M. Carrascosa, R. Gonzalez, R. Cuevas, A. Azcorra, "Are Trending Topics useful for marketing? Visibility of Trending Topics vs Traditional Advertisement," in *Proc. of ACM Conference on Online Social Networks (COSN 2013). Oct, 2013..*
- R. Farahbaksh, A. Cuevas, R. Cuevas, R. Rejaie, M. Krycka, R. Gonzalez, N. Crespi, "Investigating the Reaction of BitTorrent Content Publishers to Antipiracy Actions," in *Proc. of IEEE International Conference on Peer-to-Peer Computing (P2P 2013). Oct, 2013..*

- R. Modrzejewski, L. Chiaraviglio, I. Tahiri, F. Giroire, E. Le Rouzic, E. Bonetto, F. Musumeci, R. Gonzalez, C. Guerrero., "Energy Efficient Content Distribution in an ISP Network," in *Proc. of IEEE Global Communications Conference (GLOBECOM 2013). Dec, 2013.*

- M. Meo, Y. Zhang, Y. Hu, F. Idzikowski, . Budzisz, F. Ganji, I. Haratcherev, A. Conte, R. Bolla, O Jaramillo, R. Bruschi, A. Cianfrani, L. Chiaraviglio, A. Coiro, R. Gonzalez, C. Guerrero, E. Tego, F. Matera, S. Keranidis, G. Kazdaridis, T. Korakis, "The TREND Experimental Activities on "green" Communication Networks," in *Proc. of Tyrrhenian International Workshop 2013 on Digital Communications: Green ICT. 2013.*

- M. Lopez, I. Martinez-Yelmo, R. Gonzalez, "Analysis of relod.net, a basic implementation of the RELOAD protocol for peer-to-peer networks," in *Proc. of The XI Jornadas de Ingeniera Telemtica. (JITEL 2013), Oct. 2013.*

- I. Martinez-Yelmo, R. Gonzalez, C. Guerrero, "Benefits of an implementation of H-P2PSIP," in *Proc. of The Second International Conference on Advances in P2P Systems. (AP2PS 2010), Oct. 2010.*

Chapter 2

Related Work

The measurements and characterization of the Internet applications have attracted the attention of the research community in the last years. In this chapter we focus in the measurement studies of the large-scale applications, specifically in the ones measuring Online Social Networks.

2.1 Internet and large-scale applications workload characterisation

Many different players are interested in the analysis and modelling of network load and/or traffic of large-scale systems. For instance, Internet Service Providers (ISPs) are interested in understanding the impact of different applications in the overall traffic picture. The companies running these applications can use such information to plan and improve the infrastructure necessary for a given traffic load. Finally, the research community can design new algorithms and protocols that improve network performance, based on the knowledge acquired from measurements.

More specifically, the traffic load of relevant network applications, such as Facebook, YouTube, P2P file-sharing applications, P2P streaming applications or IPTV has been characterised in different ways. In this light, the authors of [56] studied Youtube traffic in a campus over a period of three months in 2007. A similar study was conducted by the authors in [57], and further provided a model for the system load. In the case of P2P applications, given their distributed nature and high popularity, recent studies have focused on the workload characterisation of different P2P streaming systems [58–60], and the the BitTorrent traffic associated to a large number of ISPs [61], to name a few.

Concerning the workload traffic pattern of Online Social Networks (OSNs), the authors in [62] have studied the traffic pattern of four very popular OSNs, namely Facebook, Hi5,

LinkedIN and StudiVZ, via the anonymised HTTP traces of thousands of users collected in two different ISPs. In the same direction, the authors in [63] use data from four different large Facebook communities of users in order to understand delay and load aspects of the system and study if distributed architectures could improve its performance. As in the case of YouTube, these studies only consider a minor portion of the total population of FaceBook and although they reveal important aspects of the system workload they do not estimate the overall load of the system as we do in this document.

Finally, it is also worth to mention that some studies have focused on the characterisation and modelling of network traffic in ISPs and NRENs (National Research and Education Networks) [64, 65] or the Internet Interdomain Traffic [66].

Surprisingly, there is no such work in the analysis of daily traffic patterns in Twitter, (to the best of the authors' knowledge) or the tweet arrival process.

2.2 Large-scale measurements studies in OSNs

The importance of OSNs has motivated researchers to characterize different aspects of the most popular OSNs. The graph properties of Facebook [67, 22], Twitter [68, 9] and other popular OSNs [69] have been carefully analyzed. Note that all these studies use a single snapshot of the system to conduct their analysis, instead we analyze the evolution of the G+ graph over a period of one year. In addition, some other works leverage passive (e.g., click streams) [70, 62] or active [71, 72] measurements to analyze the user activity in different popular OSNs. These papers are of different nature than ours since they use smaller datasets to analyze the behavior of individual users. Instead, we use a much larger dataset to analyze evolution of the aggregate public activity along time as well as the skewness of the contribution to overall activity across users in G+. Ding et al. proposes a collaborative way to obtain big datasets from the OSNs [73]. Finally, few works have also analyzed the users' information sharing through their public attributes in OSNs such as Facebook [74].

Previous works have separately studied the evolution of the relative size of the network elements for specific OSNs (Flickr and Yahoo 360) [75], the growth of an OSN and the evolution of its graph properties [76, 15, 77–81] or the evolution of the interactions between users [82, 83] and the user availability [84]. In this thesis, instead of looking at a specific aspect, we perform a comprehensive analysis to study the evolution of different key aspects of G+ namely, the system growth, the representative of the different network elements, the LCC connectivity and activity properties and the level of information sharing.

2.2.1 Large-scale measurement studies in Twitter

Concerning Twitter measurement studies, a number of previous studies have used the different APIs offered by Twitter to collect data and understand user behaviour in Twitter. For instance, the authors in [12] performed pioneering measurement studies on Twitter collecting data of about 100K users. Such work reported basic characteristics of Twitter users such as the correlation between the number of followers and friends of a given user or the distribution of Twitter users per continent. After this, the authors in [13] performed a Twitter graph with the relationships between 41.7 million registered users at the moment of the study (2009). The authors analysed graph topology properties as well as some other social aspects of Twitter, focusing on the social influence of certain Twitter users. Furthermore, the authors in [9] used a large dataset to analyse the dynamics of user influence across a given topic and time in Twitter. Finally, some other studies [10,11,14,85] have focused on understanding social aspects of the Twitter system.

2.2.2 Large-scale measurement studies in G+

G+ has recently attracted the attention of the research community. Mango et al. [24] use a BFS-based crawler to retrieve a snapshot of the G+ LCC between Nov and Dec 2011. They analyze the graph properties, the public information shared by users and the geographical characteristics and geolocation patterns of G+. Schiberg et al. [86] leverage Google's site-maps to gather G+ user IDs and then crawl these users' information. In particular, they study the growth of the system and users connectivity over a period of one and a half months between Sep and Oct 2011. Unfortunately, as acknowledged by the authors the described technique was not anymore available after Oct 2011. Furthermore, the authors also analyze the level of public information sharing and the geographical properties of users and links in the system. Finally, Gong et al. [23] use a BFS-based crawler to obtain several snapshots of the G+ LCC in its first 100 days of existence. Using this dataset the authors study the evolution of the main graph properties of G+ LCC in its early stage. Our work presents a broader focus than these previous works since in addition to the graph topology and the information sharing we also analyze (for first time) the evolution of both the public activity and the representativeness of the different network elements. Furthermore, our study of the graph topology evolution considers a 1 year window between Dec 2011 and Nov 2012 when the network is significantly larger and presents important differences to its early status that is the focus of the previous works. In another interesting, but less related work, Kairam et al. [87] use the complete information for more than 60K G+ users (provided by G+ administrators) and a survey including answers from 300 users to understand the selective sharing in G+. Their results show that public activity represents 1/3 of the G+ activity and that an important

fraction of users make public posts frequently. Finally, other papers have studied the video telephony system of G+ [88], the public circles feature [89], collaborative privacy management approaches [90] and the new Ripples feature [41].

2.3 Geographical aspect of the OSNs

Locality is an important aspect to be considered in large scale applications. Having it into consideration may help to improve the system design and performance as has been demonstrated for the case of p2p file-sharing applications [50, 51, 53], p2p live-streaming applications [52] or OSNs such as Facebook [54]. Although Twitter has significant different characteristics than p2p applications and slightly different than Facebook, considering the Locality effect in the system design may help to improve the performance and also the data storage procedure [55] of Twitter.

Some researchers have tried to design distributed microblogging systems. The scalability problems suffered by Twitter in the last years have attracted the attention of the research community. Some works analyse the usage a distributed solution for different services. Xu et al. have designed a distributed microblogging system called Cuckoo [49]. Before this design, Buchegger et al. presented PeerSon [47], a social network based in a peer to peer system. There are also other studies, as the one presented by Shakimov et al. [48] which is focused in privacy, cost, and availability tradeoffs in decentralized OSNs.

The position itself of the users in the OSNs has been also widely studied. Chen et al. [91] monitorize the *checkins* in Twitter and analyze the human mobility patterns. Gaito et al. [92] model the check-in behavior using social and historical ties. The authors of [93] reveal meaningful spatio-temporal patterns using about 700K Foursquare users.

2.4 Information propagation

The information propagation in the Internet has been measured and modelled in some studies [94]. De Choundry et al. [95] analyze how the data sampling strategy impacts the discovery of information diffusion in social media. Kleinberg [96] proposed a model to characterize the bursty nature of the document streams and Leskovec et al. [97] show that the most popular topics in the social media follow successive burst of popularity.

These models and measurements have been often used in order to identify the most influential users [98]. In [99] the authors demonstrate the best spreaders in complex networks are not the most connected users but the users in the *core* of the network. This fact was also demonstrated for Twitter in [9]. Some authors [100, 101] have recommend the well know *PageRank* [42] algorithm to measure the influence.

2.4.1 Improving the Internet using OSNs

The use of social data in order to improve the Internet service has been explored in the last years. Traverso et al. [102] present *TailGate* a practical system that allows the distribution of the long-tail content in an efficient way by using the social-aware scheduling algorithm. In [103], Scellato et al. study how the geographic information extracted from the social cascades can be used to know how to place multimedia content in a CDN.

Chapter 3

Measurement tools and datasets

One of the main problem researchers have to address in order to analyze different aspects of the online social networks is how to obtain enough data to consider the study representative. Three different approaches can be followed in order to solve this problem.

The best approach is always obtain the data directly from the system. This approach has been followed in many studied published by the own social networks [22,87,41,104] and in some cases also for external researchers [105, 82]. Nevertheless, it is very complicated to obtain representative datasets from the social networks for two main reasons. First, there exists clear privacy concern about use private data to third parties and this approach could be even illegal in some cases. Second, and probably more important, the companies running the online social networks consider this data very valuable and charge for the access to it [106, 107] or directly use it for advertisement purposes [108, 109].

When the option of obtain the data directly from the system is not available there still are two main possibilities in order to obtain the data needed. These approaches are to conduct passive or active measurements.

In the case of the passive measurements, the data is obtained by intercepting the communication of the users with the system. There are two basic approaches followed in this case. In the first one, if it is possible to measure over the network among the users and the social network server we can obtain the "click behaviour" [70, 62] for the users. With this technique it is possible to obtain some usage patterns but it is needed the ability to intercept the communication and the dataset obtained could be not representative of the whole system since it will represent only a subset of users usually placed close to each other. The second approach is based in install any kind of software in the final users in order to monitorize the user actions (i.e. a browser plug-in or a mobile application). In this case it is possible to obtain all the information of a determined user. Nevertheless, it is difficult to obtain a large number of users prepared to obtain the application and there exists obvious ethical concerns in use the private information.

The third option is the active measurements. In this case the data is obtained by actively querying the OSN system. Special permission or access is not needed to deploy this kind of crawlers making them the most common option for researchers without a extraordinary access to the data or ISP networks. The OSN usually provides with Application Programming Interfaces (APIs) designed to allow developers to create applications using the system. Nevertheless, these API are usually limited, either in the information provided or in the number of requests allowed. When the restrictions of the API do not allow the correct collection of the data it is still possible to obtain it by directly parsing the HTML code of the webpage or by faking the AJAX queries made by the official webpage with a web crawler. Nevertheless, the web crawlers also have limitations and on the contrary of the APIs usually the restrictions are not known before the crawling campaign.

The data used in this data has been completely obtained by using active measurement either using the system's APIs or with some custom web crawlers. The remaining of this section explain the tools developed in order to gather the data and the datasets obtained with them as well as some tips learned during the development and usage of the tools.

3.1 Twitter Crawlers and datasets

Twitter provides developers with a large API system allowing them to access the information of the system. In particular Twitter environment for developers has two independent systems the *REST API* and the *streaming API*.

The first of the systems allows applications to make queries for a specific piece of data (i.e. the profile information of an user, the list of followers of an user, the content of a tweet...) or modify it (i.e. delete a tweet) when determined permissions are granted to the application. In this API you need to query for a determined resource by using its Twitter identifier or search for the resources by using a keyword.

As many of the similar services offered for the OSN sites the *REST API* of Twitter is rate limited preventing users to obtain huge amount of information in an easy way. For this reason the usage of sophisticate tools is needed in order to avoid this limitation.

The *streaming API* works in a completely different way. In this case, the system receive a keyword and send you a stream of data containing Tweets related with this keyword in real time. This API is not rate limited, nevertheless, the service is best effort. This system is the best way to obtain a huge amount of Tweets but in practise the number of tweets obtained cannot be used as an estimator of the system load [110].

In the case of Twitter we have focused our efforts in understand two key properties of any social network. First the locality effect in the network, and for this purpose we need to collect the profile of a huge number of users, their relationships and also the bigger

number of tweets as possible. And second, the evolution of the load of the system, and for this purpose we need to collect consecutive tweets in different time moments.

3.1.1 Obtaining locality data

Our main objective is evaluate 3 different metrics: the *User, Followers*, and *Information Locality*. For this this purpose we have collected a large number of Twitter users, its geographical location, the list of its followers and then, the geographical location of them as well as a big number of tweets and retweets. This information can be obtained from the Twitter REST and Streaming API [111]. Specifically, the replay of a query for a given user-id for the REST API provides: (*i*) the user's profile information including a location-tag introduced by the user, (*ii*) a list of followers user-ids and (*iii*) other information such as the number of friends of the user and the number of tweets posted by the user so far. Rather, The Streaming API automatically send us a large number of tweets related with a certain number of terms.

Using the REST API, we have analyzed a random set of 2M users obtained from [13]. For each one of these users we have collected the geographical location of the user, the number of friends, posted tweets and followers. Furthermore we have also used the API to find the geographical location of all the followers of each analysed user. Unfortunately, Twitter limited the number of queries to be performed to 350 per hour per IP address/user-id[1]. Therefore, in order to speed up the data collection process we developed a master-slave distributed measurement architecture. This architecture counts with 1 master and 20 slaves located in different virtual machines on top of two physical machines. The master indicates to each slave the user-ids to be monitored. Moreover, each slave has its own IP address and user-id and can then perform 350 queries per hour to the Twitter API. Therefore, this distributed measurement architecture let us to perform up to 7K queries per hour. Finally, the slaves store the collected information into a redundant centralized database.

The collected user's location is the one provided by the user himself in his Twitter profile. Hence, it is not homogeneous and in some cases non-existing or meaningless. Our measurement tool filters those users that do not provide location information or provide a meaningless location. Furthermore we use the Yahoo geolocation API [112] in order to homogenize the obtained data. For instance, all those users indicating NY, NYC, New York City, etc are mapped into the same city, i.e. New York City.

In order to validate the use of the location tag we have used the Tweet Geolocation Service provided by Twitter. This service publishes along with the tweet the GPS coor-

[1]In the past Twitter gifted whitelist accounts which were allowed to perform up to 20K queries per hour. Unfortunately, these whitelist accounts are anymore available. Moreover, in the last year Twitter has changed the maximum rates to reduce the number of possible queries

dinates from where the tweet was posted. We have collected data from 140K users that have the Tweet Geolocation Service active, have a meaningful location-tag defined in their Tweeter profile and have posted at least 5 tweets with associated GPS coordinates. For each one of these users we have computed the median geographical distance between the location specified in its Twitter profile and the GPS coordinates provided in its tweets. Figure 4.37 presents the CDF of the computed distance across the analyzed users. We can observe that most of the users ($> 70\%$) typically post their tweets in a range of less than 100km from its specified location. Thus, we can conclude that in general the location-tag specified in the user's profile is a good estimator of the user location. Moreover, we can consider it even more precise if we care about a correct mapping of the user to its country as we do in this work.

We have crawled the Twitter REST API with the described distributed architecture from 10-01-2011 until 28-04-2011. The resulting dataset includes (after filtering it) 973K geolocated friends, 16.5M of geolocated followers and more than 100M of *friend→follower* relationships. This dataset has been used to analyse the *Follower Locality*.

In the case of the Streaming API, we obtain more than 400M Tweets related with some *hot topics* like "Japan", "basket" or "Obama" and with the *trending topics* existing each moment. We use 5 different virtual machines to capture this dataset from 12-03-2011 until 24-06-2011.

On the one hand, we consider all these tweets having more than 100 retweets to study the information locality effect in Twitter. This dataset is formed by 1.3K original tweets and more than 145K associated retweets.

On the other hand, to understand the *User Locality* effect, we have considered all those users posting at least 10 tweets with associated geographical coordinates. This dataset is formed by 22K different users and more than 400K tweets.

3.1.2 Obtaining the load of the system

The study of the load of an OSN is fundamental in order to understand how the design can be improved, but also in order to understand the user behaviour. While obtain the rate at which the content is consumed is almost impossible (if you are not the own OSN), obtain the rate at which the content is generated is possible in some cases by crawling consecutive Tweets. In this section we present, to the best of the author knowledge, the first system able to obtain consecutive Tweets in any determined time by taking advantage of the particular format of the Tweet IDs.

For this purpose we leverage the aforementioned REST API which, as we see before, was limited to 350 queries per hour. To overcome this problem, we have designed a

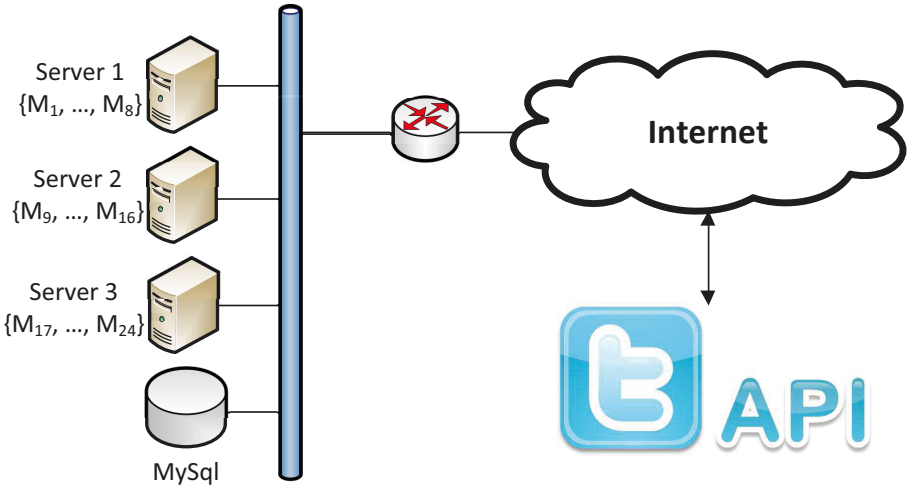


Figure 3.1: Measurement architecture review

distributed measurement infrastructure with 24 virtual machines (VMs), namely M_i, $i = 0, \ldots, 23$, each with a different IP address that collects tweets continuously (see Fig. 3.1). Hence, our measurement infrastructure generates $350 \times 24 = 8400$ queries per hour to the Twitter API at different times of the day for Monday the 16th of January 2012, GMT+1. However, this number does not translate into 8400 tweet measurements per hour, but to about half of this value, as explained in the forthcoming sections. Finally, each VM is responsible for the tweet collection at a specific hour of the same day. For example, M_1 queries for tweet arrivals at 1 a.m. and 1.30 a.m. of 16th Jan 2012 GMT+1, M_2 collects tweets at 2 a.m. and 2.30 a.m., and so on.

3.1.2.1 Tweet ID format

Before June 2010, Twitter used a sequential number to identify each message, but this revealed insufficient to cope with the exponential growth of tweet arrivals. After that, the Twitter engineers developed a new distributed system for the generation of *tweet IDs* in a more scalable way. Such a system is often referred to as Snowflake[2].

Following Snowflake, tweets are characterised by a unique 64-bit identifier, split into three fields as shown in Fig. 3.2. The first field is a 42-bit number of a custom *Timestamp* with millisecond precision. The next 10 bits mark the identifier of the machine that generated such a timestamp. This allows up to 1024 marking machines, although only five *Machine IDs* are observed from tweet ids: those with values 32 to 36. The last 12 bits comprise a *Sequence number* to allow more than one tweet with the same millisecond timestamp and marking machine.

[2]http://engineering.twitter.com/2010/06/announcing-snowflake.html

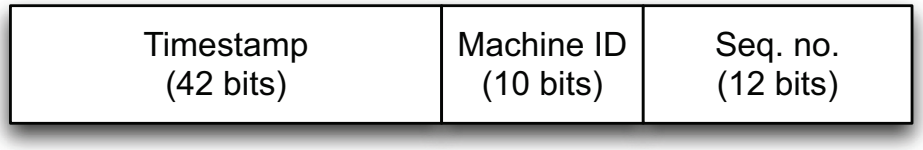


Figure 3.2: Snowflake ID schema

3.1.2.2 Measurement methodology

Algorithm 1 briefly reviews the tweet capturing process for the measurements obtained by a given virtual machine, say M_i. For example, the ID for the first tweet arrival at 10.00 a.m. on 16th Jan 2012 GMT+1, as extracted from our records, is a string of value 158850981650640896. This translates into timestamp id: 37873025334, machine id 35, seq. no. 0. Following our algorithm, M_{10} would then request the tweet whose id has the same timestamp, machine id, but seq. no. 1. After this one, the following tweet request would be that one with seq. no. 2 and so on until the API returns *Error 404*. Then, M_0 would proceed with the same timestamp, but machine id 33, seq. no. 0; then, machine id 33, seq. no. 1, and so on until a new *Error 404* message is returned. This process is repeated for all machine ids ranging from 32 to 36 to identify the number of tweet arrivals in every millisecond.

Algorithm 1 Tweet retrieving

```
time ← initTime
machineID ← 32
sequenceNumber ← 0
loop
  retrievetweetdata
  if tweetexist then
    sequenceNumber ← sequenceNumber + 1
  else
    if machineID < 36 then
      machineID ← machineId + 1
      sequenceNumber ← 0
    else
      machineID ← 32
      sequenceNumber ← 0
      time ← time + 1
    end if
  end if
end loop
```

This measuring process then produces a vector like:

$$Z_{10} = [0, 0, 1, 1, 2, 3, 4, 5, 5, 6, \ldots]$$

which gives the tweet arrival times in a millisecond time-scale for a given time of the day (10 a.m. in this case). Essentially, Z_{10} shows that two tweet arrivals occurred at $t = 0$ms (exactly 10 a.m.), two tweets arrived at $t = 1$ms (1ms after 10 a.m.), etc. Similarly, the last items of Z_{10} are:

$$[12113, 12113, 12113, 12113, 12113, 12114]$$

thus, five tweet arrivals at $t = 12113$ms, etc. Hence, this vector gives the tweet arrival times for about 12 seconds after 10.00. All Z_i data vectors contain tweet arrival times following the same structure, but with different size. For instance, Z_{10} contains information of 31964 tweets.

In conclusion, our raw data set comprises a number of vectors Z_i, where $i = 0, 1, \ldots, 23$ with the tweet arrival times at every hour, and Z_{im}, $i = 0, \ldots, 23$ with the tweet arrival times at half-past every hour. The granularity of both Z_i and Z_{im} data vectors is in the millisecond time-scale.

3.1.2.3 Crawler's measuring capacity

It is also worth benchmarking the measurement capacity of our Twitter crawler, as it follows from Alg. 1.

Let us consider the measurement vector Z_{10} of the previous section. This vector shows an average of 2.64 tweet arrivals per millisec. Following Alg. 1, it is necessary to query the API 2.64 times (one per tweet) plus another 5 extra times to get the Error 404 message for the 5 machine IDs. So, that is a total of 7.64 queries for the first millisec of Z_{10}. Hence, if our target is to get the total number of tweet arrivals for 5 seconds in every hour (both o'clock and half-past), this requires to query the REST API the following number of times:

$$(2.64 + 5) \text{ queries/ms} \times 5000 \text{ ms} = 38200 \text{ queries for } Z_{10}$$

Since our data set comprises 48 vectors (o'clock and half-past every hour), we need to query the API about: $38200 \times 48 = 1.83$ million times, just to collect 0.63 million of tweet measurements (i.e. $2.64 \times 5000 \times 48$). However, as noted before, the REST API limits the number of queries to 350 per IP address. This comprises 1.83 mill queries/350 queries/(hour*PC) = 5228 hour*PC. This means that a full day of 5s-data twice per hour (o'clock and half-past) requires a single PC launching 350

queries/hour for 5228 hours (217 days), or 24 PCs collecting data during 217 hours (9 days), as it is our case. Collecting one week of data would require the previous amount multiplied by 7, in other words, 63 days with our infrastructure.

3.1.2.4 Data set summary

Table 3.1 shows a summary of the total number of tweets in our data set, for the different times of the day. About 1.3 million tweets total have been used for this study. In the table N_i refers to the number of tweet measurements collected over T_i milliseconds. λ refers to the average number of tweets per millisecond, that is, N_i/T_i.

Hour	Z_i	N_i	T_i	λ	Hour	Z_i	N_i	T_i	λ
00:00	Z_0	61805	16741	3.6918	00:30	Z_{0m}	55500	14280	3.8866
01:00	Z_1	43106	9817	4.3910	01:30	Z_{1m}	18159	4751	3.8221
02:00	Z_2	19514	4751	4.3636	02:30	Z_{2m}	18387	4705	3.9080
03:00	Z_3	18511	4421	4.1871	03:30	Z_{3m}	18571	4666	3.9801
04:00	Z_4	22700	5549	4.0908	04:30	Z_{4m}	17585	4860	3.6183
05:00	Z_5	38535	10724	3.5933	05:30	Z_{5m}	16300	4858	3.3553
06:00	Z_6	36677	11079	3.3105	06:30	Z_{6m}	16254	5123	3.1728
07:00	Z_7	34934	11446	3.0521	07:30	Z_{7m}	14998	5239	2.8628
08:00	Z_8	14936	5169	2.8895	08:30	Z_{8m}	14412	5496	2.6223
09:00	Z_9	15168	5341	2.8399	09:30	Z_{9m}	14048	5558	2.5275
10:00	Z_{10}	31964	12114	2.6386	10:30	Z_{10m}	14337	5515	2.5996
11:00	Z_{11}	33683	11718	2.8745	11:30	Z_{11m}	15409	5306	2.9041
12:00	Z_{12}	36532	11134	3.2811	12:30	Z_{12m}	54852	17622	3.1127
13:00	Z_{13}	24228	6543	3.7029	13:30	Z_{13m}	17649	4619	3.8210
14:00	Z_{14}	18231	4502	4.0495	14:30	Z_{14m}	18440	4465	4.1299
15:00	Z_{15}	45046	9420	4.7820	15:30	Z_{15m}	19132	4318	4.4308
16:00	Z_{16}	43014	9554	4.5022	16:30	Z_{16m}	18487	4438	4.1656
17:00	Z_{17}	42845	9932	4.3138	17:30	Z_{17m}	63538	16394	3.8757
18:00	Z_{18}	41222	10136	4.0669	18:30	Z_{18m}	17731	4596	3.8579
19:00	Z_{19}	41029	10234	4.0091	19:30	Z_{19m}	17319	4546	3.8097
20:00	Z_{20}	41131	10218	4.0253	20:30	Z_{20m}	17702	4602	3.8466
21:00	Z_{21}	18924	4588	4.1247	21:30	Z_{21m}	17853	4551	3.9229
22:00	Z_{22}	18970	4589	4.1338	22:30	Z_{22m}	18214	4490	4.0566
23:00	Z_{23}	18794	4394	4.2772	23:30	Z_{23m}	63697	16405	3.8828

Table 3.1: Consecutive tweets data set summary: 16th Jan 2012, GMT+1

Fig. 3.3 shows the average tweet-arrival rate over time. As shown, between 5 a.m. and 1 p.m. GMT+1, Twitter's activity is lower than during the other times of the day.

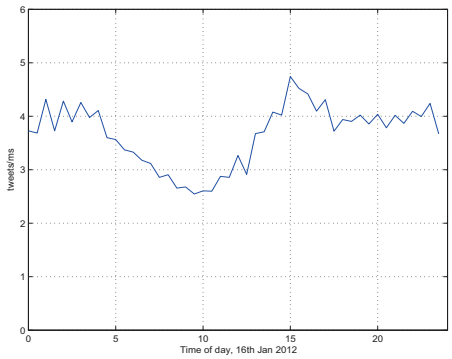


Figure 3.3: Traffic pattern on 16th Jan 2012, GMT+1

3.2 G+ Crawlers

Google has a huge API system that allows developers access the different services offered by company. Among the different API sections we can find some related with very well know services of the company as Google Maps or Youtube. Of course, Google also provides 3 different APIs for Google+, one of them to manage the Google+ Hangouts (Old Gtalk)[3], one to manage the Google+ Domains[4] and the last one to access the general information of the OSN.

The Google+ API, as the Twitter one, is a REST system which receive the unique id of an user or activity and returns some information about it.

Nevertheless, the rate limit imposed by Google in the G+ API restricts the amount of information we can obtain using this system while it seems Google allows robots to access the pages of its social networks without limits. It gives the usage of web crawlers a clear option to get some information in G+.

3.2.1 G+ Overview

After a few unsuccessful attempts (Buzz [113], Wave [114] and Orkut [115, 116]), Google launched G+ on June 28^{th} 2011 with the intention of becoming a major player in the OSNs market. Users were initially allowed to join by invitation. On September 20^{th}, G+ became open to public and the G+ Pages service was launched on November 7^{th} 2011 [117, 118]. This service imitates the *Facebook Pages* enabling businesses to connect with interested users. Furthermore, also in November 2011, the registration process was integrated with other Google services (*e.g.*, Gmail, YouTube) [119, 120].
G+ features have some similarity to Facebook and Twitter. Similar to Twitter (and

[3]http://www.google.com/+/learnmore/hangouts/
[4]https://developers.google.com/+/domains/

different from Facebook) the relationships in G+ are unidirectional. More specifically, user A can follow user B in G+ and view all of B's public posts without requiring the relationship to be reciprocated. We refer to A as B's *follower* and to B as A's *friend*. Moreover, a user can also control the visibility of a post to a specific subset of its followers by grouping them into *circles*. This feature imitates Facebook approach to control visibility of shared content. It is worth noting that this circle-based privacy setting is rather complex for average users to manage and thus unskilled users may not use it properly[5].

Each user has a stream (similar to Facebook wall) where any activity performed by the user appears. The main activity of a user is to make a "post". A post consists of some (or no) text that may have one or more attached files, called "attachments". Each attachment could be a video, a photo or any other file. Other users can react to a particular post in three different ways: *(i) Plusone*: this is similar to the "like" feature in Facebook with which other users can indicate their interest in a post, *(ii) Comment*: other users can make comments on a post, and *(iii) Reshare*: this feature is similar to a "retweet" in Twitter and allows other users to resend a post to their followers.

G+ assigns a numerical user ID and a profile to each user. The user ID is a 21-digit integer where the highest order digit is always 1 (*e.g.*, 113104553286769158393). Our examination of the assigned IDs did not reveal any clear strategy for ID assignment (*e.g.*, based on time or mod of certain numbers). Note that this extremely large ID space (10^{20}) is sparsely populated (large distance between user IDs) which in turn makes identifying valid user IDs by generating random numbers impractical. Similar to other OSNs, G+ users have a profile that has 21 fields where they can provide a range of information and pointers (*e.g.*, to their other pages) about themselves. However, providing this information is not mandatory (except for the sex) for creating an account and thus users may leave some (or all) attributes in their profile empty. Furthermore, users can limit the visibility of specific attributes (even for the sex) by defining them as "private" and thus visible to a specific group[6]. For a more detailed description of G+ functionality we refer the reader to [122, 123].

3.2.2 Capturing LCC Structure

To capture the connectivity structure of the Largest Connected Component (LCC), we use a few high-degree users as starting seeds and crawl the structure using a breadth-first search (BFS) strategy. Our initial examination revealed that the allocated users IDs are very evenly distributed across the ID space. We leverage this feature to speed

[5]A clear example of this complexity is the diagram provided to guide users to determine their privacy setting in [121].

[6]Note that it is not possible to distinguish whether a non visible attribute is private or not specified by the user.

Name	#nodes	#edges	Start Date	Duration (days)
LCC-Dec11*	35.1M	575M	11/11/12	46
LCC-Apr12	51.8M	1.1B	15/03/12	29
LCC-Aug12	79.2M	1.6B	20/08/12	4
LCC-Sep12	85.3M	1.7B	17/09/12	5
LCC-Oct12	89.8M	1.8B	15/10/12	5
LCC-Nov12	93.1M	1.9B	28/10/12	6
LCC-Dec12	105.1M	2.2B	12/11/12	8
LCC-Jan13	119.8M	2.5B	28/12/12	9
LCC-Feb13	134.8M	2.8B	11/02/13	10
LCC-Mar13	149.0M	3.0B	21/03/13	11
LCC-Apr13	155.1M	3.1B	12/04/13	11
LCC-May13	173.1M	3.5B	22/05/13	12
LCC-Jul13	190.0M	3.8B	12/07/13	13

Table 3.2: Main characteristics of LCC snapshots

up our crawler as follows: We divide the ID space into 21 equal-size zones and assign a crawler to only crawl users whose ID falls in a particular zone. Given user u in zone i, the assigned crawler to zone i collects the profile along with the list of friends and followers for user u. Any newly discovered users whose ID is in zone i are placed in a queue to be crawled whereas discovered users from other zones are periodically reported to a central coordinator. The coordinator maps all the reported users by all 21 crawlers to their zone and periodically (once per hour) sends a list of discovered users in each zone to the corresponding crawler. This strategy requires infrequent and efficient coordination with crawlers and enables them to crawl their zones in parallel. The crawl of each zone is completed when there is no more users in that zone to crawl. After some tuning, the average rate of discovery for each crawler reached 800K users per day or 16.8M users per day for the whole system[7]. With this rate, it takes 4-13 days to capture a full snapshot of the LCC connectivity and users' profiles. Table 3.2 summarizes the main characteristics of our LCC datasets. We obtained the LCC-Dec snapshot from an earlier study on G+ [24]. We examined the connectivity of all the captured LCC snapshots and verified that all of them form a single connected components.

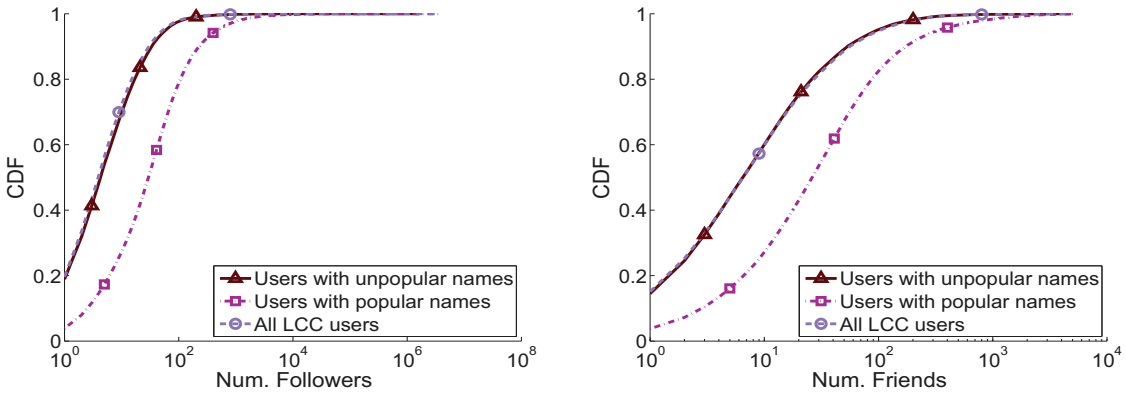
3.2.3 Sampling Random Users

Our goal is to collect random samples of G+ users for our analysis. To our knowledge, none of the prior studies on G+ achieved this goal. The sparse utilization of the extremely large ID space makes it infeasible to identify random users by generating random IDs. To cope with this challenging problem, we leverage the search function of the G+ API to efficiently identify a large number of seemingly random users. The function provides a list of up to 1000 users whose name or surname matches a given input keyword. Careful

[7]LCC-Apr snapshot was collected before this tuning and therefore took longer.

Name	#nodes	#edges	Start Date	Duration (days)
Rand-Apr12	2.2M	145M	08/04/12	23
Rand-Oct12	5.7M	263M	15/10/12	10
Rand-Nov12	3.5M	157M	28/10/12	13
Rand-Jan13	5.0M	321M	08/01/13	8
Rand-Mar13	1.1M	77M	15/03/13	4
Rand-Apr13	3.6M	249M	23/04/13	7
Rand-Jul13	3.0M	234M	12/07/13	8

Table 3.3: Main characteristics of Random datasets



(a) Num Followers (b) Num Friends

Figure 3.4: Distribution of #followers (a) and #friends (b) for users collected from the search function of G+ API with popular surnames (>1000 users), users collected with unpopular surnames (< 1000 users), and all LCC users (Reference)

inspection of search results for a few surnames revealed that G+ appears to order the reported users based on their level of connectivity and activity, *i.e.* users with a higher connectivity or activity (that are likely to be more interesting) are listed at the top of the result. Since searching for popular surnames most likely results in more than 1000 users, the reported users are biased samples. To avoid this bias, we selected a collection of 1.5K random American surnames from the US[8] 2000 census [124] with low to moderate popularity and used the search function of the API to obtain matched G+ users. We consider the list of reported users only if it contains less than 1000 users. These users are assumed to be random samples because G+ must report all matched users, and there is no correlation between surname popularity and the connectivity (or activity) of the corresponding users. Table 3.3 summarizes the main characteristics of our random datasets. Note that the timing of each one of the random datasets is aligned with a LCC dataset. To validate the above strategy, we collect two groups of more than 140K samples from the search API, users whose name match popular and unpopular (< 1000 users) surnames,

[8]US is the most represented country in G+ [24, 86]. Furthermore, the high immigration level of US allows to find surnames from different geographical regions.

Users	Postings	Attachments	Plusones	Replies	Reshares
32.5M	541M	444M	1B	408M	140M

Table 3.4: Main characteristics of Aticvities among active users in LCC (collected in Jul-Oct 2013)

in Sep 2012. We focus on samples from each group that are located in the LCC since we have a complete snapshot of the LCC that can be used as ground truth. In particular, we compare the connectivity of samples from each group that are located in LCC with all users in LCC-Sep snapshot. Figure 3.4 plots the distribution of the number of followers and friends for these two groups of samples and all users in the LCC, respectively. These figures clearly demonstrate that only the collected LCC samples from unpopular surnames exhibit very similar distributions of followers and friends with the entire LCC. A Kolgomorov-Smirnov test confirms that they are indeed the same distribution. The collected samples from popular surnames have a stronger connectivity and thus are biased.

3.2.4 Capturing Users Activity

We consider user activity as a collection of all posts by individual users and the reaction (*i.e.* Plusones, Comments and Reshares) from other users to these posts. User activity is an important indicator of user interest and thus the aggregate activity (and reactions) across users is a good measure of an OSN popularity. Despite its importance, we are not aware of any prior study that examined this issue among G+ users. Toward this end, we focus on user activity in the most important element of the network (*i.e.* the LCC). We leverage the G+ API to collect all the public posts and their associated reactions for all LCC-Jul13 users between G+ release date (Jun 28^{th} 2011) and the date our measurement campaign started (Jul 3^{th} 2013), *i.e.* roughly 2 years. Given the cumulative nature of recorded activity for each user, a single snapshot of activity contains all the activities until our data collection time. Furthermore, since each post has a timestamp, we are able to determine the temporal pattern of all posts from all users. Note, that G+ API limits the number of daily queries to 10K per registered application. Then, we use 303 accounts to collect the referred data in 68 days. Table 3.4 summarizes the main features of the activity dataset. In particular, note that only 32.4M (out of 190M) LCC-Jul13 users made at least one public post in the analysis period.

3.2.5 Activity propagation in G+

The activity unit in G+, similarly to Facebook, is the *post*. When a user publishes a post, his followers can forward (i.e., propagate) that post to their respective followers by means of a *reshare*. The followers of the followers can also reshare the post and so on.

Then, an original post along with all its reshares can be organized in a tree that we refer to as *reshare tree* or *propagation tree*. Moreover, the posts published can contain external content as a Youtube video or a link to a online newspaper. All the *propagation trees* referring to the same external content can be grouped in *propagation forests*.

By analyzing the main properties of the *reshare tree* associated to a large number of posts in G+ we can characterize the information propagation in G+. Moreover, the analysis of the *propagation forests* allows us to understand the importance of external factors (as the external popularity of the content) in the propagation of the information in the network.

In this section we describe our measurement methodology to collect all public posts and its public reshares in G+ that form the basic data to conduct our characterization study. Furthermore, we also present the filtering techniques used to process the collected data.

Our aim is to collect all posts in G+ and its associated reshare trees and the propagation forests associated to a given external content. To this end, we use all the activity IDs discovered in the previous section and we leverage a public feature of G+ named *Ripples* [41]. Each public post reshared at least once in G+ has an associated Ripple page in which the reshare tree associated to the post is available including relevant information such as the id and the language (if available) of each user, the timestamp of each reshare and the parent-child relationships within the reshare tree. We also leverage this G+ feature in order to obtain the propagation forest associated to each external content published in G+. We use a web crawler that retrieves the previous information for the reshare trees associated to each public post (with at least one reshare) obtained in the second phase.

Using the previous methodology we obtain a dataset formed by 29.6M reshare trees that overall include 90M nodes. We refer to this dataset as *G+ reshares*. Furthermore, our *G+ forests* dataset is composed for 34.7M propagation forests referring to a content that has been shared more than one time in the Online Social Network. All these propagation forests together are composed by more than 615M nodes.

Finally, we want to clarify that both the original posts and the reshares collected with our tool are public since neither the G+ API nor the Ripples provide information about private posts or reshares.

In a first manual inspection of our dataset we discovered the presence of an important fraction of large reshare trees in which the original post and most of the reshares were done by the same user. In some cases, the same user reshared its post more than 1K times. We suspect that these users are bots (an example of such users can be found at `https://plus.google.com/u/0/112555830876915762462`.)

Label	OSN	Date	Info
TW-Pro	Twitter	Jul 2011	Profile (80K rand. users)
TW-Con [9]	Twitter	Aug 2009	Connectivity (55M users)
TW-Act [80]	Twitter	Jun 2010	Activity (895K rand. users)
TW-Retweets	Twitter	Jun 2013	Activity propagation (2.3M tweets)
MS-Act [80]	MySpace	Jan 2010	Activity (239k rand. users)
FB-Pro	Facebook	Jun 2012	Profile (480K rand. users)
FB-Con	Facebook	Jun 2012	Connectivity (75K rand. users)
FB-Act	Facebook	Sep 2012	Activity (16k rand. users)

Table 3.5: Features of other datasets in our analysis

The goal of our document is to characterize the information propagation in G+ and thus if a user reshares its own post no propagation event occurs. Then, we filter all links in which the parent and child are the same user by merging both nodes in a single one in the propagation tree.

3.3 Other Dataset used

There are a few other datasets for Twitter, Facebook and MySpace that we have either collected or obtained from other researchers. Table 3.5 summarizes the main features of these datasets. In the absence of any public dataset for Facebook, we developed our own crawler and collected the profile (FB-Pro) connectivity (FB-Con) and activity (FB-Act) for random Facebook users. We also collect the profile (TW-Pro) for random Twitter users and a dataset including the number of retweets for 2.3M tweets collected from more than 17K randomly selected users (TW-retweets).

3.4 Lessons learned

During the design and development of the aforementioned tools some valuable lessons have been learned. This section describe the main tips to follow in order to properly obtain data from different online social networks.

3.4.1 Avoiding the APIs rate limit

As we told before, the access to the system APIs is usually limited in the number of queries we can perform. Moreover, the way in which the systems limit the access differ from one platform to another. Facebook or G+ APIs stablish a limitation based in user account while Twitter used to limit the number of queries made per IP address. Nevertheless, if the system have a limitation per user account usually apply also a limitation per IP address and it will ban the user accounts/IP addresses if they discover a "not human" behaviour.

If the system require us to use different user accounts we should create a big number of account as well as we need a big number of IP address. If we have enough IP addresses to improve the speed of our measurement to the desired level we can use as many virtual machines as IP addresses we have. However, this approach will not scale if we need a huge number of accounts/IP address since, on the one hand, the number of IP addresses is limited, and on the other hand, the effort needed to configure a virtual machine is not negligible and the resources needed only to run the virtual machine itself are high.

A second possible approach is based in using different proxies in order to pretend to be in different places. In this case the main problem we have to face is how to find reliable proxies. There exist a lot of different list of proxies over the internet, one of the most reliable is *Hide My Ass*[9]. Nevertheless, the proxies listed in these webpages typically start failing just a few minutes after they are published and we should control them. To avoid this churn we can install our own proxies system in a more reliable system as planetlab [125], where we count with almost 1000 different machines with different IP addresses distributed around the world.

Furthermore, every system limit the access to their API but not all of them control the access (at least in the same way) to their webpage. In particular, in Twitter or Facebook you can access to the public part of the webpage without needing to be logged in as an user but they control the robots banning the IP addresses if they detect a not human behaviour while it seems G+ does not limit the access to their webpage at all allowing us to crawl information from the system as fast as our resources allow us to do it.

3.4.2 Crawling the web

When we want to crawl the web we have first to query for the webpage and them parse the html code in order to obtain the information.

While the first part seems to be obvious, it is not that easy in all the cases. A common practise in the OSN systems nowadays is to host a skeleton of the webpage in a CDN

[9] `http://www.hidemyass.com/`

service as Akamai and then make AJAX queries in order to download the information. In this case we only download the HTML code but we do not interpret it we would not have any valuable information. In this case, the best way to obtain the information is by directly fake the AJAX queries.

In order to parse the code we can follow two different approaches, in the first one we can use an HTML parser to access the information using the Xpath, in the other one we can only search for specific patterns inside the code.

To finalize this section very useful tools used to develop our crawlers are presented:

Firebug[10] is a Firefox add-on designed to help developers debug their webpages. In our case it allows us to see the requested source code navigating across the different elements in order to obtain the Xpath or the elements we should focus on to obtain the information.

Tamper data[11] is a Firefox add-on that intercepts all the communication done by the browser. In this case this tool is basic in order to understand the AJAX interaction of the webpage with the back-end of the OSNs system.

Jsoup[12] is an HTML parser for Java that allows us to easily request and parse complete webpages.

Selenium[13] is a tools that allows to automatize actions over the browser. While this approach is not as stable as a pure Java (or any other general purpose language) software it allows us to develop simple crawlers in a very easy way over a real web browser.

[10] `https://addons.mozilla.org/en-US/firefox/addon/firebug/`
[11] `https://addons.mozilla.org/en-US/firefox/addon/tamper-data/`
[12] `http://jsoup.org/`
[13] `http://docs.seleniumhq.org/`

Chapter 4

Online Social Networks Characterization

In this chapter we analyze the datasets presented in the previous chapter. We first focus in Google+ and we perform a complete characterization of the OSN system by analyzing the evolution of the system during its first 2 years. We start our analysis understanding the composition of the Google plus network. In this case we can observe how the number of users isolated from the main component of the graph has grown during the whole year 2012 while it seems to be stable during 2013. Anyway, we can observe how the percentage of the network composed by this nodes is much bigger than in Facebook or Twitter.

We then analyze the activity in G+ and the relation of the activity with the reactions attracted and the connectivity properties of the users. Moreover, we analyze how the information propagates in Google plus comparing it with the information propagation in Twitter and analyzing the most influential users in the network.

After that, we change our view to a system that has had a lot of scalability properties in the past like Twitter. For the case of Twitter we try to understand 2 variables that have a key importance in order to design an scalable and probably distributed version of the system. In this case we first analyze the load of the system by modeling the tweet arrival process as a Gaussian distribution.

We finally analyze the locality properties of the Twitter network by studing the locality of the users, the locality of the information and the locality the relationships.

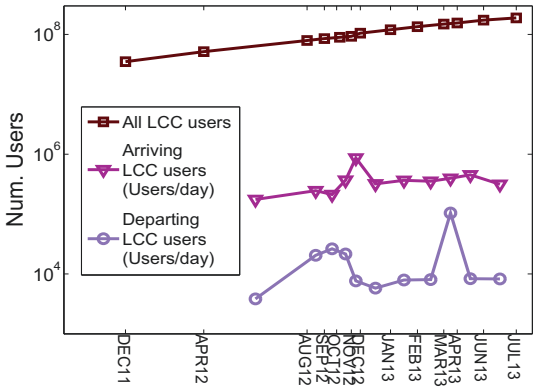


Figure 4.1: Evolution of total size and #arriving and #departing LCC users over time

4.1 Analysis of the network properties of Google Plus

4.1.1 Macro-Level Structure & Its Evolution

The macro-level connectivity structure among G+ users should intuitively consist of three components: (*i*) The largest connected component (LCC), (*ii*) A number of partitions that are smaller than LCC (with at least 2 users), and (*iii*) Singletons or isolated users. We first examine the temporal evolution of LCC size and then discuss the relative size of different components and their evolution over time.

Evolution of LCC Size: Having multiple snapshots of the LCC at different times enables us to examine the growth in the number of LCC users over time and determine the number of users who depart or arrive between two consecutive snapshots as shown in Figure 4.1 using log scale for the y axis. This figure illustrates that the overall size of the LCC has increased from 35M to 105M during 2012 at an average growth rate of 176K users per day. This average rate has even increased to 350K users per day during the first half of 2013 resulting on an average gowth rate of 263K users per day during the whole studied period (Dec 2011- Jul 2013).

The connectivity of these users to LCC is a clear sign that they have intentionally joined G+ by making the explicit effort to connect to other users (*i.e.*, these are interested users). While the average daily increase of 263K new interested users is impressive, it is roughly 3 times smaller than the average ∼650K daily new users registed in G+ between July 2011 and October 2013 obtained from official figures reported by Google [126]. The difference between the rate of growth for the overall system and LCC must be due to other components of the network (small partitions and singletons) as we explore later in this section.

We observed some short term variations in the growth rate of LCC users (as shown in Figure 4.1) which is consistent with the reported results by another recent study on another large OSN [78]. Figure 4.1 also shows that LCC users have been departing the LCC

(a) Fraction of G+ users

Element	% users						
	Apr12	Oct12	Nov12	Jan13	Mar13	Apr13	Jul13
LCC	43.5	32.3	32.2	28.1	28.0	27.4	26.9
Partitions	1.4	1.7	1.5	3.6	3.1	3.7	4.2
Singletons	55.1	66.0	66.3	68.3	69.0	69.0	68.9
All	100	100	100	100	100	100	100

(b) Fraction of G+ active users

Element	% active users (at least 1 public post)						
	Apr12	Oct12	Nov12	Jan13	Mar13	Apr13	Jul13
LCC	8.9	7.0	6.9	5.8	5.9	5.7	5.7
Partitions	0.1	0.2	0.2	0.4	0.3	0.4	0.4
Singletons	1.4	1.6	1.6	1.6	1.8	1.8	1.9
All	10.4	8.8	8.7	7.8	8.0	7.9	8.0

(c) Fraction of G+ users with public attributes

Element	% users with at least 1 public attribute						
	Apr12	Oct12	Nov12	Jan13	Mar13	Apr13	Jul13
LCC	27.4	17.9	17.6	13.7	13.0	12.4	11.2
Partitions	0.5	0.6	0.5	0.1	0.7	0.9	0.8
Singletons	1.8	5.7	6.2	3.0	2.75	2.6	2.2
All	29.7	24.2	24.3	17.6	16.4	16.0	14.2

Table 4.1: Fraction of G+ users (a), active users (b) and users with public attributes (b) across G+ components along with the evolution of these characteristics from April 2012 to July of 2013 (based on the corresponding Random datasets)

at an average rate of 10.1K users per day. We carefully examined these departing users and discovered two points: *(i)* all of the departing users have removed their G+ accounts, and *(ii)* the distribution of #followers, #friends and public attributes of departing users is very similar to all LCC users, however most of them are inactive. This seems to suggest that the departing users have lost their interest due to the lack of incentives to actively participate in the system.

Evolution of the Main Components: To estimate the relative size of each component and its evolution over time, we determine the mapping of users in a random dataset to the three main components of the G+ structure. The LCC users can be easily detected using the corresponding LCC snapshot for each random data set (*e.g.*, LCC-Oct12 for Rand-Oct12). For all the users outside the LCC, we perform a BFS crawl from each user to verify whether a user is a singleton or part of a partition, and in the latter case determine the size of the partition. Table 4.0(a) presents the relative size of all three components using our random datasets in Apr, Oct and Nov 2012 and Jan, Mar, Apr and Jul 2013[1]. The results show that the relative size of LCC has dropped from 43% (in Apr12) to 27% (in Jul13) while the relative size of singletons has increased from 55% to 69% during the same period. Note that this drop in the relative size of LCC occurs despite

[1]It is possible that our approach incorrectly categorizes user u as a singleton if u has a private list of friends and followers and, all of u's friends and followers also have a private list of followers and friends. However, we believe this is rather unlikely. Indeed our BFS crawl on the LCC identified about 7.5% users with private friend and follower lists who were detected through their neighbors.

the dramatic increase in the absolute size of LCC (as we reported earlier). This simply indicates an even more significant increase in the number of singletons. We believe that this huge increase in the number of singletons is a side effect of the integrated registration procedure that Google has implemented. In this procedure, a new G+ account is implicitly created for any user that creates a new Google account to utilize a specific Google service such as Gmail or YouTube[2]. The implicit addition of these new users to G+ suggests that they may not even be aware of (or do not have any interest in) their G+ accounts. The relatively small and decreasing size of LCC for G+ network exhibits a completely different characteristic that was reported for LCC of other major OSNs. For instance, 99.91% of the registered Facebook users were part of LCC as of May 2011 [67] and LCC of Twitter reported to include 94.8% of the users with just 0.2% Singletons in August 2009 [9]. Furthermore, Leskovec et al. [127] showed that the relative size of the LCC of other social networks (*e.g.*, the arXiv citation graph or an affiliation network) has typically increased with time until it included more than 90% of their users.
Partitions make up only a small and rather stable fraction (1.5%) of all G+ users. We identified tens of thousands of such partitions and discovered that 99% of these partitions have less than 4 users in all snapshots. The largest partition was detected in Rand-Apr snapshot with 52 users.

Tables 4.0(b) and 4.0(c) present the fraction of all G+ users that have any public posts or provide public attributes in their profiles and the breakdown of these two groups across different components of G+ network, respectively. We observe that the fraction of users that generate any post dropped from 10% to 8% during 2012 remaining stable during 2013, and the majority of them are part of LCC. Similarly, the fraction of users with any public attributes have dropped from roughly 30% to 14.2% over the same period. A large but decreasing fraction of these users are part of LCC and a smaller but growing fraction of them are singletons. Since the LCC is the well connected component that contains the majority of active users, we focus our remaining analysis only on the LCC.

In summary, the absolute size of LCC in G+ network has been growing by 150-350K users/day while its relative size has been decreasing. This is primarily due to the huge increase in the number of singletons that is caused by the implicit addition of new Google account holders to G+. In July of 2013, the LCC made up 1/4rd and the rest of the network mostly consists of singletons. Around 8% of G+ users generate any post, and less than 15% provide any public attribute, and a majority of both groups are LCC users.

4.1.2 Public Activity & Its Evolution

To investigate user activity, we characterize publicly visible (or in short "public") posts by LCC users as well as other users' reactions (including users outside LCC) to

[2]In fact, we examined and confirmed this hypothesis for new Gmail and YouTube accounts.

these public posts[3]. An earlier study used ground-truth data to show that more than 30% of posts in G+ were public during the initial phase of the system [87]. However, the proposed setting by Google encourages users to generate public posts and reactions since only these public activities are indexable by search engines (including Google), and thus visible to others (apart from Google) for various marketing and mining purposes [128]. Therefore, characterizing public posts and their reactions provides an important insight about the publicly visible part of G+.

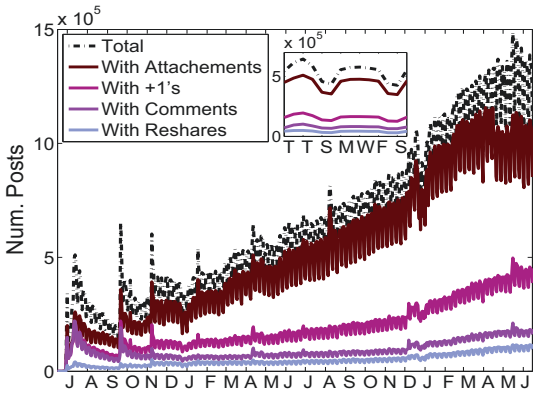
We recall that the main *action* by individual users is to generate a "post" that may have one or more "attachments". Each post by a user may trigger other users to react by making a "comment", indicate their interest by a "plusone" (+1) or "reshare" the post with their own followers. To maintain the desired crawling speed for collecting activity information, we decided to only collect the timestamp for individual posts (but not for reactions to each post). Therefore, we use the timestamp of each post as a good estimate for all of its reactions because most reactions often occur within a short time after the initial post. To validate this assumption, we have examined the timestamp of 4M comments associated to 700K posts and observed that more than 80% of the comments occurred within the 24 hours after their corresponding post.

Temporal Characteristics of Public Activity: Having the timestamp for all the posts and their associated reactions enables us to examine the temporal characteristics of all public activity among LCC users during the first 2 years of G+ operation.

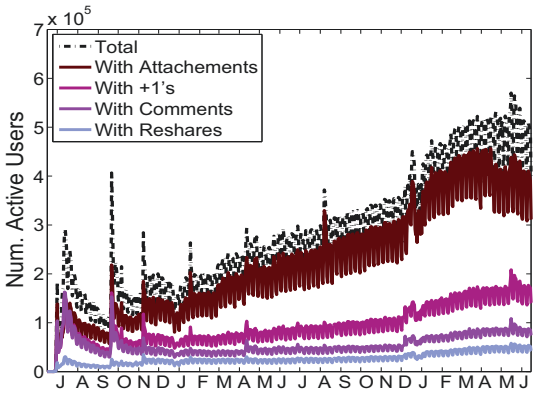
Figure 4.2(a) depicts the total number of daily posts by LCC users along with the number of daily posts that have attachments, have at least one plusone, have been reshared or have received comments. Note that a post may have any combination of attachments, plusones, reshares and comments (*i.e.*, these events are not mutually exclusive). The pronounced repeating pattern in this figure (and other similar results) is due to the weekly change in the level of activity among G+ users that is significantly lower during the weekend and much higher during weekdays as shows the inner figure in Fig. 4.2(a). The timing of most of the observed peaks in this (and other related) figure(s) appears to be perfectly aligned with specific events as follows[4]: *(i)* the peak on Jun 30 2011 caused by the initial release of the system (by invitation) [129]; *(ii)* the peak on Jul 11 2011 is due to users reaction to a major failure on Jul 9 when the system ran out of disk [130]; *(iii)* the peak on Sep 20 2011 caused by the public release of the system [129]; *(iv)* the peak on Nov 7 2011 is due to the release of G+ Pages service [118]; *(v)* the peak on Jan 17 2012 is caused by the introduction of new functionalities for auto-complete and adding text in photos [131, 132]; *(vi)* on Apr 12 2012, caused by a major redesign of G+ [133].

[3]We are not aware of any technique to capture private posts in G+ for obvious reasons. It might be feasible to create a G+ account and connect to a (potentially) large number of users in order to collect their private posts. However, such a technique is neither representative nor ethical.

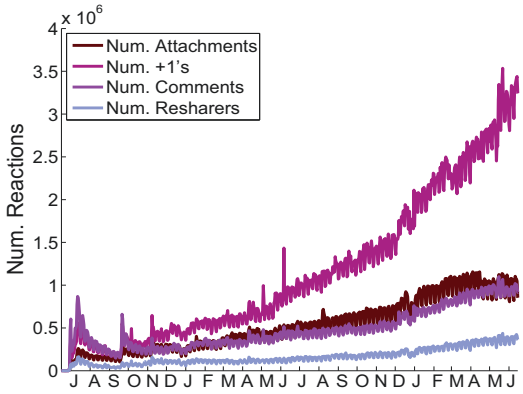
[4]We could not identify any significant event at the time of the peaks on May 3rd, Jun 4th and Aug 7th 2011



(a) The num. of daily posts, and num. of daily posts with different types of reactions



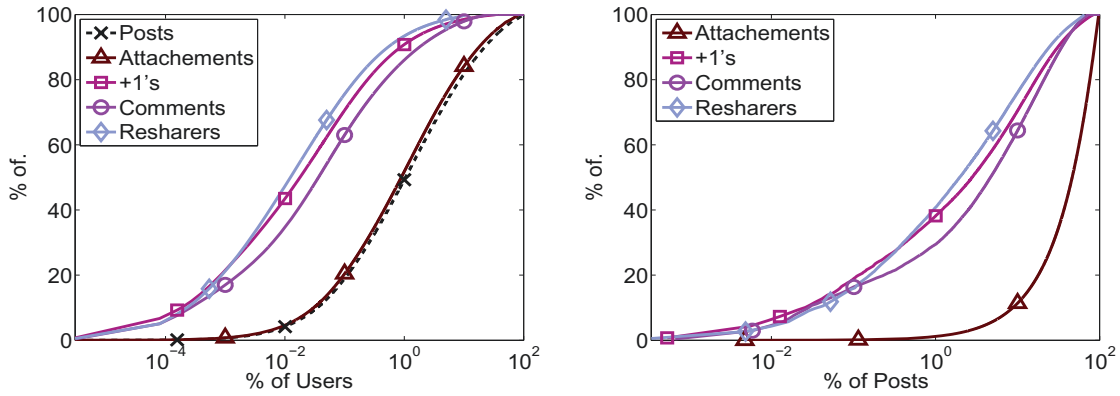
(b) The num. of daily active users and num. of daily active users whose posts received each type of reaction



(c) The num. of daily attachements, plusones, reshares and comments.

Figure 4.2: Evolution of different aspects of public user activity during the 2 years operation of G+ (July 2011 to June 2013)

Figure 4.2(a) also demonstrates that the aggregate number of daily posts has steadily increased after the first five months (*i.e.*, the initial phase of operation). We can observe that a significant majority of the posts have attachments but the fraction of posts that trigger any reaction by other users is much smaller, in addition plusones is the most common type of reaction. Note that Figure 4.2(a) presents the number of daily posts with attachment or reactions but does not reveal the total daily number of attachment or reactions. To this end, Figure 4.2(c) depicts the temporal pattern of the aggregate daily rate of attachments, plusones, comments and reshares for all the daily posts by LCC users, *i.e.*, multiple attachments or reactions to the same post are counted separately. This figure paints a rather different picture. More specifically, the total number of comments and specially plusone reactions have been rapidly growing after the initial phase. Figure 4.2(c) illustrates that individual posts are more likely to receive multiple plusones than any other type of reaction, and mostly have single attachment. Figure 4.2(b) plots the temporal



(a) % of posts, attachements, plusones, reshares, comments associated to top x% users

(b) % of attachements, plusones, reshares, comments associated to top x% posts

Figure 4.3: Skewness of actions and reactions contribution per user and post

pattern of user-level activity by showing the daily number of active LCC users along with the number of users for whom their posts have attachments or triggered at least one type of reaction. This figure reveals that the total number of daily active users with a public post has been steadily growing (after the initial phase) roughly at the rate of 670 users per day. However, this rate of growth in daily active users is significantly (roughly 392 times) lower than the daily rate of new users joining the LCC of G+. While a large fraction of these users create posts with attachments, the number of daily users whose posts trigger at least one plusone, comment or reshare has consistently remained below 200K, 100K and 50K, respectively, despite the dramatic growth in the number of LCC users.

Skewness in Activity Contribution: We observed that a relatively small and stable number of users with interesting posts receive most reactions. This raises the question that "how skewed are the distribution of generated posts and associated reactions among users in G+?". Figure 4.3(a) presents the fraction of all posts in our activity dataset that are generated by the top x% of LCC users during the life of G+ (the x axis has a log-scale). Other lines in this figure show the fraction of all attachments, plusones, comments and reshares that are associated with the top x% users that receive most reactions of each type. This figure clearly demonstrates that the contribution of the number of posts and the total number of associated attachments across users is similarly very skewed. For example, the top 10% of users contribute 82.7% of posts. Furthermore, the distribution of contribution of received reactions to a user's posts is an order of magnitude more skewed than the contribution of total posts per user. In particular, 1% of users receive roughly 86% of comments and 91% of plusones and reshares. These findings offer a strong evidence that *only a very small fraction of the active users (around 5M) create most posts and even a smaller fraction of these users receive most reactions from other users to their posts, i.e., both user action and reaction are centered around a very small fraction of users.* We also repeated a similar analysis at the post level to assess how skewed are the

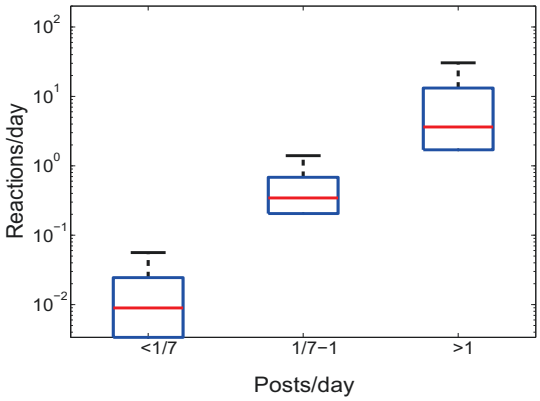
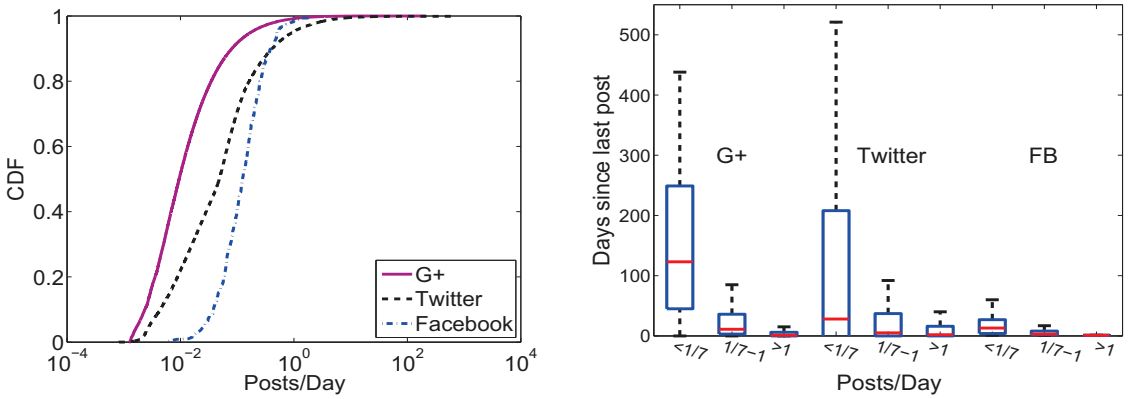


Figure 4.4: Post-rate (x axis) vs aggregate reaction rate (y axis) correlation



(a) Avg. Post/Tweet rate

(b) Recency of Activity

Figure 4.5: Comparison of activity metrics for G+, Twitter and Facebook

number of reactions to individual posts. Figure 4.3(b) shows the fraction of attachements, plusones, comments and reshares associated to the top $x\%$ posts. The distribution for attachements is rather homogeneous which indicates that most posts have one or a small number of attachments. For other types of reactions, the distribution is roughly an order of magnitude less skewed that the distribution of reaction across users (Figure 4.3(a)). This is a rather expected result since reactions tend to spread across different posts by a user.

Correlation Between User Actions and Reactions: Our analysis so far has revealed that actions and reactions are concentrated on a small fraction of LCC users. However, it is not clear whether users who generate most of the posts are the same users who receive most of the reactions. For example, a celebrity may generate a post infrequently but receives lots of reaction to each post. To answer this question, first we examine the correlation between the rate of posts and the aggregate reactions rate for different groups of users grouped based on their average level of activity as follows:

-*Active* users who post at least once a day (>1),

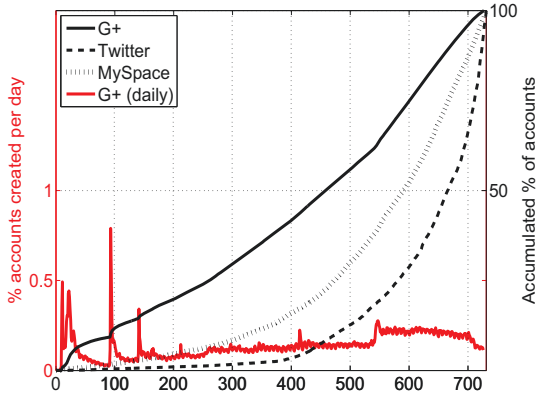


Figure 4.6: Relative number of active users in G+, Twitter and MySpace during the first two years of each OSN

	posts	plusones	comments	reshares
posts	-	0.49	0.39	0.4
plusones	0.49	-	0.55	0.46
comments	0.39	0.55	-	0.39
reshares	0.4	0.46	0.39	-

Table 4.2: Rank Correlation between actions (posts) and reaction (plusones, comments, reshares) as well as between different type of reactions associated to active users.

-*Regular* users who post less than once a day but more than once a week ($\frac{1}{7}$-1), and
-*Casual* users who post less than once a week ($< \frac{1}{7}$).
Figure 4.4 shows the summary distribution of daily reaction rate among users in each one of the described groups using boxplots. This figure reveals that the reaction rate grows exponentially with the user posting rate. Therefore, *the small group of users that contribute most posts is also receiving the major portion of all reactions.*

To gain further insight in the correlation between users' actions and reactions, Table 4.2 shows the result for the Rank Correlation (RC) -a.k.a. Spearman Correlation- [134] between users' actions and different types of reactions for our activity dataset. Note that RC shows the correlation between the rank of a group of users by two different parameters. It offers values between -1 (ranks are reversed) and 1 (ranks are the same), where 0 indicates that ranks are independent. Note, that due to the large size of our activity dataset the p-value is $\sim$0 in all cases and thus we confirm that there exist a correlation between the studied parameters. The RC reveals that *there is a notable positive correlation between users' actions (post) and the different types of reactions* (0.39-0.49).
Finally, we explore whether the capacity of users to attract different types of reactions presents any correlation. To this end, Table 4.2 presents the RC associated to each pair of reaction types in our activity dataset. We observe that the capacity of attracting plusones is highly correlated with the capacity of attracting comments (0.55) and reshares

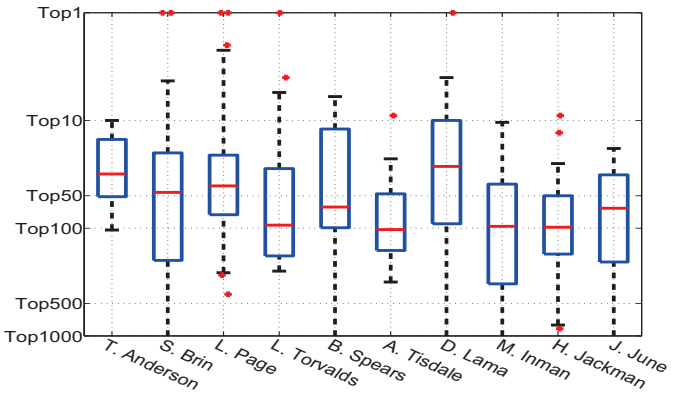


Figure 4.7: Distribution of the position in the ranking of reactions attracted for the main users of G+

(0.46). However, the correlation between comments and reshares although significant is less marked (0.39).

4.1.2.1 Identity of Most Active Users and Users Attracting More Reactions

We have identify the top 1000 users with the largest number of public post as well as those that receive a largest number of reactions each month. The analysis of the first group does not present any interesting result since the top of the publisher have a very high variation from one month to the next one and the top users are usually normal users.

For the case of the users attracting more reactions the variability is smaller. Figure 4.7 presents the distribution of the position in the Top for the 10 most common users among the months in our study. We can consider these users as the users with a more constant presence in the system. Four of these users are well known persons directly related with the Internet business, these users are consistently among the top200 indicating the importance of this kind of profiles in Google+. Moreover we can define another user in this group as Internet professional since Matthew Inman is the creator of the comic and article webpage theOatmeal.com. Inman, on the contrary of the previous ones, show important differences among different months. Inman is some montha among the top 10 while other months (like May12) he is not even in the top1000.

The other users in this group are 3 celebrities, the Dalai Lama and the adult model Jessi June. In this case, Britney Spears (one of the users with more followers in the system) and Dalai Lama are almost always among the top 100. Nevertheless, the other two celebrities show a bigger variability falling even below the top 500 in some months. It is also very surprising to find a porn start among the users attracting more reaction in G+, moreover when the activity of Jessi June in G+ seems to be the posting of semi-nude photos and people seems to react publicly to them without caring about any privacy implication.

4.1.2.2 Comparison with Other OSNs

We examine a few aspects of user activity (*i.e.*, generating posts or tweets) among G+, Twitter and Facebook users to compare the level of user engagement in these three OSNs. For this comparison, we leverage TW-Act, FB-Act datasets (described in Table 3.5) that capture activity of random users in the corresponding OSNs. In our analysis, we only consider the active users in each OSN that make up 17%, 35%, and 73% of all users in G+, Facebook and Twitter, respectively.

Activity Rate: Figure 4.5(a) shows the distribution of average activity rate per user across all active users in each OSN. The activity rate is measured as the total number of posts or tweets divided by the time between the timestamp of a user's first collected action and our measurement time. This figure reveals the following two basic points in comparing these three OSNs: *(i)* the activity rate among Facebook and G+ users are more homogeneous than across Twitter users, *(ii)* Facebook users are the most active (with the typical rate of 0.19 posts/day) while G+ users exhibit the least activity rate (with the typical rate of 0.06 posts/day).

Recency of Last Activity: An important aspect of user engagement is how often individual users generate a post. We can compute the *recency* of the last post by each active user as the time between the timestamp of last post and our measurement time. The distribution of this metric across a large number of active users provides an insight on how often active users generate a post. Figure 4.5(b) depicts the distribution of recency of the last post across G+, Twitter and Facebook users. We have divided the users from each OSN into three groups of casual, regular and active users based on their average activity rate ($< \frac{1}{7}$, $\frac{1}{7}$-1, >1 post/day) as we described earlier. We observe that among casual users in all three OSNs, Facebook and Twitter users typically generate posts much more frequently (*i.e.*, have lower median recency) than casual G+ users. Regular users in different OSNs exhibit the same relative order in their typical recency of last post. Finally, for active users, it is not surprising to observe that all three OSNs show roughly the same level of recency.

Growth Rate of Active Users: Our TW- and MySpace-Act datasets [80] include information about the evolution of the aggregate number of active users that joined TW and MySpace in the two first years after their releases. Hence, comparing these datasets with our G+ activity dataset we can derive interesting conclusions regarding the evolution of the aggregate number of users in G+ compared to two examples of 1^{st} generation OSNs.

First, if we focus in the total number of active users, two years after its release G+ accounts with 32.4M users that have been at some point active in the network. This value is 2.3 and 8.6 times larger than the equivalent in MySpace and Twitter, respectively. Hence, we can conclude that 2^{nd} generation OSNs such as G+ has been able to attract a

larger volume of active users than some of the most important 1^{st} generation OSNs. This indicates that users show a higher interest in OSNs nowadays than few years ago.
However, in addition to the volume of users, it is important to characterize the growth pattern of G+ in number of active users in comparison to Twitter and MySpace. To this end, Figure 4.6 depicts in the left-Y axis the % of active users that become active in every day of our measurement period whereas the right-Y axis represents the cumulative percentage of active users along the first two years for G+, Twitter and MySpace. The curve of % of new active users per day presents the same spikes observed in Figure 4.2. Interestingly, these spikes do not appear for Twitter or MySpace. We conjecture that this is a sign of maturity of the OSN market. OSN users have became savvy and are able to identify relevant events such as the release of a new OSN or an specific service within an OSN (e.g., the pages service) and rapidly react to these events leading to the reported spikes on volume of activity (Figure 4.2) or new users becoming active (Figure 4.6).
Furthermore, the cumulative growth rate of active users reveals a clearly different growth pattern in G+ compared to Twitter and MySpace. In particular, the growth pattern of G+ presents a much closer-to-linear shape than Twitter and MySpace. Then, the relative growth of G+ in the next months/years is expected to be significantly smaller than it was for other 1^{st} generation OSNs such as Twitter or MySpace.

In summary, the analysis of different aspects of user activity in G+ resulted in the following important points: (i) *The number of daily active LCC users has steadily grown but roughly 475 times slower than the whole LCC population.* (ii) *Around 10% of the active LCC users generate a majority of all posts and only 1/10th of these users receive most of all the reactions of any type to their posts (86% of the comments and more than 90% of the plusones and resharers). This is due to the fact that the rate of receiving reaction is strongly correlated with the user posting rate.* (iii) *The comparison of user activity for G+ with Facebook and Twitter revealed that Facebook and Twitter users exhibit a higher rate of generating posts.*(iv) *the number of active users has grown faster in G+ than other 1^{st} generation OSNs, but its growth pattern presents flash-crowd episodes and a much linear shape that leads to an expected limited relative growth in the next years. These results seems to be indication that the OSN market is becoming mature.*

4.1.3 Public User Attributes

We compare the willingness of users in different OSNs to publicly share their attributes in their profile. This is an indicator of user engagement and interest in an OSN. Roughly 48% of all the LCC users in G+ were providing at least one extra attribute different to their sex in April 2012. This ratio rapidly decreased to 44% at the endof 2012, reaching eventually 30% in our last snapshot in Jul 2013.

We further examine the distribution of the number of visible attributes across LCC

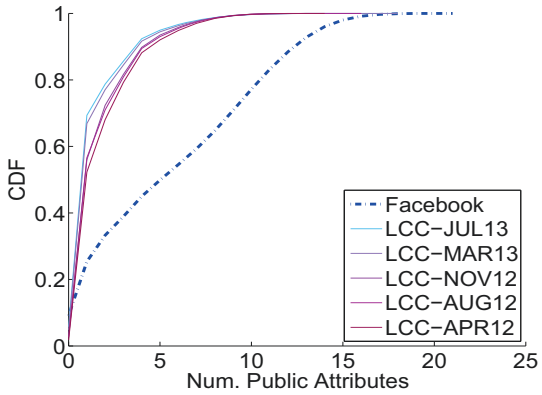


Figure 4.8: Distribution of number of public attributes for G+ and Facebook

users for different LCC snapshots and compare them with 480K random Facebook users (in FB-Pro dataset from Table 3.5) in Figure 4.8. We recall that there are 21 different attributes in both G+ and FB profiles. Figure 4.8 shows that the distribution for all LCC snapshots is very similar. Also G+ users publicly share a much smaller number of attributes compare to Facebook users. In particular, half of the users publicly share at least 6 attributes on Facebook while less than 10% of G+ users share 6 attributes. Twitter profile only has 6 attributes and 3 of them are mandatory. Examination of TW-Pro dataset shows that 69% and 13% of Twitter users share 0 and 1 non-mandatory attribute, respectively. In short, G+ users appear to share more public and non-mandatory attributes than Twitter users but significantly less than Facebook users.

Table 4.3 presents a more detailed view by showing the fraction of LCC users that provide public information in each specific field of their profile for different snapshots. As we can see the percentage of people with each attribute public is usually decreasing over the time. Only in the case of "Introduction", "Bragging Rights" and "Places Lived" we observe an increment in the percentage of people making them public until AUG12 when they also start decreasing following the general trend.

In addition, overall, users seem more inclined to share attributes related to the professional aspects such as "Studies", "Location", "Profiles" and "Profession". In contrast, they are less willing to share attributes that reveal rather more private aspects of their life such as their relationships (*e.g.*, single, married) or what they are looking for? (*e.g.*, friendship, love). This may be an indication that Google+ is being used for professional purposes (or by professional rather then average users). To double-check this hypothesis we have retrieved the identity of the 20 most popular users from Twitter, Facebook and Google+ (*i.e.*, users with more followers in Twitter and Google+, and Facebook pages with more fans) and manually inspected their professions. We observe that in Twitter and Facebook all the Top 20 publishers are *celebrities* (politicians, musicians, actors, soccer players, etc) and some companies (*e.g.*, YouTube, Twitter, FaceBook). However, in Google+ we

attribute	LCC-Dec*	LCC-Apr12	LCC-Aug12	LCC-Nov12	LCC-Mar13	LCC-Jul13
Alias	-%	0,00%	0,01%	0,01%	0,01%	0,01%
Bragging rights	3,90%	3,77%	3,93%	3,14%	2,57%	2,38%
Contact (home)	0,21%	0,21%	0,26%	0,23%	0,44%	0,55%
Contact(Work)	0,22%	0,34%	0,16%	0,25%	0,27%	0,28%
Contributor to	13,15%	11,95%	11,59%	8,10%	5,85%	4,95%
Education	27,11%	24,32%	24,72%	20,13%	16,80%	15,67%
Employment	13,27%	11,47%	13,32%	17,36%	14,92%	14,04%
Gender	97,67%	95,82%	95,76%	94,41%	93,02%	92,08%
Indexable	-%	0,00%	0,00%	0,00%	0,00%	0,00%
Introduction	7,80%	8,42%	9,74%	6,52%	5,03%	4,56%
Links	3,63%	3,26%	3,30%	2,51%	1,88%	1,68%
Looking for	2,74%	2,64%	2,61%	2,03%	1,58%	1,41%
Occupation	13,27%	11,47%	13,32%	8,93%	7,10%	6,40%
Other names	4,39%	4,08%	4,20%	3,34%	2,74%	2,50%
Other Profiles	13,48%	10,70%	10,54%	17,44%	5,87%	5,08%
Places	26,75%	26,98%	28,36%	24,15%	20,83%	19,70%
Relationship	4,31%	3,94%	3,99%	3,11%	2,63%	2,46%
Skills	-%	-%	-%	-%	0,10%	0,46%
Tagline	-%	-%	-%	-%	6,75%	6,12%
Web	-%	1,22%	1,10%	1,15%	0,81%	0,78%

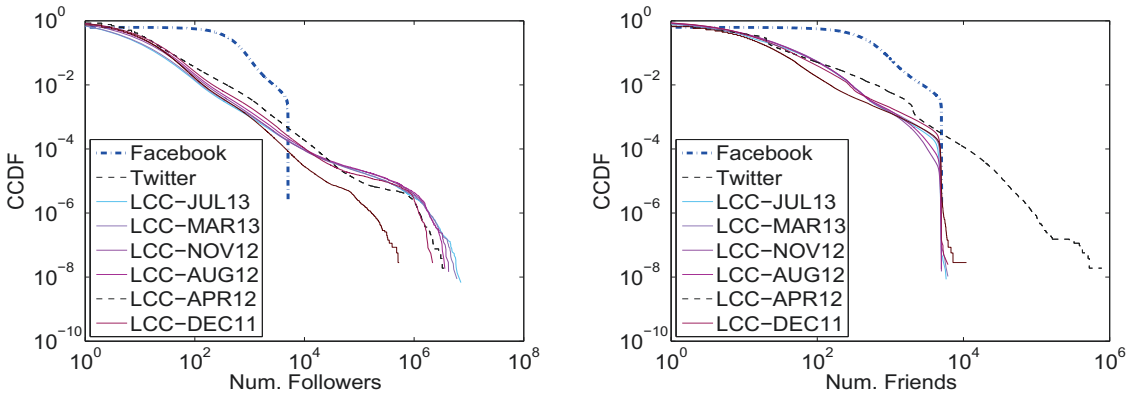
Table 4.3: Percentage of LCC users that make public each attribute for each dataset

found, along with some celebrities, professionals from the hi-tech sector (*e.g.*, Google CEO, Virgin CEO, Myspace founder), photographers or even less famous Google products as the Google Art project in the Top 20. These observations along with the results regarding the users attracting more reactions obtained in Section 4.1.2.1 confirms that Google+ seems to be acquiring a more professional focus despite of have been launched as a general purpose OSN. Some of these observations are aligned with results presented in [24].

4.1.4 LCC Connectivity & Its Evolution

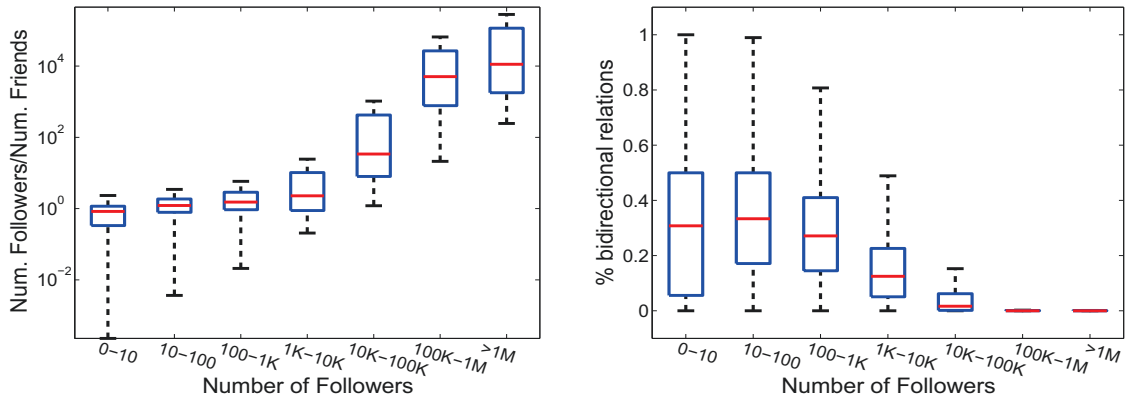
In this section, we focus on the evolution of different features of connectivity among LCC users over time as the system becomes more populated, and compare these features with other OSNs.

Degree Distribution: The distribution of node degree is one of the basic features of connectivity. Since G+ structure is a directed graph, we separately examine the distribution of the number of followers in Figure 4.9(a) and friends in Figure 4.9(b). Each figure shows the corresponding distribution across users in each one of our LCC snapshots, among Twitter users in TW-Con snapshot, and the distribution of neighbors for



(a) Num Followers (b) Num Friends

Figure 4.9: Degree Distribution for different snapshots of G+, Twitter and Facebook



(a) #followers/#friends (b) % bidirectional relationships

Figure 4.10: The level of imbalance and reciprocation for different group of users based on their number of followers.

random Facebook users in FB-Con snapshots[5]. This figure demonstrates a few important points: First, the distributions of followers and friends for G+ users can be approximated by a power law distribution with $\alpha = 1.21$ and 1.42 in LCC-Jul13 snapshot, respectively. A similar property has been reported for degree distribution of other OSNs including Twitter [68], RenRen [82], and Flickr or Orkut [69]. Second, comparing the shape of the distribution across different LCC snapshots, we observe that both distribution look very similar for all LCC snapshots. The only exception is the earliest LCC snapshot (LCC-Dec) that has a less populated tail. This comparison illustrates that the shape of both distributions has initially evolved as the LCC became significantly more populated and users with larger degree appear, and then the shape of distributions has stabilized after 14 months since G+ release. Third, interestingly, the shape of the most recent distribution

[5]Note that Facebook forces bidirectional relationships. Therefore, the distribution for Facebook in both figures is the same.

of followers and friends for G+ users is very similar to the corresponding distribution for Twitter users. The only difference appears is in the tail of the distribution of number of friends which is due to the limit of 5K friends imposed by G+ [135]. *The stability of the distribution of friends and followers for G+ users in recent months coupled with their striking similarity with these features in Twitter indicates that the degree distribution for G+ network has reached a level of maturity.* Fourth, while the distributions for Facebook are not directly comparable due to its bidirectional nature, Figure 4.9 shows that the distribution of degree for Facebook users does not follow a power law [67] as they generally exhibit a significantly larger degree than Twitter and G+ users. Specifically, 56% of Facebook users have more than 100 neighbors while only 3.6% (and 0.8%) of the G+ (and Twitter) users maintain that number of friends and followers.

Balanced Connectivity & Reciprocation: Our examination shows that the percentage of bidirectional relationships between LCC users has steadily dropped from 32% (in Dec 2011) and became rather stable in the last month of our study around 22.4% (in Jul 2013). Again, we observe that this feature of connectivity among LCC users in G+ seems to have reached a quasi-stable status after the system have experienced a major growth. Interestingly, Kwak et al. [68] reported a very similar fraction of bidirectional relationships (22%) in their Twitter snapshot from July 2009. This reveals yet another feature of G+ connectivity that is very similar to the Twitter network and very different from the fully bidirectional Facebook network. In order to gain deeper insight on this aspect of connectivity, we examine the fraction of bidirectional connections for individual nodes and its relation with the level of (im)balance between node indegree and outdegree. This in turn provides a valuable clue about the user level connectivity and reveals whether users exchange or simply relay information. To quantify the level of balance in the connectivity of individual nodes, Figure 4.10(a) plots the summary distribution of the ratio of followers to friends (using boxplots) for different group of users based on their number of followers in our most recent snapshot (LCC-Jul13). This figure demonstrates that only the low degree node (with less than 100 followers) exhibit some balance between their number of followers and friends. Otherwise, the number of friends among G+ users grows much slower than the number of followers.

We calculate the percentage of bidirectional relationships for a node u, called $BR(u)$, as expressed in Equation 4.1 where Friend(u) and Follower(u) represent the set of friends and followers for u, respectively. In essence, $BR(u)$ is simply the ratio of the total number of bidirectional relationships over the total number of unique relationships for user u.

$$BR(u) = \frac{Friend(u) \cap Follower(u)}{Friend(u) \cup Follower(u)} \tag{4.1}$$

Figure 4.10(b) presents the summary distribution of $BR(u)$ for different groups of G+ users in LCC based on their number of followers using LCC-Jul13 snapshot. The results

for other recent LCC snapshots are very similar. As expected, popular users (> 10k followers) have a very small percentage of bidirectional relationships. As the number of followers decreases, the fraction of bidirectional relationships slowly increases until it reaches around 35% for low-degree users (< 1K followers). In short, even low degree users that maintain a balanced connectivity, do not reciprocate more than 40% of their relationships. Our inspection of 5% of LCC users who reciprocate more than 90% of their edges revealed that 90% of them maintain less than 3 friends/followers and less than 5% of them have any public posts. *These results collectively suggest that G+ users reciprocate a small fraction of their relationships which is often done by very low degree users with no activity.*

Clustering Coefficient: Figure 4.11 depicts the summary distribution of the undirected version of the clustering coefficient (CC) among G+ users in different LCC snapshots

This figure clearly illustrates that during the two and a half year period (from Dec 2011 to Jul 2013), the CC among the bottom 90% of users remained below 0.6 and continuously decreases, moreover, the percentage of users with clustering coefficient 0 has grew from 20% to more than 50% in one year and a half. On the other hand, the CC for the top 10% of users has been very stable. In essence, *the G+ structure has become less clustered as new users joined the LCC over the two and a half year period.* A similar trend in cluster coefficient has been recently reported for a popular Chinese OSN [77] which indicates such an evolution in CC might be driven by underlying social forces rather than features of the OSNs. We also notice that, if we remove the growing amount of users with CC=0, the distribution of CC among G+ users also exhibit only minor changes between Aug 2012 and Jul 2013 which is another sign of stability in the connectivity features of G+ network. Compared to Twitter network where CC is less than 0.3 for 90% of users, G+ is still more clustered. Furthermore, using the approximation presented in [24], we conclude that just 1% of the nodes in a complete Facebook snapshot collected in May 2011 [67] have a CC larger than 0.2 in comparison with the 16% and 30% in Twitter and G+ (using LCC-Nov snapshot). In summary, as the population of G+ has grown, its connectivity has become less clustered but it is still the most clustered network compared to Twitter and Facebook.

Path Length: Figure 4.12 plots the probability distribution function for the pairwise path length between nodes in different LCC snapshots for G+ and a snapshot of Twitter (TW-Con). We observe that roughly 99% of the pairwise paths between G+ users are between 2 to 7 hops long and roughly 70% of them are 4 or 5 hops. The diameter of the G+ graph has increased from 17 hops (in April) to 21 hops (in July of 2013). The two visibly detectable changes in this feature of G+ graph as a result of its growth are: a small decrease in typical path length (from April 2012 to July 2013) and the increase of its diameter in the same period. Table 4.4 summarizes the average and mode path length, the

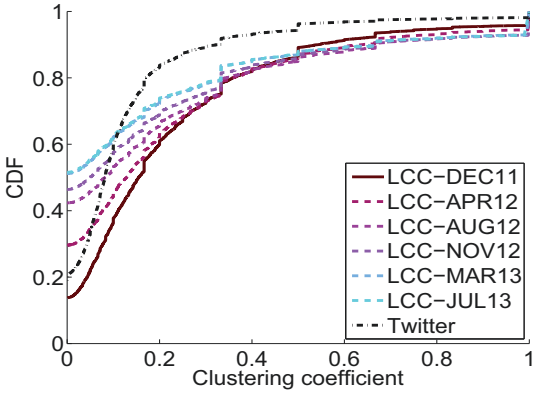


Figure 4.11: Clustering Coefficient

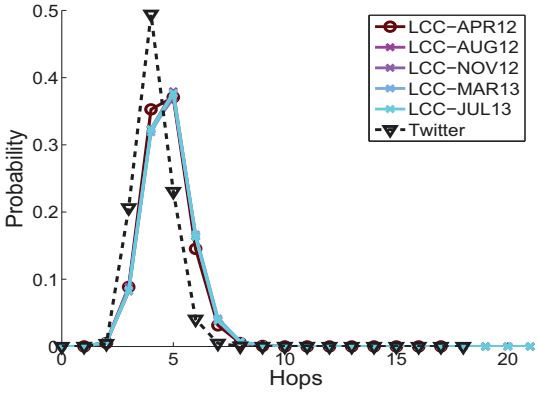


Figure 4.12: Average Path Length

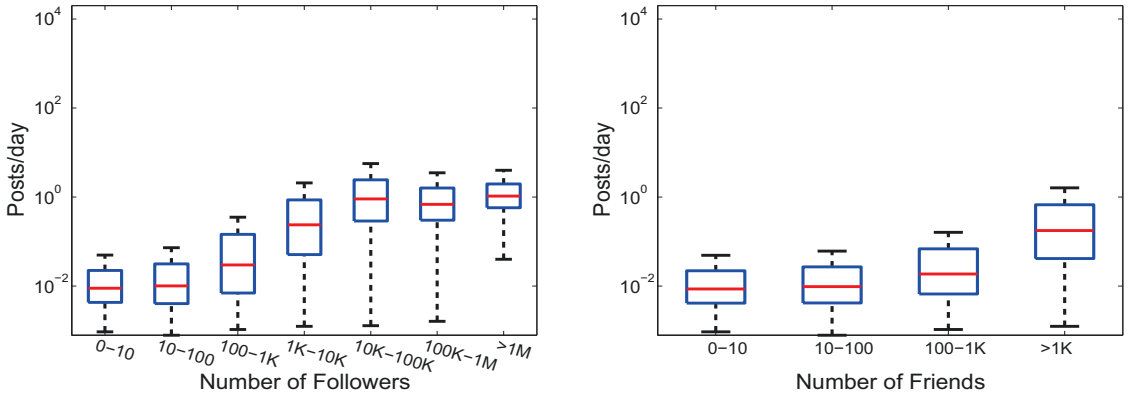
diameter and the efficient diameter [127] (*i.e.*, 90 percentile of pairwise path length) for the G+ network (using LCC-Jul13), Twitter (using TW-Con) and a Facebook snapshot from [22]. We observe that G+ and Facebook exhibit similar average (and mode) path length but Facebook has a longer diameter. One explanation is the fact that the size of Facebook network is roughly one order of magnitude larger than G+ LCC. Twitter has the shortest average and mode path length and diameter among the three. We conjecture that this difference is due to the lack of restriction in the maximum number of friends that leads to many shortcuts in the network as Twitter users connect to a larger number of friends.

	LCC-Jul13	FB	Twitter
Path Length (Avg)	4.75	4.7	4.1
Path Length (Mode)	5	5	4
Eff. Diameter	6	-	4.8
Diameter	21	41	18

Table 4.4: Summary of path length and diameter characteristics for G+, Facebook and Twitter

	num friends	num followers
num posts	0.22	0.21
num attach.	0.17	0.19
num plusones	0.25	0.33
num comments	0.34	0.33
num reshares	0.16	0.23

Table 4.5: Rank Correlation



(a) #followers vs Avg. Post Rate

(b) #friends vs Avg. Post Rate

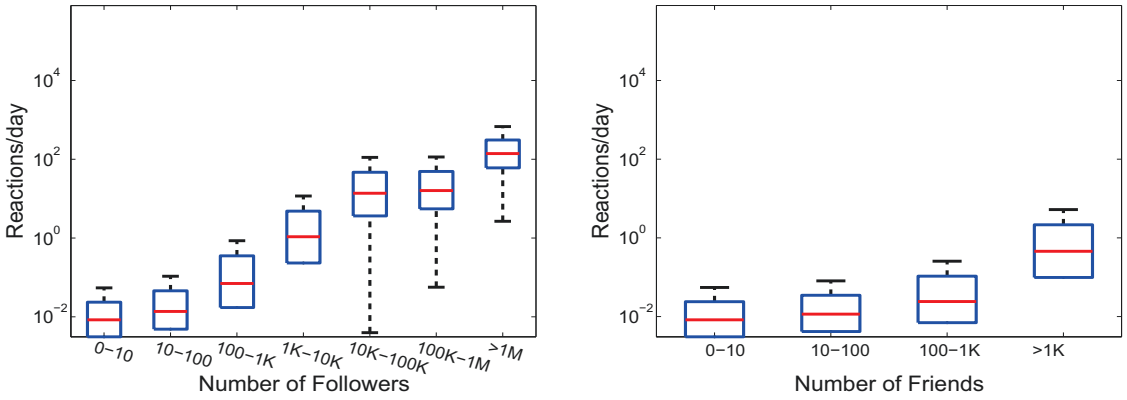
Figure 4.13: Correlation between Post Rate and Connectivity (#followers and #friends) properties in Google+

In summary, our analysis on the evolution of LCC connectivity led to the following key findings: (i) As the size of LCC significantly increased over the past year, all connectivity features of LCC (excepting the Clustering Coefficient) have initially evolved but have become rather stable in recent months despite its continued growth. (ii) Only low degree and non-active users may reciprocate a moderate fraction of their relationships. (iii) Many key features of connectivity for G+ network (e.g., degree distribution, fraction of bidirectional relationships) have striking similarity with the Twitter network and are very different from the Facebook network. These connectivity features collectively suggest that G+ is primarily used for message propagation similar to Twitter rather than pairwise users interactions similar to Facebook.

4.1.5 Relating User Activity & Connectivity

In earlier sections, we separately characterize different aspects of user activity and connectivity. One interesting question is whether and how different aspects of connectivity and activity of individual users are related. We tackle this question at both broad and more detailed levels.

To determine how correlated the connectivity of a user (#followers, #friends) are with



(a) #followers vs Avg. Reaction rate

(b) #Friends vs Avg. Reaction rate

Figure 4.14: Correlation between Aggregate Reaction Rate and Connectivity (#followers and #friends) properties in Google+

different aspects of its activity (#Posts, #Plusones, #Comments, #Reshares), we compute the Rank Correlation (RC) between all 8 pairs of these properties across active users using our last LCC snapshot and show it in Table 4.5. The results suggest that users' popularity (#followers) is more correlated with two specific types of reactions, #plusones and #of comments (0.33), than with the users direct activity, #posts (0.22). Furthermore, we observe similar results for the #friends.

To take a closer look at the relationship between user connectivity and activity, we examine how the distribution of actions and reactions among a group of users change if we divide users into groups based on their #followers or #friends. The two plots in Figure 4.13 show the summary distribution of posts/day for different groups of users based on #followers, and #friends using log scale for both axis. Figure 4.13(a) illustrates that the rate of generated posts by users rapidly increases with their number of followers and the rate of increase is especially large as we move from users with 100-1K followers to those with 10K-100K followers. Figure 4.13(b) shows that there is also a positive correlation between #friends and rate of posts. However, the rate of increase is much smaller than what we observed for grouping based on #followers in Figure 4.13(a).

Figure 4.14 presents the summary distribution of average aggregate reaction rate (*i.e.*, for 3 types of reactions) for different group of users based on #followers and #friends. Again, we observe a very strong correlation between the reaction rate to a user and its number of followers especially for users with more than 100 followers. The reaction of users does increase with the number of friends but at a much lower rate. The stronger correlation between #followers and the rate of reaction by others is reasonable since only the followers of a user see her posts (without taking any action) and thus have the opportunity to react.

In summary, given that users with many followers, that in turn are the most active, have a small fraction of bidirectional edges (as we shown in Section 4.1.4), we can confirm that G+ users use the system primarily for broadcasting information as suggested our separated study of activity and connectivity properties in previous sections.

4.2 Characterization of the information propagation in Google plus

4.2.1 Basic Characterization of Information Propagation in G+

Our goal in this section is to characterize the information propagation in G+. To this end, we analyze a set of spatial and temporal properties along with other metrics associated to the propagation trees in our *G+ reshares* dataset. Furthermore, in order to put our results into a meaningful context we compare them with those reported for TW in [13] or obtained from our *TW-reshares* dataset.

4.2.1.1 Fraction of Propagated Information

The first step to characterize the information propagation in an OSN is to understand what fraction of the available information in the system actually propagates. To this end, we have computed the percentage of posts (tweets) in our *G+ reshares* (*TW reshares*) dataset that have at least 1 reshare (retweet). The results indicate that just a small fraction of posts is propagated in both networks. In particular, only 6.8% and 3.3% of the posts/tweets are reshared in G+ and TW, respectively. However, despite both percentages are small, *it is important to highlight that the probability of getting a post reshared in G+ is roughly double than in TW*. We conjecture that this is due to the fact that the overall volume of activity is over an order of magnitude larger in TW than in G+ [136] and then TW presents a much longer tail of non-propagated tweets that leads to the reported result.

4.2.1.2 Public vs. Private information propagation

In Twitter most of the available information is public due to its broadcasting nature [13]. However, in G+ (similar to FB) users can set up different privacy configurations and decide whether their posts are public (available to anyone) or private (accessible just to some selected users). An early study revealed that around 30% of the posts published in G+ are public [87].

We can accurately compute the percentage of public reshares for each post within our *G+ reshares* dataset. As indicated in Section 3.2.4, the G+ API provides the total number of reshares (private and public) for each post whereas the Ripples functionality only reports the public reshares. Then, we can divide the number of public reshares by the number of total reshares to obtain the fraction of public reshares for each post in our dataset.

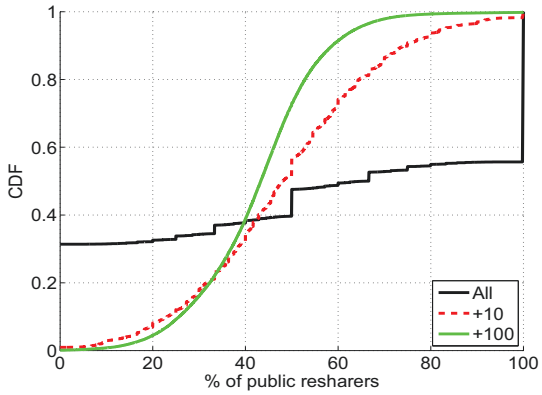


Figure 4.15: CDF of percentage of public resharers per post. We plot the results for three set of posts grouped according to the number of resahres they attract. (*i*) All posts (*All*), (*ii*) All posts with 10 or more reshares (*+10*), and (*iii*) All posts with 100 or more reshares (*+100)*

Our results indicate that, overall, 51% of the reshares in our dataset are public. This suggests that roughly half of the propagated information in G+ is disseminated in a public way. Furthermore, Figure 4.15 presents the CDF of the percentage of public reshares for the posts in our *G+ reshares* dataset. In particular, we consider three groups of posts for our analysis: *All* represents all posts that have at least 1 reshare in our dataset; *+10* includes all posts that have at least 10 reshares in our dataset (i.e., mid-popular posts) ; *+100* has all posts that have at least 100 reshares in our dataset (i.e., popular posts). For *All* we observe that more than 50% of posts have either 0 or 100% public reshares. Most of these are posts with just a single reshare that can be either private or public. In addition, as we increase the popularity (i.e., number of reshares) of the group of posts under consideration there is a reduction in the fraction of public reshares. This suggests that popular posts tend to keep a larger fraction of their propagation trees private.

Note that, unless otherwise stated, in the rest of this section we analyze the public part of the propagation trees associated to the posts within our dataset. Therefore, most of our results refer to the propagation of public information in G+, that as reported above represents roughly half of the whole public propagated information.

To the best of the authors knowledge, analyzing the propagation of private information is a very challenging task due to: first, the ethical issues associated to the collection of private information and second, the obvious difficulty of collecting private information in a scalable manner. In anycase, only public activity is indexable by search engines (including Google), and thus visible to others (different than Google) for various marketing and mining purposes [128]. Hence, characterizing the distribution of public information provides an important insight about the publicly visible part of G+ and helps to extend our knowledge about information propagation in OSNs.

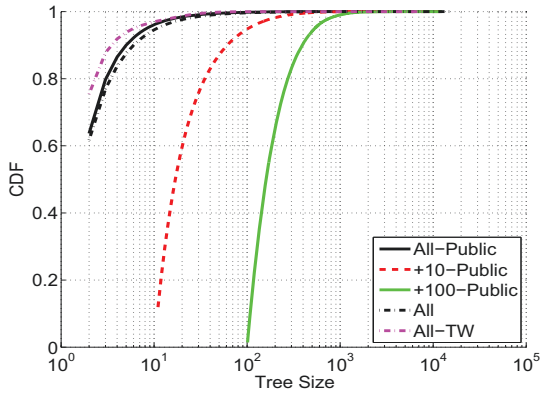


Figure 4.16: CDF of the tree size per post for different groups of posts within our *G+ reshares* and *TW retweets* datasets.

4.2.1.3 Spatial Properties of Propagation Trees in G+

In this subsection we study two spatial properties that are essential to properly characterize the information propagation phenomenon in G+:

-*Tree Size* is defined as the total number of nodes that form the propagation tree of a post. This is, the original post and all the reshares. This metric captures the popularity of a post.

-*Tree Height* is defined as the number of levels forming the longest branch of a tree. The node that publishes the original post is located at Level 1 and we refer to this node as *root node*. Nodes which reshare from the root node are located at Level 2; nodes that reshare from nodes in Level 2 are located in Level 3, an so on. Different branches of a tree may have different number of levels. The height is, then, equal to the number of levels included in the longest branch. This metric captures how far the information travels from the root node.

Tree Size Analysis Let us start by analyzing the distribution of the size of propagation trees in G+. To this end Figure 4.16 shows the CDF of the size for the propagation trees within our *G+ reshares* dataset. In particular we consider the following groups of posts: *All* includes all posts with at least 1 reshare in our dataset; *All-Public* includes all posts in our *G+ reshares* dataset with at least 1 public reshare; *+10-Public* includes the posts in our *G+ reshares* dataset with at least 10 public reshares (mid-popular posts); *+100-Public* includes the posts in our *G+ reshares* dataset with at least 100 public reshares (popular posts). Furthermore, the figure presents the distribution of the size for the propagation trees of tweets with at least 1 retweet in our *TW-retweets* dataset. We refer to this group as *TW-All*.

The results indicate that 90% of the trees have a size ≤ 5 and ≤ 6 for *All-Public* and

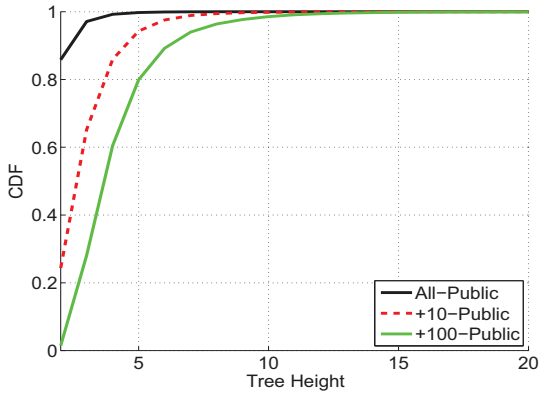


Figure 4.17: CDF of Tree Height for different groups of posts within our *G+ reshares* dataset.

All, respectively. Surprisingly, this value is 3 in the case of *All-TW*. Finally, there is not any remarkable observation to mention for *+10-Public* or *+100-Public*.

Therefore, our results indicate that the propagated information attracts more reshares in G+ than in TW.

Tree Height Analysis Now we focus in analyzing the height for the propagation trees in G+. Figure 4.17 presents the CDF of the tree height for the propagation trees in our *G+ reshares* dataset. In this case we only have information for the groups of posts including public reshares (*All-Public*, *+10-Public* and *+100-Public*). Furthermore, our *TW retweets* dataset does not include information about the height of the propagation trees. Then, we refer to the results obtained by Kwak et al. [13] for the comparison with TW.

We observe that 6.8% of *All-Public* trees have a height ≥ 1 in G+ in front of the 3.3% reported for TW. Furthermore, it is interesting to notice that the highest tree in our G+ dataset presents 129 levels whereas Kwak et al. [13] report a maximum height equal to 11 for Twitter[6].

In short, our results indicate that information travels longer paths in G+ than in TW.

4.2.1.4 Temporal Properties of Propagation Trees in G+

In this subsection we analyze the following two temporal metrics that will provide important insights in the speed of information propagation in G+:

[6]We would like to remind that the dataset of [13] was collected in 2009, three years after the release of Twitter. Our dataset has been collected two years after the release of G+.

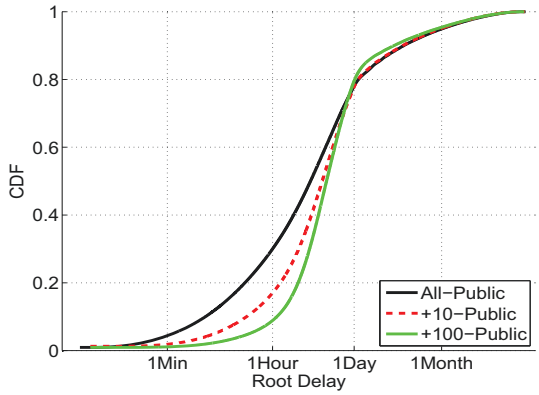


Figure 4.18: CDF of root delay. We plot the results for three sets of public posts grouped according to the number of public reshares they attract: (*i*) All public posts (*All-public*), (*ii*) posts with 10 or more public reshares (*+10-public*), and (*iii*) posts with 100 or more public reshares (*+100-public)*

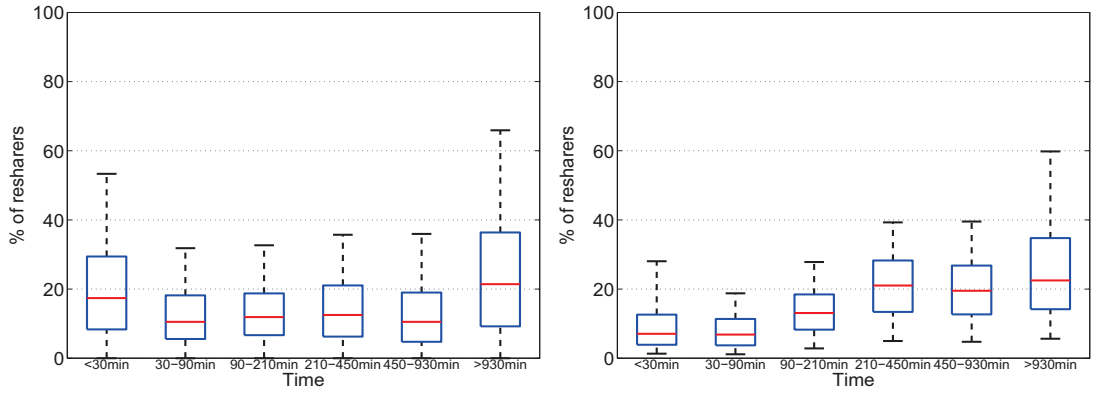
- *Root Delay* is defined as the time elapsed between the instant a node reshares a post and the original posting time. This metric captures the overall propagation delay of a post across the entire reshare tree.

- *Transition Level Delay* is computed as the time difference between the timestamps of the node's reshare and its parent's reshare. This is the time that the post needs to traverse the node's level. This metric gives us detailed information regarding the propagation time for different levels of the reshare tree.

Note that, as it occurred for the case of the tree height, we can only obtain the value of these temporal metrics for the public reshares in our dataset and then we present the results for *All-Public*, *+10-Public* and *+100-Public*. Furthermore, our *TW-retweets* dataset does not include information regarding these temporal metrics. Thus we will refer to the results reported by Kwak et al. [13] for the comparison between G+ and TW, as we did for the discussion of trees' height.

Root Delay Analysis We start our analysis of the temporal properties by looking at the root delay. Figure 4.18 shows the distribution of the root delay for nodes in *All-Public*, *+10-Public* and *+100-Public*. The results show that 80% of all public reshares happen in the first 24 hours after the original post was published and the median root delay is equal to 4.4 hours. Furthermore, Kwak et al. report a median root delay lower than 1 hour for Twitter.
Hence, we conclude that information propagates faster in Twitter than in G+.

Interestingly, groups of more popular posts represented by *+10-Public* and *+100-Public* seem to propagate more slowly. This suggests that OSNs behave differently to other popular Internet applications such as Peer-to-Peer file-sharing systems in which



(a) +10-Public (b) +100-Public

Figure 4.19: Boxplot of the percentage of reshares per tree in different delay time windows

.

popular items lead to flash-crowd events in which the users activity is concentrated close to the item publishing instant. In order to investigate this issue, we have computed the percentage of reshares for each post that occur within a given time window from the original posting time. Figure 4.19 shows the obtained results for *+10-Public* and *+100-Public*. In particular, the boxplot[7] for a given time window (e.g., 30-90 min) shows the distribution of the percentage of reshares taking place in that window for each of the posts in *+10-Public* (Figure 4.19(a)) or *+100-Public* (Figure 4.19(b)). Surprisingly, popular posts (+100) present an opposite behaviour to what we would expect from a flashcrowd reaction. Indeed, the larger fraction of reactions happen in the further time windows. In addition, for mid-popular posts (+10) we observe that the first and last time windows include around 20% of reshares while other windows in between account for 10-15% of reshares.

These results confirm that the concept of popularity in the context of information propagation in G+ (and maybe in other OSNs) translates in a longer life of the post instead of a flashcrowd reaction. To the best of the authors knowledge, this is the first time that the effect that popularity has in the temporal properties of information propagation in OSNs has been analyzed.

Transition Level Delay Analysis In this subsection we characterize the transition level delay for the group of *All-Public* posts in our *G+ reshares* dataset. In particular, we consider the levels between 2 and 10 of the propagation trees that aggregately account with 99.97% of all reshares. Figure 4.20 shows the distribution of the transition level delay for these levels in the form of boxplots.

First of all, we observe the slowest transition level delay at the second level with a median

[7]The box represents the 25, 50 and 75 percentiles and the whiskers show the 5 and 95 percentiles. Unless otherwise stated all boxplots in the thesis follow this definition.

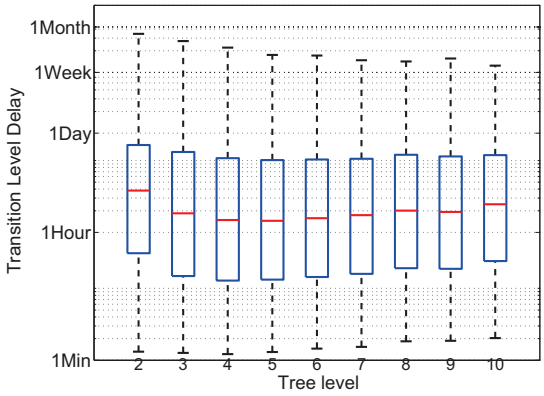


Figure 4.20: Transition Delay at different levels of the reshare tree

value of 4 hours. This level accumulates most of the reshares (91%) since all propagation trees with a single reshare (the most common) are included in this level. Following, the transition level delay continuously decreases for the next levels up to level 5 which presents a median transition level delay of around 1 and a half hour. The delay increases again from level 6 to level 10 in which its median value is roughly 2 and a half hours. Therefore, the median transition delay depicts a *convex* curve across the studied levels. Interestingly, the same *convex* pattern has been reported by Kwak et al. for Twitter. However, in the case of Twitter the median transition level delay is smaller than 1 hour for levels 2 to 10, while in G+ it ranges between 1.5 and 4 hours.
Therefore, the conducted analysis confirms that information propagates significantly faster in Twitter than in G+ also at the granularity of tree levels.

Furthermore, we want to analyze how the popularity of posts affects to the transition delay at different levels. For that purpose, Figure 4.21 shows the distribution of the transition level delay for levels 2 to 10 considering different groups of posts based on their popularity. In particular, we group the posts in the following buckets based on their associated tree size: 2-10, 10-10^2, 10^2-10^3 and 10^3-10^4.

First of all if we compare the different boxplots for a given level we observe that the transition level delay increases as we move from the lowest to the highest popularity bucket. *This confirms that in G+ the higher popularity of a post maps into a longer life span* as it has been shown by Figure 4.19.

Furthermore, if we consider the boxplots for a given popularity bucket, in general, we observe the convex evolution of the transition level delay from level 2 to level 10 reported in Figure 4.20[8].

[8]There are few exceptions such as: (*i*) in the popularity bucket "2-10" the transition delay for level 10 is significantly smaller than for other levels; (*ii*) in the popularity bucket "10^3-10^4" Levels 6 and 7 breaks the convexity of the curve.

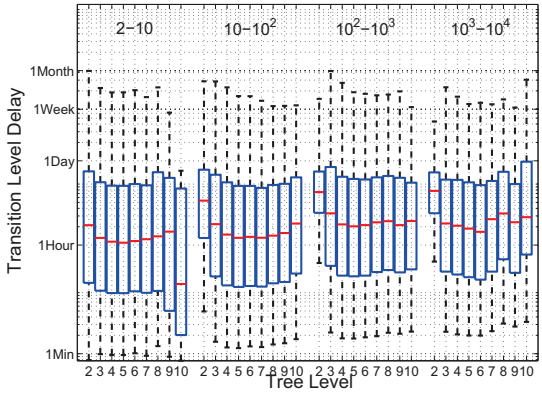


Figure 4.21: Transition Level Delay at different levels of the reshare tree for four sets of posts grouped by their popularity (i.e., number of public rehsares they attract)

4.2.1.5 Favouritism in Information Reshare

In this subsection we conduct an analysis in order to investigate whether the reshare events made by the followers[9] of a user are evenly distributed among them or, contrary, few of the followers concentrate most of the reshares of the user's posts. Furthermore, we also analyze this issue from the complementary perspective. This is, we study whether a user reshares evenly posts from all its friends[10] or contrary have a favouritism for few of them.

To address this issue we follow the methodology proposed by Kwak et. all [13] and analyze the disparity [137] in the reshare trees.

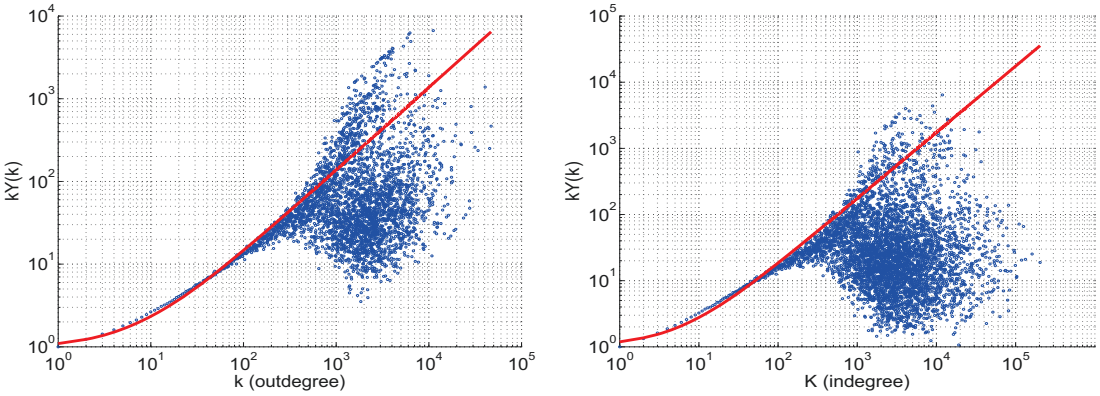
For each user i with k followers we define $|r(i,j)|$ as the number of reshares from user j. The disparity, Y(k,i), is computed as follows:

$$Y(k,i) = \sum_{j=1}^{k} \left(\frac{|r(i,j)|}{\sum_{l=1}^{k} |r(i,l)|} \right)^2 \tag{4.2}$$

Furthermore, Y(k) represents the average value of the disparity across all users with k outgoing (incoming) relationships. Note that kY(k) $\sim$ 1 indicates an homogeneous distribution whereas kY(k) $\sim$ k implies an unbalanced distribution in which few followers are responsible for most of the reshares of a user (or the user only reshares from few of its friends). Figure 4.22 shows the obtained results for the reshare trees in our *G+ reshares* dataset. We observe a linear correlation up to few hundreds followers (friends). However,

[9] In this case the set of followers of a user is composed by those users that reshared at least 1 post from the former user.

[10] In this case the set of friends of a user is form for all those users from which the former user have reshared at least one post.



(a) outdegree (b) indegree

Figure 4.22: Disparity in reshare trees

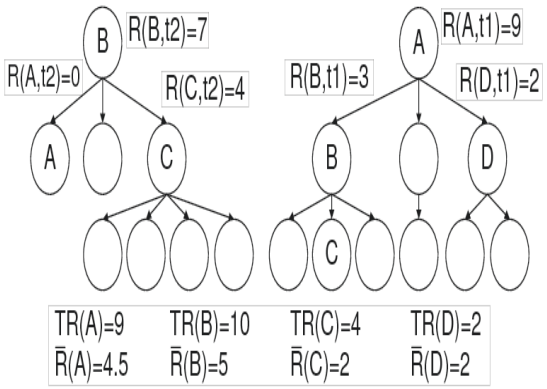


Figure 4.23: Graphical example to explain the metrics Reach, Total Reach (TR), Avg. Reach ($\overline{R}$) using two propagation trees.

the step of the associated line is 0.137 and 0.175 for the outdegree and the indegree, respectively, and thus kY(k) < k in both cases. In contrast, Kwak et al. show a linear correlation for TW in which kY(k) $\sim$ k up to 1k followers (friends).
Therefore, the obtained results demonstrate that the contribution of reshares across a user's followers is significantly more homogeneous in G+ than in TW.

4.2.1.6 Users participation in Resharers trees

The contribution of different users to the information propagation in major OSNs such as Twitter has been reported to be skewed [9]. Indeed, the capacity of a user to disseminate information is dictated by its *influence*. In this section we study the skewness in the contribution of users to the information propagation in G+.

In order to conduct our analysis we rely on a metric that we refer to as *Reach (R)*. The Reach of user u in a tree t is computed as the number of nodes in t located in the

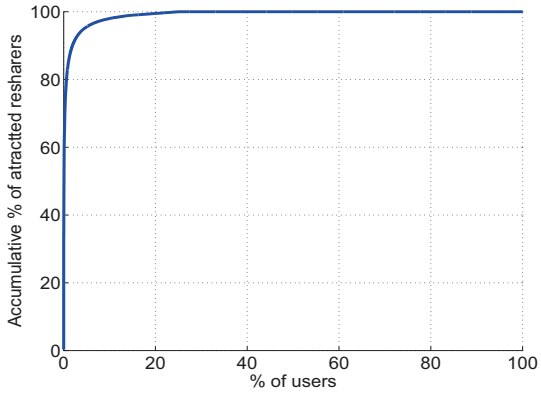


Figure 4.24: Skewness of the Total Reach across G+ users

subtree below u. If u is the root node, then $R(u,t)$ is equal to the tree size - 1.

Using this basic concept we define two metrics that capture two different types of user's influence:

-*Total Reach (TR)*: This metric is computed as the sum of the Reach of user u across all the propagation trees in which u participates. The formal expression of TR for a user u that has participated in T trees is as follows:

$$TR(u) = \sum_{t=1}^{T} R(u,t) \tag{4.3}$$

-*Avg. Reach ($\overline{R}$)*: This metric is computed as the average Reach of user u across all the propagation trees in which u participates (including those original posts from u without reshares). The formal expression of $\overline{R}$ for a user u that has participated in T trees is as follows:

$$\overline{R} = \frac{1}{T} \sum_{t=1}^{T} R(u,t) \tag{4.4}$$

In Figure 4.23 we present a graphical example with two propagation trees in order to further clarify the introduced metrics. We compute the Reach for nodes A, B, C and D in both trees and present it beside these nodes. In addition, we include a table at the bottom of the figure that shows the Total Reach and Average Reach for those nodes.

The Total Reach and the Avg Reach present complementary versions of a user's influence. On the one hand, TR of a user u captures the aggregate number of people that u has reached with all her posts and reshares. Thus, it measures the overall capacity of a user to propagate information.

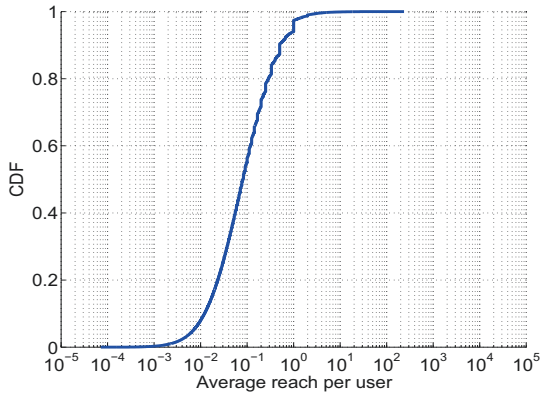


Figure 4.25: CDF of the Average Reach for G+ users

We start our analysis by studying the contribution of different users to the propagation of information in G+. A skewed contribution would reveal the presence of influential users. In particular, we study the distribution for the two defined metrics, TR and $\overline{R}$.

Figure 4.24 shows the portion of reshares included in our dataset (y-axis) associated to a given percentage of users (x-axis). In other words, it depicts the skewness of the distribution of Total Reach across G+ users. We observe that there are few users (1%) with a very high Total Reach that concentrate most of the reshares (85%).

Figure 4.25 presents the CDF of the Avg. Reach across G+ users. We observe that just 140 and 31 users present an $\overline{R}$ $\geq$50 and $\geq$100, respectively.

4.2.1.7 Summary

In this section we have characterized the main properties of information propagation in G+ and compared them with those of another major OSN such as TW. The main outcomes of our analysis are:

- A common characteristic of propagation information in major OSNs is that a very small fraction ($< 7\%$) of the information available in these systems propagates. This provides an indicator of the fraction of interesting[11] information available in major OSNs.

- Information propagates faster in Twitter than in G+ but it gets more rehsares and travels longer paths in G+. To explain this phenomenon we leverage the results from [136] that demonstrate that the overall daily volume of information available is over an order of magnitude higher in TW than in G+. In other words, a user in Twitter is exposed to a higher volume of information that changes more frequently. Then, it is more likely that a user changes his attention to a different conversation, or simply misses some information

[11] "Interesting" refers to a piece of information interesting enough for someone to share it with his/her friends or followers.

due to the high frequency of new received tweets (i.e., a user who does not connect to Twitter in a period of few hours may miss some tweets of his interest that were published during this time). Furthermore, the contribution of reshares across the followers of a user is more homogeneously distributed in G+ than in TW.

- The higher popularity of a post in G+ translates into a longer lifespan of that post in the system. As future work, it would be interesting to analyze this aspect in other major OSNs in order to confirm if this is a common property of major OSNs. This property differentiates G+ (and possibly other OSNs) from other popular applications in the Internet (e.g., p2p file-sharing) in which popularity is mapped into flash-crowd events.

- We observe a very skewed distribution for the Total Reach and Average Reach among the G+ users. It indicates only a few users are attracting the attention of the network.

4.2.2 Basic Characterization of the Content Propagation in G+

In this section we characterize how a given external content propagates on the network. For this purpose we analyze some spatial properties of the propagation forests generated when different users post the same URL in G+. It is important to remark the trees composing the propagation forests are a subset of the trees analyzed in the previous subsection.

We use the results obtained in this subsection in order to shed light on the importance of the external factors in the information propagation inside a social network.

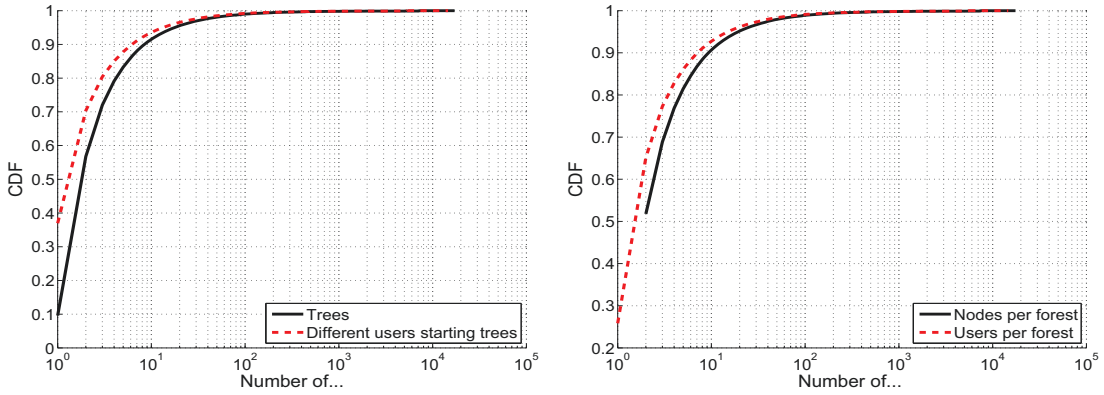
4.2.2.1 Spatial Properties of Propagation Forests in G+

In this subsection we study two of the main spatial properties of the propagation forests in G+:

-*Number of trees* per forest is the number of times a content has been originally posted in the social network. This variable give us an intuition of the external content popularity, since the social network activity does not affect to the number of times an external content is originally posted.

-*Forest size* is the number of nodes forming the forest. This includes, the original posts and all the reshares generated by any of them. This metric captures the popularity of a given content inside the social network.

It is worthy to remark from the more than 113M out of the 148M forests in our dataset have been shared by a single users who have not attracted any reshare. We do not take these "single node" forest into account for the following analysis.



(a) Num. of Trees per Forest (b) Forest size

Figure 4.26: Forest Composition

Number of trees per forest Figure 4.26(a) shows the CDF for the number of trees per forest and the number of different root users generating these trees. We can observe only 5% of the forest contains only one tree. It gives us a clue that the propagation of a given content depends in external factors since if a content is popular enough to attract the attention of more than one person it usually will be originally shared more than one time. Moreover, only 6.7% of the forests have more than 10 trees and only a 0.8% of them have more than 100. Finally, the widest forest includes more than 10K trees

Focusing now in the number of users who originally share the same content we can observe for about 30% of the forests an unique user have originated all the reshared trees. Since we have previously observe that only a 15% of the forests contains only one tree it seems that at least for 15% of the forests under study a single user is sharing more than one time the same content. Nevertheless as in the case of the number of trees we can found forest in which more than 10K users are originally sharing the same content.

Forest Size Next we analyze our other spatial metric, the *forest size*. Figure 4.26(b) present the CDF of the number of nodes per forest as well as the CDF for the number of different users participating in each forest. It is possible to see almost 55% of the nodes are composed for only two nodes. Moreover, the distribution is very similar to the distribution obtained for the number of trees. It means the number of trees and nodes per forest is usually very similar, this is, the forest are composed for very small trees.

Forest Size vs. Number of Trees The previous results suggest the number of trees and the size of the forests is very similar, indicating most of the forest are composed by very small trees. To confirm this intuition in this subsection we analyze the relation between the two metrics.

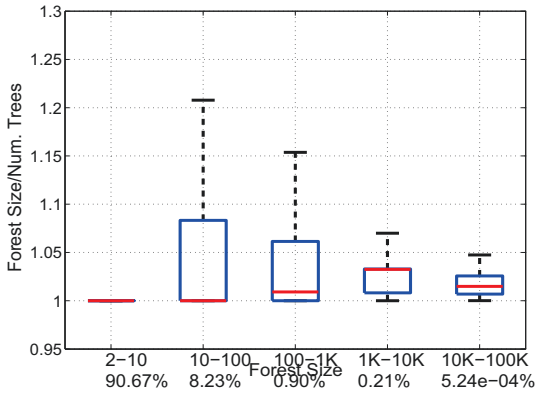
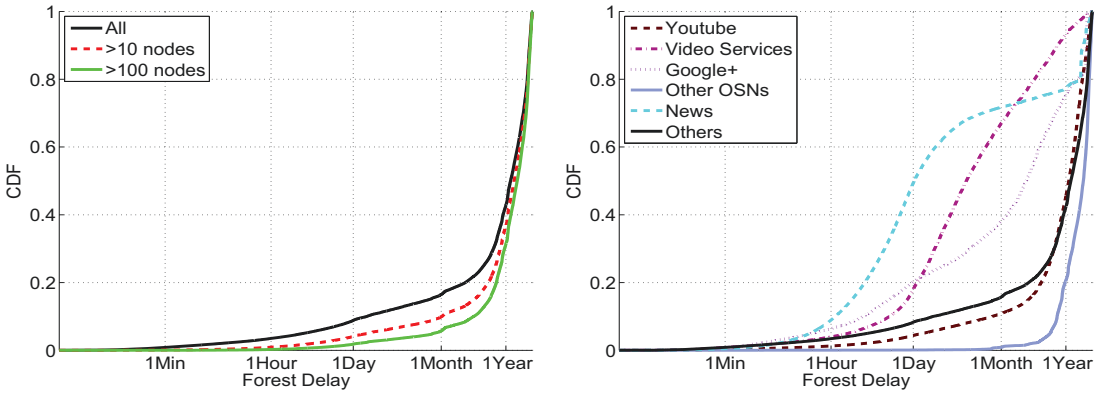


Figure 4.27: Average tree size per forest



(a) Divided by forest size

(b) Divided by content type

Figure 4.28: CDF of the Forest Delay

Figure 4.27 presents the boxplot for the average tree size dividing the population under study in different groups using the size of the tree. The results in any case give us a average tree size among 1 and 1.25. This result confirms our previous hypothesis: the forest are usually composed by a big number of small trees.

4.2.2.2 Temporal Properties of Propagation Forests in G+

As we did in the previous subsection we analyze the temporal properties of the propagation forests in G+. In this case we analyze the Forest Delay, a metric analogue to the Root Delay. This metric is defined as the time elapsed between the instant when an user share or reshare a content and the instant when this content was posted for the first time. This metric captures the overall propagation delay of a post across the entire reshare forest.

Figure 4.28(a) shows the CDF of the forest delay for all the nodes across our dataset. On the contrary of the case of the root delay, where 80% of the reshares where done

during the first 24 hours, for the forest delays we obtain bigger values. We can observe that after 1 month only 20% of the nodes has been published. While in the reshares trees the social graph was very important since for an user is difficult to find the post of other user after some time, this constrain does not apply when the users post external content. Moreover, the content published usually does not have an importance based in the novelty, since half of the nodes has been posted at least one year after the content was published for the first time.

In order to confirm the previous assumption we have divided the post in categories using the domain from where the content comes. Figure 4.28(b) shows the CDF for the forest delay for each one of the defined category. We observe for news webpages like *cnn.com* or *nytimes.com* 50% of the propagation is made during the first day. While this result is far from the observed in the previous section for a single post it indicates this kind of content tend to be consumed in the first hours after it is generated. Finally, we observe two not expected results: first, the lifespan of videos shared in Youtube is much longer than the lifespan of videos shared in other services like *vimeo.com* or *dailymotion.com* and second we observe how the links to other social networks like *Twitter*, *Facebook* or *Linkedin* have a very high lifespan but it does not means they are continuously present in the network, but they are posted after a long time. On the contrary, the links to the other pages inside G+ tend to propagate faster than the average content.

4.2.2.3 Users participation in Resharers forests

The results obtained above indicate there exists some users sharing the same content several times. This subsection analyze the behaviour of the users who have participated in the Forests. For this purpose and following the methodology used in the previous subsection we analyze two variables:

-The *Number of times* a user has participated in the forests indicates the activity level of a given user. In this sense, it is also important to analyze whether the user share the same content more than one time, or if they tend to share content that usually come from the same domain.

-The *User Reach* as defined in subsection 4.2.1.6. In this case we analyze the average reach ($\overline{R}$) and total reach (TR).

How much the users post? Figure 4.29 presents the CDF of the number of times the same user appears in our *G+ forests* dataset, the number of different contents this user has shared in the system and from how many different domains this content come. For this purpose, we use as domain the hostname of the shared URL removing the sub-domains starting with *m.* or *www.* (*i.e.*, m.youtube.com and www.youtube.com count as youtube.com while play.google.com and feedproxy.google.com count as different domains).

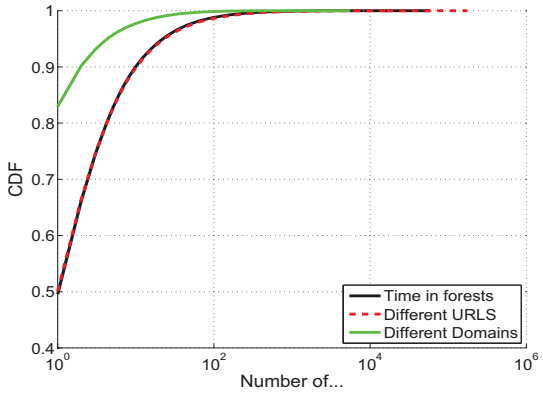


Figure 4.29: Links per user

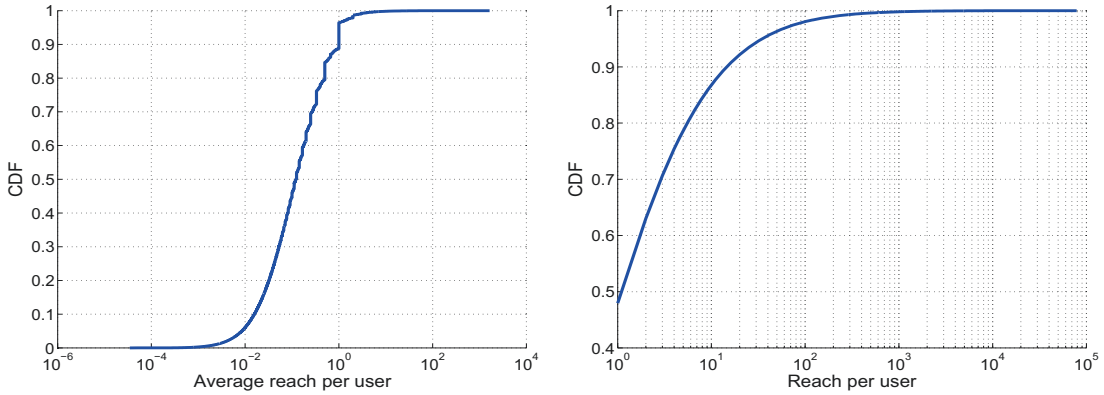
We observe that about 50% of the users appears only one time in our dataset. Moreover, the distribution for the number of times and the number of different contents is very similar. It indicates the users usually share each content only once. To explain why we observed in the previous subsection a huge number of trees where the same user share the same content several times we should focus in the tail of the distribution. A manual inspection of this users allows us to identify some users who seems to be robots/spammers that automatically post a huge amount of links several times, as an example the user `https://plus.google.com/106373251267437926474/` appears more than 8M times in our dataset, but it shares less than 52K different URLs coming from less than 220 different domains.

Moreover, we can observe the number of domains from where the information comes is smaller than the number of different URLs. More than 80% of the users posting only from one domain. While the effect of the aforementioned robots/spammers is important in this metric, it is also worthy to mention that youtube.com is the most popular domain accounting for 37.8% of the total appearances, 18% of the different URLs in our dataset while the second one, ow.ly, only represents 1% of the appearances, 2% of the different URLs and 0.2% of the users.

Capacity of the users to attract resharers Our previous analysis demonstrate the existence of a huge number number of trees with a single node inside the studied forest. It suggest for standard G+ users is very difficult to obtain reshares in their posts.

First, it is important to remark from 37M of users who appears in our forests only 1.3M (3.43%) have obtained at lest one reshare across the reshare forest where they have participated. Figure 4.30 presents the CDF of the $\overline{R}$ and TR for this 1.3M users receiving at least 1 reshare.

The $\overline{R}$ for 88.75% of the users is smaller than 1. It means 88.75% of the users under



(a) Avg. Reach (b) Total Reach

Figure 4.30: CDF of the avg. and total reach per user

study receive less than one reshare per post. Nevertheless, there is a small number of users (5.3K) receiving more than 5 resharers per publication.

Finally, we can observe the value of TR is equal to 1 for about 50% of the users who only manage to obtain one reshare among all their posts. In this case only 2% of the users have attracted more than 100 resharers summing up all their posts.

4.2.2.4 Importance of the social graph in the content propagation

The previous results indicates external aspects are more influential in the content publishing than the relation of an user with its followers since the content is usually originally shared more often than reshared from other user in the network. In this subsection we try to understand if at least the information propagated inside the system follows the underlying social graph. To this end, we have checked if each link created when a user reshare the content of other users represents an existing link in the public social graph or not.

Figure 4.31 presents a boxplot for the percentage of reshares received per user that have been done by the user's followers. The users have been divided in different groups depending on the number of different users from which they have received reactions. Intuitively, we can expect non popular users will have a strong dependency on the social graph in order to disseminate the information posted, nevertheless, we can see how only for 25% of the users attracting reshares from a unique user, this user is her follower. Moreover, when the users receive a higher number of reshare the percentage of their followers resharing them decrease. A manual inspection to the post receiving this *non-follower* resharers suggest the effect of the Google+ communities and the *Hot Topics* as a possible cause for this non expected results.

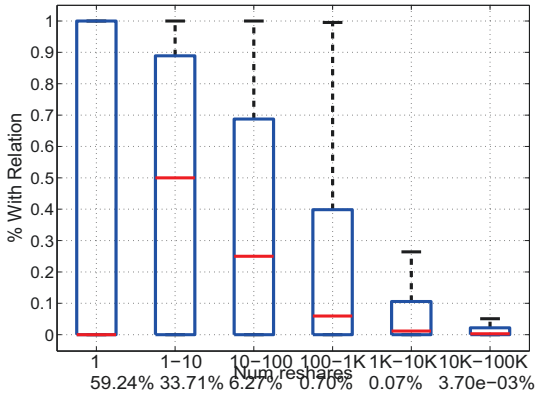
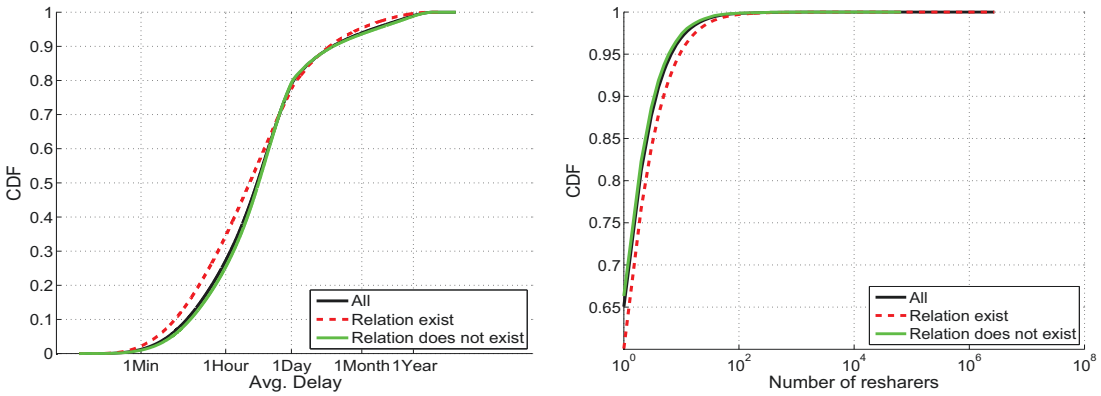


Figure 4.31: % of resharers made by the followers of the user



(a) Average delay in the reshare

(b) Num resharers per relation

Figure 4.32: Effect of the social graph in the propagation of the content

A deeper inspection is needed to understand the role played by the social graph in the propagation of the information. Figure 4.32(a) presents the CDF of the avg. time needed to reshare a post from other user when the relation exist and when the relation does not exist. It is easy to observe the propagation is faster during the first hours when the relation between the users exists. This is a expected result since any user can see her friends activities in her own *wall* just after the publication has been posted. However, when the relation does not exist the information propagates faster during the first days after after the initial hours. This effect can be caused for the time needed for a post to become a hot topic or the time different users need to access the G+ community page.

Next, we analyze the number of times a given user has reshared from other given users. For example, if the user A has reshared 4 different times content published by the user B, we will have a value of 4 for this metric. Figure 4.32(b) shows the CDF of this metric. We can observe the propagation is more common when the social link exists. Nevertheless, the difference observed is smaller than the one we could expect, for example, when the social link exist, 40% of times the users reshare more than one content from the same

users whereas this percentage only decreases to 35% if the social link does not exist.

4.2.2.5 Summary

In this subsection we have characterized the main properties of the external content propagation in G+ and studied the effect of the social graph on it. The main outcomes of our analysis are:

- The content propagation forests in Google+ are composed basically for a big amount of small trees instead of some bigger trees as we could expect in a social network. It suggest external factors as the content popularity have a key importance in order to understand the content propagation inside a social network.

- On the contrary of previous studies and our own previous subsection the life time of a content inside the social network is usually very long.

- The standard users does not usually share the same content more than one time and it is easy to identify a big number of robots/spammer only identifying users who post the same content several times. Nevertheless it is common for standard users to post different URLs belonging to the same domain (*i.e.*, youtube.com).

- The role played by the social graph in the content propagation is important during the first hours after the content has been published. Nevertheless, after this initial phase other factors as the G+ Hot Topics or the communities are more important.

4.2.3 Analysis of the context importance in the user's influence

So far most studies have evaluated the influence of a user considering the reaction to its direct activity (i.e, posts) or based in the user's connectivity properties (e.g., number of followers) [9, 33–40]. However, we believe that influence metrics should take into account the context in order to properly capture the influence. Understanding the complete context around a user is very complex and out of the scope of this thesis. Instead, we would like to present a proof of concept to demonstrate that context matters. In particular, we consider the content popularity as a context condition that affects the influence of a user. For instance, a user A who posts 10 contents that any other user posts and receive 10 reshares would be considered, by traditional metrics, as influent as a user B who post 10 contents published by another 1K users and receiving 10 reshares. We claim that A should be considered more influential than B.

While the definition of the users influence can change we are going to define the influence of an users in the content dissemination as *"the amount of content that wouldn't be disseminated if the given users didn't exist"*. Following this definition the influence of

an user who post a very popular video (posted several times in G+) is smaller than the influence of an user who obtain the same reach with a non popular video (posted only by this user).

Our previous results suggest the popularity of the content is very important on its diffusion over the online social network. Our goal in this subsection is to show how the usage of the content popularity change the concept of the user influence in the network. For this purpose we define 3 different metrics that takes into account the external popularity of the content in order to weight the user's influence. While the results obtained using this metrics are meaningful, we want to remark we do not claim to present the perfect metrics since there are other context variables in addition to the content popularity (*i.e.*, the language used, the time of the day when a given user post or nature of the messages shared) but we present this case to demonstrate the importance of this external variables.

4.2.3.1 Metric Definition

Top	Num. Followers	PageRank	Num. +1s	Num. Replies	Total Reach	metric1	metric2	metric3
1	Lady Gaga	Travel Channel	Rena Matsui (SKE48)	Miyuki Watanabe (NMB48)	YouTube	CNET	CNET	YouTube
2	Britney Spears	Victoria Justice	Miyuki Watanabe (NMB48)	Yamamoto Sayaka (NMB48)	The Verge	The Verge	The Verge	Mike Elgan
3	YouTube	Lisa Bettany	Yamamoto Sayaka (NMB48)	Rena Matsui (SKE48)	Mike Elgan	Susan Stone	Android Police	Google
4	Larry Page	Angry Birds	Yui Yokoyama (AKB48)	Nana Yamada (NMB48)	Vic Gundotra	Android Police	Positive Inspirational Quotes	Vic Gundotra
5	Google Art Project	VEJA	Nana Yamada (NMB48)	Akari Suda (SKE48)	Google	Positive Inspirational Quotes	WIRED	Guy Kawasaki
6	David Beckham	Miniclip	Sashihara Rino (HKT48)	Akari Yoshida (NMB48)	Guy Kawasaki	Rah e azadi	Google Glass	The Verge
7	Snoop Dogg	FriendsEAT	Airi Furukawa (SKE48)	Fuuko Yagura (NMB48)	Robert Scoble	WIRED	Aruna Kojima (AKB48)	Pete Cashmore
8	Trey Ratcliff	Pitbull	Akari Suda (SKE48)	Airi Furukawa (SKE48)	Android Police	Esquire Network	NASA	Robert Scoble
9	Madonna	British Airways	Togasaki Tomnobu (AKB48 manager)	Key Jonishi (NMB48)	Positive Inspirational Quotes	Linda Lawrey	Susan Stone	Mashable
10	Thomas Hawk	Terra Naomi	Akari Yoshida (NMB48)	Yui Yokoyama (AKB48)	Linda Lawrey	glyn moody	Esquire Network	Android Police

Table 4.6: Top10 users using each one of the defined metrics

This subsection present the proposed metrics and the top10 users using these metrics is presented in the table 4.6. In the definition our metrics we use the following variables:

$U_i = \{N_{i1}, N_{i2}, N_{i3}...\}$ = activities posted by the user i
R_{in} = reach of the node in
S_j = Size of the forest j
NT_j = Number of trees in the forest j
$maxTS_j$ = Size of the biggest tree in forest j

- The formula 4.5 define the metric1.

$$Metric1i = \sum_{n=1}^{U_i} \frac{R_{in}}{NT_j} \tag{4.5}$$

This simple metric takes into account the external popularity of the content to weight the TR obtained by the user. In this case we assume the popularity of a content itself (independently of the OSNs propagation) can be estimated by taking into account the number of times a user has posted the content in G+ without resharing it from a different users, this is, the number of trees composing the forest originated for each content. To measure this effect the number of nodes reached for each user is divided by the number of different trees in this forest.

In the top10 of this metric we find popular publications pages as *CNET*, *The Verge* or *WIRED*. This pages usually generate their own content attracting a lot of engagement among their followers. Nevertheless, we also find in the top position of this rank users like *Susan Stone* whose popularity is based on attracting a few resharers in a huge number of activities (More than 50K) instead of a big attention over each ones of this activities. This metric add a very high penalty to the content with external popularity, however, it is able to find meaningful results, and in the top50 we find other well know pages as *NASA*, *Google Glass*, *E! Entertainment* or *TIME*.

- The formula 4.6 define the metric2.

$$Metric2i = \sum_{n=1}^{U_i} \frac{R_{in}}{NT_j} * \frac{R_{in}}{maxTS_j} \tag{4.6}$$

In the second metric presented a second component is added. In this case we divide the reach obtained for each user for the maximum reach in this forest. This second component adds a relative factor of the importance of the user in the distribution of a given content.

The Top10 users obtained using this metric are very similar to the previous one, nevertheless, we observe how the pages that generate their own content are in a best position in this ranking finding in the Top10 users like *NASA*, *Google Glass* or the Japanese singer *Aruna Kojima*.

- The formula 4.7 define the metric3.

$$Metric3i = \sum_{n=1}^{U_i} R_{in} \left(\frac{R_{in} - \frac{S_j}{NT_j}}{R_{in} + \frac{S_j}{NT_j}} \right) \tag{4.7}$$

For the third presented metric we follow a different approach. In this case we weight the Reach obtained for each user with a factor depending in how far from the average is the contribution of the user to the distribution of a given content. In this case this factor

will be negative when the user contribution is smaller than the average and close to one when a user is responsible of a big part of the distributed content.

In the previous metric each activity has a positive influence in the final value allowing users/robots posting a huge amount of activities to reach a good position in the ranking, however, this metric penalize the appearance in trees where the users is not very popular filtering out this users. In the Top10 for this metric we found people and pages well known outside G+. Nevertheless when we focus in the next 15 users we found users like *Wil Wheaton*, *Felicia Day* or *Scott Beale* whose activity make them more popular in G+ than outside the social network.

4.2.3.2 Comparison with traditional metrics

In this subsection we compare the aforementioned metrics with other traditional metrics as the number of followers, the amount of +1's or comments attracted, the pageRank and the total reach.

Since the value provided by our metrics try to establish an order, and the value itself it is probably meaningless we are going to compare the different metrics for the Top10K and Top10 users of G+.

Comparing the top10K To compare the top10K obtained with each metric we are compute the Spearman Rank Correlation among each couple of metrics. Table 4.7 present the value for the rank correlation, Figure 4.33 presents a graphical representation of the same values and the p-values obtained are smaller than 0.002 in all the cases. To calculate this table we use the variable in the left to obtain the top10K users and then we obtain the rank correlation of this variable with each one of the other variables (*i.e.*, the first row present the rank correlation among the number of followers and each one of the other variables for the 10K users with more followers.).

A first inspection shows the new metrics are correlated with the total reach. This effect is reasonable, since obtaining a big total reach indicates a big attention of the community to the content published by the users, however, the value is always smaller than 0.8 and in some cases smaller than 0.5. It indicates important changes in the ranking when we take into account the context in which the content has been published.

If we focus now in the new metrics we observe they are usually correlated among them, specially metric1 and metric2 since they are constructed following the same principles. In a deeper inspection we observe a special situation in the relation of the new defined metrics. For the 10K users with a higher value for the metric3, the value of this metric is correlated with the metric1 and the metric2 (close to 0.6). Nevertheless, when we look

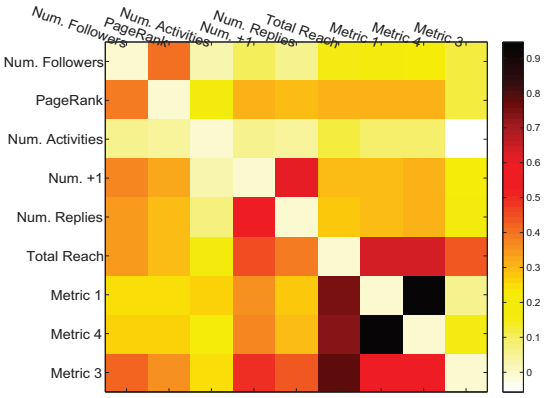


Figure 4.33: Rank Correlation among the top10000 users of different metrics. The top10000 has been calculated using the variable in the left (Every row use the same population).

at the relation the other way around, obtaining the studied population using metric1 and metric2 we can observe how the rank correlation is close to zero. This change is produced because the metric3 filter out the users who obtain a small reach in a big number of trees. Thus, users with a high value in metric1 or metric2 can have a small or even negative for metric3 while the users with a high value in metric3 will usually have a high value for the metric1 and metric2.

When we compare the new defined metrics with the traditional ones we observe they are usually not correlated. It indicates adding the content popularity gives us a different view of the user's influence.

	num followers	pagerank	num activities	num plusones	num replies	total reach	metric1	metric2	metric3
num followers	-	0.41192	0.031042	0.10085	0.068163	0.1587	0.19576	0.2119	0.11978
pagerank	0.39232	-	0.17515	0.31567	0.30351	0.31375	0.31244	0.31607	0.11694
num activities	0.058369	0.043055	-	0.068131	0.044423	0.11737	0.098366	0.097685	-0.056631
num plusones	0.37522	0.33227	0.036823	-	0.6075	0.30047	0.30101	0.31377	0.22187
num replies	0.34809	0.30108	0.077974	0.57303	-	0.28713	0.29308	0.30403	0.20201
total reach	0.34794	0.30283	0.19086	0.45969	0.387	-	0.63768	0.64247	0.43081
metric1	0.24821	0.24442	0.27174	0.35422	0.28208	0.74378	-	0.94553	0.068089
metric2	0.26677	0.25828	0.22336	0.37024	0.29784	0.73853	0.94391	-	0.15406
metric3	0.41504	0.36144	0.24729	0.4951	0.43976	0.77726	0.57235	0.5787	-

Table 4.7: Rank Correlation among the top10000 users of different metrics. The top10000 has been calculated using the variable in the left (Every row use the same population).

Comparing the top10 The different metrics analyzed above provide different results, however every one of these metrics is able to identify popular and active users in the social network. Table 4.6 presents the top10 users using the classic metrics and the new ones. The users with more followers or a bigger PageRank are usually well know outside G+ but not necessarily very active or influential in the network. Moreover this users are different very different to the users obtained taking into account the new metrics. For instance, *Lady Gaga*, the user with more followers in G+ when we collected the data is

not among the Top400 users for any of the metrics that takes into account the content popularity. It indicates this metrics do not necessarily represent the influence of an user in the network but the popularity of the users outside G+.

When we focus on the top10 using the number of comments or +1's we surprisingly find the top position of the top are filled by Asiatic singers of groups like AKB48, NMB48 or SKE48. These singers fill 23 of the top25 positions if we rank them using the number of +1's and 97 of the top100 when we use the number of comments on their activities. Again, this ranking is very different than the ranking found with the new metrics. To explain this difference we should focus on the kind of content published by this users. In this case the Asiatic singers tend to share content inside the social network as their own photos or text, but not external content.

Finally, if we compare the ranking obtained with the new defined metrics and the total reach we observe metric3 provide similar results to the TotalReach while metric1 and metric2 provide similar results among them. Only *The Verge* and *Android Police* appears in the Top10 for the fore ranking. These pages share a huge amount of external content, sometimes own generated, and usually obtain a good number of resharers in each one of them.

It is worthy to remark the important presence of Google products and workers (as the former vice-president Vic Gundotra) among the top users of G+ following each one of the ranking. It indicates the users of G+ tend to be also interested in other Google products.

4.2.3.3 Summary

In this subsection we present three new metrics that can be used to identify the most influential users of a social network taking into account the content popularity. The main outcomes of this subsections are:

- The metrics defined, while they do not pretend to be perfect, present meaningful results by taking into account the effect of the content popularity.

- The ranking obtained for the traditional metrics like the number of followers represents the popularity of the users outside G+ but not the influence of the users in the network. Moreover, the rank correlation observed for the new metrics is usually very high.

- There are a big number of Google related users among the most influential ones. It indicates the users of Google+ are very interested in other Google products.

4.3 Analysis of the arrival process of messages to twitter

4.3.1 Data analysis

Now, consider the hypothesis that the tweet arrival times at Twitter for the i-th hour follows a Poisson process with rate λ_i. This assumption comes from the well-known Palm-Khintchine that states that *the aggregation of multiple independent low-rate counting processes converges to a Poisson process*. Here, each user is considered an independent low-rate counting processes. In like of this, the number of aggregated tweet arrivals $k = 0, 1, \ldots$ within a given time window $[0, t]$ follows:

$$P(N(t) = k) = \frac{(\lambda_i t)^k}{k!} e^{-\lambda_i t}, \quad k = 0, 1, 2, \ldots$$

Now, consider Z_{10} (i.e. the tweet arrival times at 10 a.m.), a first estimate of the value of $\hat{\lambda}_{10}$ is:

$$\hat{\lambda}_{10} = \frac{N_i}{T_i} = \frac{31964}{12114} = 2.64 \text{ tweets/ms}$$

that is, the total number of tweet arrivals divided by the arrival time of the last one (see Table 3.1).

To check the Poisson assumption, let $X_{10}^{(T_{\text{samp}})}$ refer to the number of tweet arrivals within a sampling window of T_{samp} units of time. To obtain $X_{10}^{(T_{\text{samp}})}$, we generate a vector of, for instance 100 random sample times t_j, $j = 1, \ldots, 100$ and then we count the number of tweet arrivals within the time interval $(t_j, t_j + T_{\text{samp}})$. For example, for $T_{\text{samp}} = 1$ms such vector would be:

$$X_{10}^{(1\text{ms})} = [3, 4, 3, 3, 4, 3, 2, 3, 2, 3, \ldots]$$

Next, our hypothesis is that $X_i^{(T_{\text{samp}})} \sim Poiss(\lambda_i T_{\text{samp}})$. Fig. 4.34 shows the histogram and the Poisson fit for $X_{10}^{(T_{\text{samp}})}$ for different values of T_{samp}: 1ms, 5ms, 10ms and 25ms. Visually, the Poisson fit is very accurate for values of T_{samp} below 10ms.

It is also worth remarking that the Poisson distribution with parameter λT approaches the Gaussian distribution with λT mean and variance, for large values of λT, as it follows from the Central Limit Theorem (CLT). Hence $X_i^{(T_{sample})} \sim N(\lambda_i T_{sample}, \lambda_i T_{sample})$ as well, when λT_{samp} is large. Fig. 4.35(a) shows the normalised histogram for $X_{10}^{(5\text{ms})}$ with the Poisson fit. Fig. 4.35(b) shows the CDF of $X_{10}^{(5\text{ms})}$ together with the Poisson and Gaussian CDFs with estimated parameters. As shown, the two CDFs are visually close, which allows to assume that the Gaussian PDF is also suitable for modelling the tweet arrival process. Finally, Fig. 4.35(c) shows the QQ-plot for $X_{10}^{(5\text{ms})}$ with the Gaussian fit. As shown, the Gaussian fit is accurate near the mean, but not so accurate on the Gaussian tails. This will have an impact on the normality tests obtained in the next section.

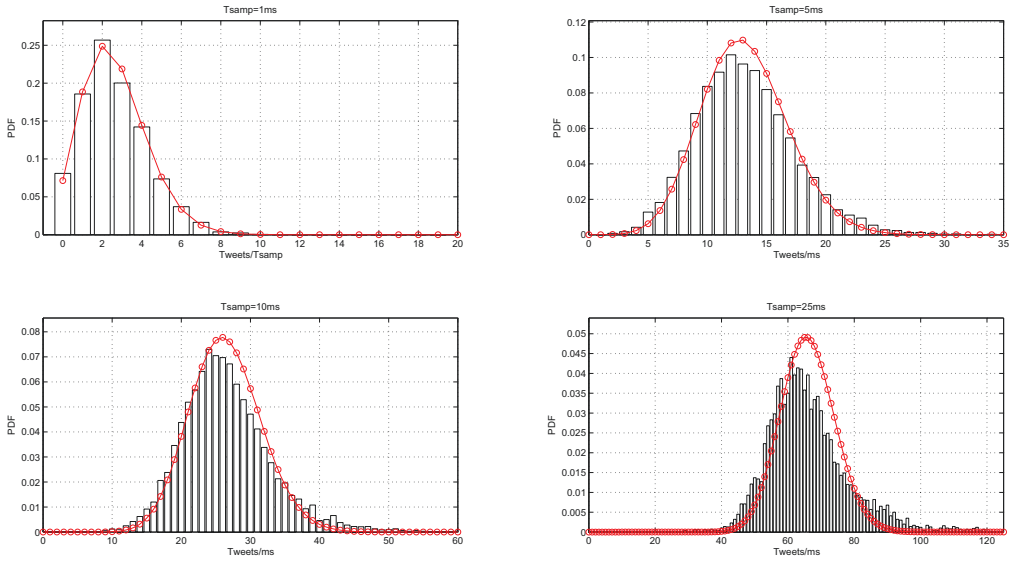


Figure 4.34: Histogram and Poisson fit for different T_{samp} values: 1ms Top-Left; 5ms Top-Right; 10ms Bottom-Left; 25ms Bottom-Right

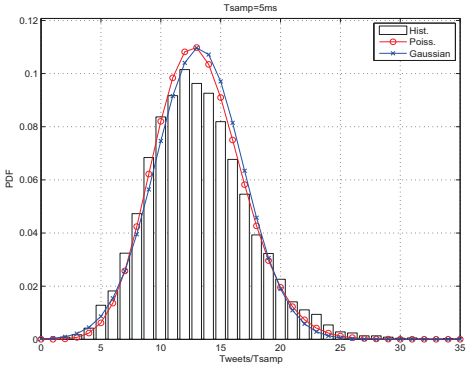
Next section provides a more quantitative approach to assessing on the Normality of $X_i^{(T_{\text{samp}})}$ for $i = 0, \ldots, 47$, using well-known normality tests.

4.3.1.1 Goodness-of-fit tests applied to X_i

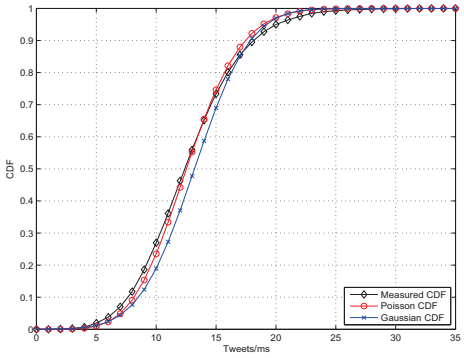
The literature offers a large number of normality goodness-of-fit tests to check whether or not a sample vector of measurements collected from some random variable X can be considered normally distributed, with some degree of confidence α (typically $\alpha = 0.05$). We have applied the most popular normality tests to the measurement set of tweet arrivals at different times of the day. These tests are: Chi-square, Kolmogorov-Smirnov, Lilliefors, Anderson-Darling, Shapiro-Wilks, D'Agostino-Pearson and Jarque-Bera.

The results of Table 4.8 give the percentage of rejected normality tests at each hour. Essentially, for every hour Z_i, we have generated 1000 random vectors $X_i^{(5\text{ms})}$ and applied the seven normality tests, with result $H = 0$ (pass) and $H = 1$ (fail). Thus, the values in the table show the percentage of times that the tests were rejected. As shown the reject ratios highly depend on the tests and the hours. Additionally, the Chi2 and KS tests, which give less importance to the tail fit of the Gaussian distribution show a larger percentage of passed tests than the other tests which focus on the tail fit.

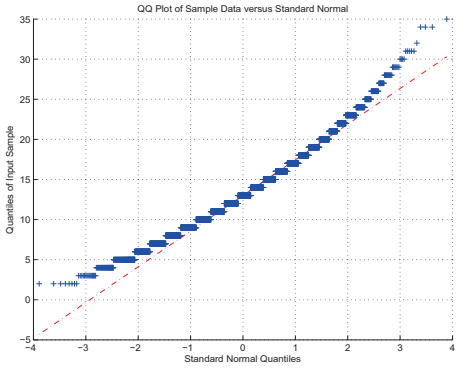
Finally, Table 4.9 shows the average number of rejected tests (aggregated for all hours) for different values of T_{samp}: 1ms, 2ms, 5ms, 8ms, 10ms, 15ms, 20ms, 25ms, 50ms, 75ms



(a) Histogram



(b) CDF



(c) QQ-plot

Figure 4.35: Visual assessment of normality: (a) Histogram, (b) CDF and (c) QQ-plot

and 100ms. As shown, normality is more accurate for T_{samp} values between 5ms and 10ms.

4.3.1.2 Multivariate Gaussian distribution

Given the above results, we cannot say that the tweet arrival process can be characterised by a Gaussian process, but we still may use a Gaussian process as a reasonable approximation of it. To do so, we can use the multivariate Gaussian distribution for approximating the tweet arrival process at every hour:

$$f_{\mathbf{X}}(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\right] \tag{4.8}$$

Here, $\mathbf{X} = \{X_0, \ldots, X_{23m}\}$ is a multivariate ($k = 48$ dimensions) random variable, whose items refer to the tweet arrivals per T_{samp}. Such a vector $\mathbf{X}$ is characterised by its vector mean μ and covariance matrix Σ, and $k = 48$ refers to the size of vector $\mathbf{X}$.

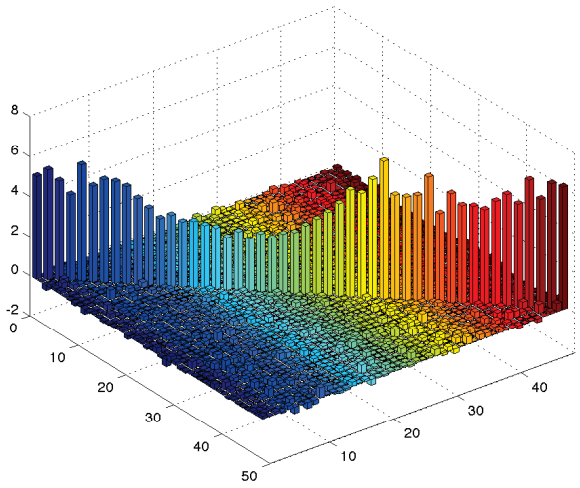


Figure 4.36: Covariance matrix

Fig. 4.36 shows the 48×48 covariance matrix for vector $\mathbf{X}$ estimated from the measurement set. As shown, there is little covariance between the different hours since $Cov(X_i, X_j)$, $i \neq j$ approach zero, thus we may assume independence of the tweet arrival process on each hour. Hence, every hour may be modelled with a univariate Gaussian distribution characterised by its mean μ_i and variance σ_i^2.

4.3.2 Applications

The Gaussian approximation proposed in this work may be applied to the following further research lines:

4.3.2.1 Dimensioning and upgrading

As shown, the daily traffic pattern reveals an average tweet arrival rate between 2.5 to 5 tweet/ms, for a number of about 200 million users total. This comprises an average of:

$$\frac{5 \text{ tweets/ms}}{200 \text{ million users}} = 2.16 \text{ tweets/(user } \cdot \text{ day)}$$

that is, every user generates about 2 tweets per day for the peak hour. This rule of thumb can be used to plan a network upgrade as the social network grows (in terms of number of users). It is also worth remarking that a posted tweet may translate to a large number of copies to his/her followers, depending on the user's popularity (number of followers for that user).

4.3.2.2 Detection of outlier events

Thanks to the Gaussian approximation, we may identify outliers by monitoring and estimating the tweet arrival rate in real-time. For example, consider that we observe 10 tweet arrivals in a ms-timeslot at 1 a.m., while the average number of tweet arrivals for that time should be: 4.391 tweets/ms (see Table 3.1). Hence, given the Gaussian model at 1 a.m. characterised by mean $\mu = \lambda T$ and $\sigma = \sqrt{\lambda T}$, timeslots with 10 or more tweet arrivals occur with probability:

$$\begin{aligned} P(X_1 > 10) &= \int_{x=10}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= \int_{\frac{10-\mu}{\sigma\sqrt{2}}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-z^2} dz \end{aligned}$$

after changing variable: $z = \frac{x-\mu}{\sigma\sqrt{2}}$.

Hence, this gives:

$$P(X_1 > 10) = \frac{1}{2}\text{erfc}\left(\frac{10 - 4.391}{\sqrt{4.391}\sqrt{2}}\right) = 3.7 \cdot 10^{-3}$$

The same number of tweet arrivals at time 10 a.m. is even more unlikely since the Gaussian model for 10 a.m. is characterised by: $N(2.64, \sqrt{2.64})$. Hence:

$$P(X_{10} > 10) = \frac{1}{2}\text{erfc}\left(\frac{10 - 2.64}{\sqrt{2.64}\sqrt{2}}\right) = 2.95 \cdot 10^{-6}$$

Thus, the unlikelihood of such a case can be used to trigger an alarm to the network administrator to see whether this belongs to an important social event (e.g. Michael Jackson's death) or a Denial-of-Service attack (DoS).

Similarly, the same analysis can be performed to identify extremely low tweet arrival rates, which may suggest a failure in the system operation or network connectivity performance.

4.3.2.3 Data Migration and Replication

The Twitter Infrastructure comprises several datacenters located at both US East and West Coasts, a usual practice followed by other social networks (e.g. Facebook). Thus, as suggested in previous studies, the daily traffic pattern analysis may be used to schedule delay tolerant replication and migration processes during low-loaded times of the day (in our case, 10 a.m. for instance). Such a strategy may lead to a better use of the network and datacenter infrastructure.

4.3.2.4 Energy-efficiency

The traffic pattern exhibited by the Twitter load may also be used to define energy-efficiency strategies such as powering off equipment (servers, switches, etc) at the off-peak hours, and vice versa. Important energy savings and carbon footprint reduction may be achieved by a dynamic power on/off strategy.

4.3.3 Discussion and further work

In a nutshell, the main contributions of this section are:

- A detailed measurement methodology to enable researchers plan their measurement experiments, highlighting the Twitter API limits.
- The statistical analysis of the collected data demonstrates that the aggregated tweet arrival process can be approximated (but not accurately modelled) by a Gaussian process, which may further permit statistical inference and forecasting.
- Further applications of the above two findings are highlighted, and may comprise interesting furure research lines.

Finally, further work shall study the daily traffic pattern with measurements collected from multiple days to evaluate whether or not the traffic patterns repeat themselves in consecutive days.

Hour/Test	Chi2	KS	Li	AD	SW	DP	JB
00.00	20.20%	10.70%	72.00%	61.90%	61.50%	62.00%	62.40%
00.30	12.20%	3.60%	53.40%	37.90%	40.10%	41.20%	42.80%
01.00	20.40%	11.10%	72.00%	69.00%	70.80%	74.40%	75.30%
01.30	19.80%	9.80%	73.20%	64.50%	59.70%	61.80%	62.50%
02.00	16.80%	6.80%	66.10%	63.60%	57.80%	61.50%	62.50%
02.30	11.90%	2.50%	52.80%	37.80%	36.20%	40.20%	41.70%
03.00	23.90%	14.70%	78.80%	79.40%	78.20%	80.60%	81.10%
03.30	12.20%	5.00%	58.10%	50.80%	52.70%	58.20%	59.30%
04.00	15.00%	7.70%	64.00%	60.90%	63.80%	64.80%	65.60%
04.30	23.90%	11.20%	76.50%	67.60%	61.70%	60.10%	61.70%
05.00	17.30%	7.20%	70.30%	60.30%	58.20%	59.60%	60.60%
05.30	20.30%	6.50%	67.80%	52.30%	45.30%	43.20%	44.20%
06.00	16.90%	4.10%	63.00%	44.20%	35.90%	30.80%	31.90%
06.30	16.90%	6.70%	68.70%	49.10%	42.20%	37.00%	37.80%
07.00	20.30%	6.80%	70.10%	52.80%	39.00%	34.50%	35.20%
07.30	23.10%	11.10%	78.30%	63.50%	49.80%	44.90%	45.80%
08.00	22.50%	12.60%	79.80%	61.70%	42.50%	26.80%	28.70%
08.30	22.70%	6.10%	74.60%	57.10%	45.10%	29.80%	31.10%
09.00	20.40%	5.20%	67.30%	55.60%	43.90%	34.00%	36.00%
09.30	19.10%	8.30%	75.00%	54.10%	38.30%	31.70%	32.80%
10.00	19.60%	9.50%	76.60%	60.00%	48.70%	44.00%	44.80%
10.30	21.30%	12.20%	79.40%	63.50%	50.70%	41.10%	42.20%
11.00	18.50%	6.40%	66.20%	44.50%	28.50%	24.00%	25.20%
11.30	20.90%	5.10%	66.10%	45.00%	30.00%	20.10%	21.00%
12.00	17.30%	5.50%	64.70%	47.90%	34.80%	30.90%	32.10%
12.30	18.60%	3.00%	61.40%	38.20%	25.60%	19.30%	20.40%
13.00	16.40%	6.80%	66.00%	57.10%	57.30%	58.10%	59.70%
13.30	26.00%	16.30%	79.70%	77.00%	75.70%	78.20%	78.70%
14.00	11.00%	6.10%	59.70%	51.20%	53.70%	58.90%	59.20%
14.30	8.40%	0.80%	33.70%	23.00%	14.40%	12.50%	13.90%
15.00	10.40%	2.30%	48.70%	40.40%	37.90%	40.80%	43.00%
15.30	7.60%	1.20%	34.40%	24.30%	17.60%	19.00%	20.00%
16.00	6.30%	0.90%	36.60%	23.90%	17.50%	18.60%	20.70%
16.30	5.30%	0.30%	21.70%	10.60%	6.50%	5.90%	6.40%
17.00	10.10%	1.30%	44.70%	30.40%	23.10%	25.40%	27.10%
17.30	13.70%	2.50%	45.70%	32.30%	30.00%	32.60%	34.00%
18.00	12.80%	5.10%	56.40%	49.40%	49.00%	53.70%	54.80%
18.30	6.80%	1.00%	35.30%	25.20%	28.40%	32.50%	33.30%
19.00	9.40%	3.40%	51.00%	37.70%	36.70%	40.30%	42.10%
19.30	9.00%	1.90%	44.10%	33.90%	29.30%	31.40%	32.50%
20.00	10.30%	2.90%	48.20%	34.80%	33.60%	38.00%	39.60%
20.30	11.20%	2.90%	46.50%	36.80%	34.20%	38.50%	39.70%
21.00	6.20%	0.90%	32.60%	24.30%	19.20%	22.00%	23.20%
21.30	8.30%	0.70%	35.00%	22.30%	15.60%	18.30%	19.60%
22.00	5.40%	0.50%	25.30%	12.70%	7.10%	8.20%	8.60%
22.30	6.80%	0.30%	28.40%	19.30%	12.70%	14.00%	14.80%
23.00	8.70%	0.20%	30.40%	23.70%	18.40%	4.80%	6.00%
23.30	9.90%	2.20%	47.20%	32.20%	30.70%	32.80%	35.00%

Table 4.8: Univariate normality tests $T_{\text{samp}} = 5$ms. Percentage of rejected tests.

T_{samp}	Chi2	KS	Li	AD	SW	DP	JB
1ms	52.02%	72.63%	99.97%	100.00%	98.93%	53.86%	56.43%
2ms	42.67%	22.60%	91.28%	83.68%	64.19%	41.11%	42.56%
5ms	14.83%	5.41%	57.24%	45.12%	39.37%	38.35%	39.51%
8ms	14.03%	4.66%	50.93%	46.90%	43.46%	43.58%	45.31%
10ms	15.62%	5.39%	51.41%	51.17%	47.30%	46.59%	48.89%
15ms	23.51%	8.52%	56.83%	62.44%	57.77%	53.19%	55.88%
20ms	32.76%	12.28%	62.99%	70.35%	65.05%	57.04%	59.75%
25ms	41.26%	16.80%	67.76%	75.53%	70.42%	59.62%	62.08%
50ms	62.80%	35.50%	78.69%	85.77%	82.22%	63.38%	65.67%
75ms	71.88%	47.01%	83.36%	89.99%	87.52%	64.08%	66.69%
100ms	75.69%	53.45%	84.94%	91.03%	89.53%	64.20%	67.02%

Table 4.9: Goodness-of-fit tests applied for different T_{samp} values

4.4 Analisys of the locality effect in Twitter

In this section we analyze the locality effect in Twitter, for this purpose we first analyze if the Twitter users tend to *tweet* from a unique place or if they use the system from different regions. Then we analyze if the followers are usually geographically close to the users they follow and finally we study if the information travels or remain in the place where it is generated.

4.4.1 Twitter Users' Locality

Our goal in this section is characterizing the locality properties associated to the activity of Twitter users. For this purpose we define the concept of *coverage area*. We define the coverage area as the geographical location (or set of locations) from where a Twitter user performs her activity. The activity of a Twitter user is divided into two major tasks: posting (producing) and reading (consuming) tweets. Although the coverage area from where these two tasks are performed may not be perfectly correlated at a low granularity level (*e.g.*, specific address from where both activities are performed), it is reasonable to think that the location of both types of activity is highly correlated when we consider larger geographical areas such as a city or a country. Therefore, we assume that the set of geographical locations (*e.g.*, city, country) from where a user either post or read tweets accurately defines the coverage area of this user.

To the best of our knowledge, there is any proposed technique that allows to retrieve the location from where a large number of Twitter users consume tweets. However, the methodology described in Section 3.2.4 enables us to collect the location from where hundreds of thousands users post their tweets. Therefore, in this section we use our *Users Dataset* to characterize the coverage area of Twitter users. For this purpose we first explore the geographical distance between the location tag provided by a user and the geolocation coordinates associated to her tweets. Second, we map the GPS coordinates of a user's tweets to different GCP communities (*i.e.*, country, region/state and city). Finally, we analyze the fraction of tweets that a user posts from the different locations that form the user's coverage area.

4.4.1.1 Geographical Distance of the Coverage Area

For each user in our *Users Dataset* we consider the location tag as the user's reference location. We compute the distance between the location specified in the location tag and the location defined by the GPS coordinates for each one of the user's tweets. Figure 4.37 presents the CDF of the median distance between the location tag and the location of the different tweets for a user. The result shows that a large fraction of users ($> 70\%$)

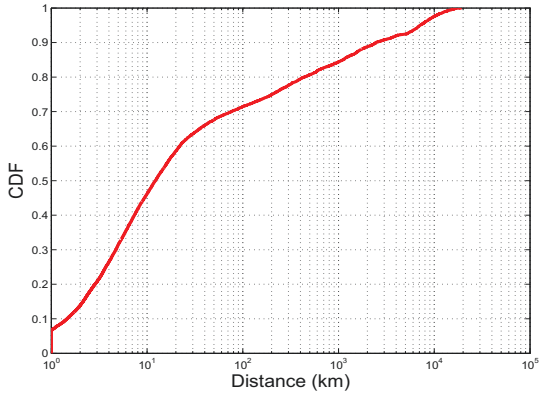


Figure 4.37: Median distance between the user's location tag and the user's tweets GPS coordinates

typically post their tweets in a range of less than 100Km from the location indicated in their profile. This suggests that: (*i*) The location tag can be safely used as an accurate location for a major portion of Twitter users. (*ii*) A major fraction of Twitter users shows a coverage area in the order of few hundred Kms.

4.4.1.2 Geopolitical Composition of the Coverage Area

In this subsection, for each users within our *User Dataset* we map the GPS coordinates of all her tweets to different GCP communities with different granularity, namely countries, regions[12] and cities.

Figure 4.38 shows the distribution of number of cities, regions and countries from where users within our *Users Dataset* send tweets. Note that the box represents the 25, 50 and 75 percentiles and the two external bars represent the 5 and 95 percentiles for the considered metric.

The obtained results show that the coverage area of Twitter users is formed by 3 cities in median whereas just 25% of users send tweets from more than 5 cities. Furthermore, if we consider carefully the other two more coarse metrics, we observe that 75% of users send their tweets from just one or two regions and a single country.

4.4.1.3 Distribution of User's Activity across different locations.

In the previous subsections we have analyzed the coverage area of Twitter users. Specifically, we have analyzed its size and the number of countries, regions and cities included in each user's coverage area. However, in order to fully characterize the locality

[12]We define a region as an GCP community smaller than a country and larger than a city. For instance states in US or Germany or administrative regions in France.

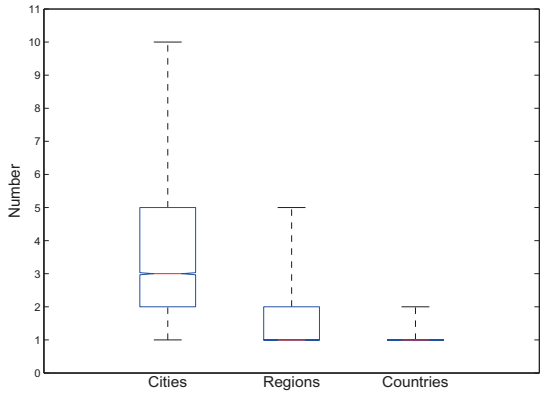


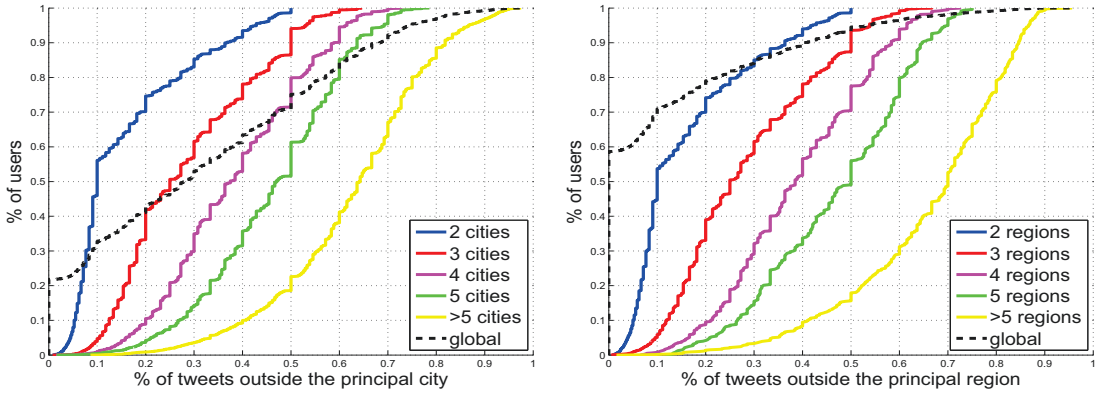
Figure 4.38: Distribution of number of cities, regions and countries from where users send tweets

associated to users' activity it does not suffice with knowing from how many locations (*e.g.*, cities) they perform their activity (*i.e.*, post tweets), rather we need to analyze what is the fraction of the activity performed from each location. We address this issue in this subsection.

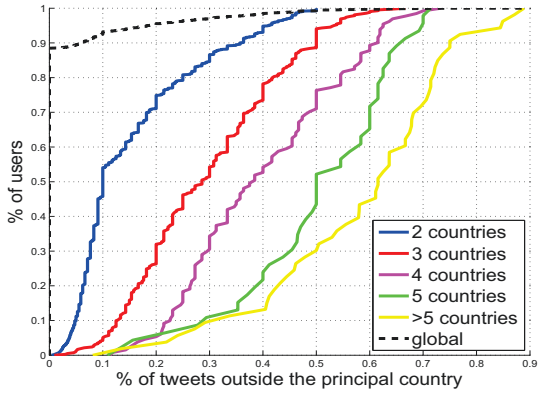
Figure 4.39 shows the cumulative percentage of users (y axis) that send at least x% of their tweets (x axis) from outside their main location using three types of GCP communities with different granularity: city, region and country. We group users by the number of associated locations (n) in the following groups: 2 locations, 3 locations, 4 locations, 5 locations, more than 5 locations and *global* that includes all users in our dataset. Note that the group of users with an unique location (n = 1) is not included in the figure since they send all their tweets from that single location.

Let us focus first on the *global* group that includes all the users. The results show that 90% of users send all their tweets from a single country. This percentage shrinks to 60% and 20% for regions and cities respectively.

If we now consider the other groups the results reveal two important observations: (*i*) the main location is significantly more used by the user than the other ones. For instance, 50% of users post at least 50% of their tweets from the main location for all groups (excepting for n > 5) and all types of locations (city, region or country); (*ii*) In general, the users do not tweet from sporadic locations, rather they tweet from locations that they visit frequently. We refer to a sporadic location as that one that the user visit just one (or few times) and from where she posts just few tweets (*e.g.*, during a business trip). Note that if these sporadic locations were common, their presence would influence more to those groups having larger values of n. Then, the separation between the curves should become significantly smaller as we increase the number of locations.



(a) Cities (b) Regions



(c) Countries

Figure 4.39: Percentage of users vs. percentage of tweets sent from a different location than the principal one (city, region and country)

4.4.1.4 Summary

The obtained results suggest that around 3/4 of Twitter users perform their activity from a relatively reduced coverage area within a country that covers few hundred Km including few (≤ 5) cities and an even smaller (≤ 2) number of regions . Hence, these results reveal that the area from where most Twitter users perform their activity is highly localized. In addition, our study on the activity distribution across users' locations reveals that there is typically a predominant location (city, region or country) from where the user post a significant portion of her tweets. At the same time, users seem to rarely post tweets from "sporadic" locations. Finally, our analysis reveal that the user's location tag accurately define the location of a user (at least at the country level).

4.4.2 Twitter Relationships' locality

In this section we study the geographical properties associated to Twitter relationships, *i.e.*, *friend→follower* links. For this purpose, we rely on our *Relationship Dataset*

Country	Language	Friends (num / %)	Followers (num / %)	Originated Friend→Follower Links (num / %)	Received Friend→Follower Links (num / %)
US	EN	528K / 54.24%	7.37M / 44.59%	60.1M / 59.82%	57.1M 56.84%
UK	EN	70.6K / 7.27%	987K / 1.41%	7.18M / 7.15%	6.94M 6.90%
BR	PO	61.7K / 6.34%	1.81M / 10.94%	6.46M / 6.42%	6.74M 6.70%
CA	EN/FR	39.4K / 4.05%	565K / 3.42%	4.74M / 4.72%	4.55M 4.53%
AU	EN	20.3K / 2.09%	232K / 1.40%	2.50M / 2.48%	2.40M 2.38%
DE	DE	21.7K / 2.23%	331K / 2.00%	2.02M / 2.01%	2.26M 2.25%
IN	IN/EN	18.8K / 1.93%	442K / 2.67%	1.28M / 1.28%	1.52M 1.51%
NL	NL	14.9K / 1.53%	334K / 2.02%	1.22M / 1.22%	1.26M 1.25%
ES	SP	8.7K / 0.89%	277K / 1.68%	0.90M / 0.89%	904K 0.90%
FR	FR	10.8K / 1.11%	232K / 1.41%	0.82M / 0.82%	840K 0.84%
ID	ID	12.1K / 1.24%	862K / 5.22%	0.64M / 0.64%	1.09M 1.09%
MX	SP	5.5K / 0.56%	234K / 1.41%	0.55M / 0.55%	657K 0.65%
IT	IT	7.1K / 0.73%	159K / 0.96%	0.49M / 0.48%	637K 0.63%
JP	JP	6.9K / 0.71%	192K / 1.16%	0.48M / 0.48%	597K 0.59%
TOP 14	-	827K / 85.00%	13.37M / 80.31%	89.9M / 88.95%	88.06M 87.08%
ALL	-	973K / 100%	16.53M / 100%	100.5M / 100%	100.5M 100%

Table 4.10: Contribution of the Top 14 countries to the *Relationships Dataset*, sorted by the number of originated Friend→Follower Links

that includes more than 100M relationships in which both friend and follower have a location tag.

In order to perform the analysis, we group the friends in our dataset by country. We have selected the country criteria since it perfectly matches the concept of GPC community, *i.e.*, friends having a close geographical location, a similar cultural profile and the same language. Furthermore, as we have demonstrated in the previous section, we can map a user to a country with a very low error probability[13].

We first study the demographic composition of our dataset. Then we characterize the geographical properties of the Twitter relationships by carefully studying the fraction of intra and inter *friend→follower* relationships for the most relevant countries in our dataset.

4.4.2.1 Twitter demographics

Table 4.10 shows the number of friends, the number of followers and the number of originated and received *friend→follower* links for the Top 14 countries in our dataset, that are those that contribute more than 100K users. Note that overall these 14 countries are responsible for around 85% of all the friends, followers and relationships within our *Relationships Dataset*. Furthermore, US is clearly a predominant country in Twitter responsible for around half of the friends, followers and links. Among the other countries we observe two clear profiles from a language perspective. On the one hand, we have those countries whose official (or co-official) language is the English such as US, Canada, UK, India and Australia. On the other hand, we find those countries with a different official language than English such as Brazil, Spain, Germany, France, Italy, Indonesia, Japan

[13]We could perform the same analysis using GPC communities at different granularities (*e.g.*, regions or cities). However, as we will see our analysis based on countries reveals important insights, then we leave the analysis with other GPC communities for future work.

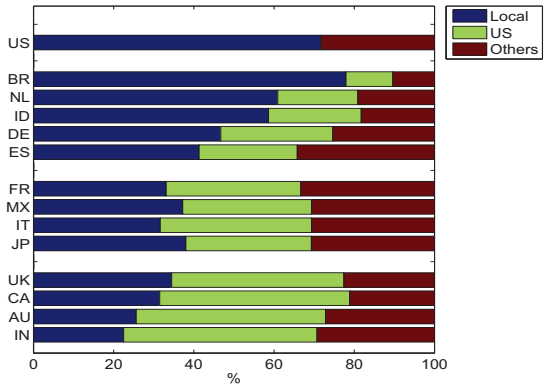


Figure 4.40: Percentage of *friend→follower* relationships originated in each one of the Top 14 countries that remain local, go to US or go to another country different than US

and The Netherlands. Finally, it is worth to note the presence of developing countries such as Brazil, India and Mexico in the list. This is mainly due to the big population of these countries that enables to contribute a large number of users but also indicates the interest of their population on new social ways of communication such as Twitter.

Once we have analyzed the basic demographics of our dataset, in the rest of the section we focus on analyzing the fraction of intra- and inter-country relationships for each one of the Top 14 countries. For this purpose we rely on both the GPC community information (*i.e.*, user's country) and the geographical distance of the *friend→follower* links.

4.4.2.2 Geopolitical Analysis

For each *friend→follower* link within our *Relationships* dataset we identify the country of the friend and the follower involved in the relationship. This allows us to study the destination of all the relationships originated in a given country. In particular, we perform a twofold analysis. First, we study the aggregate percentage of relationships generated in a country that go to different destinations. We refer to this analysis as *link-level* analysis. However, the behaviour of unpopular users might not be well captured in such analysis since those popular users are the ones responsible for a larger portion of the relationships generated in a country. Therefore, in the second part of our analysis we study the percentage of links associated to each individual user that go to different destinations. We refer to this analysis as *user-level* analysis.

4.4.2.3 Link-level Analysis

For each one of the Top 14 countries, we compute the percentage of *friend→follower* links originated in the country that: (*i*) remain within the country, (*ii*) go to US (predominant country) and, (*iii*) go to a different country than US. Figure 4.40 depicts the

obtained results that show the presence of significantly different behaviors across the studied countries. Specifically, we can distinguish the next four different profiles:

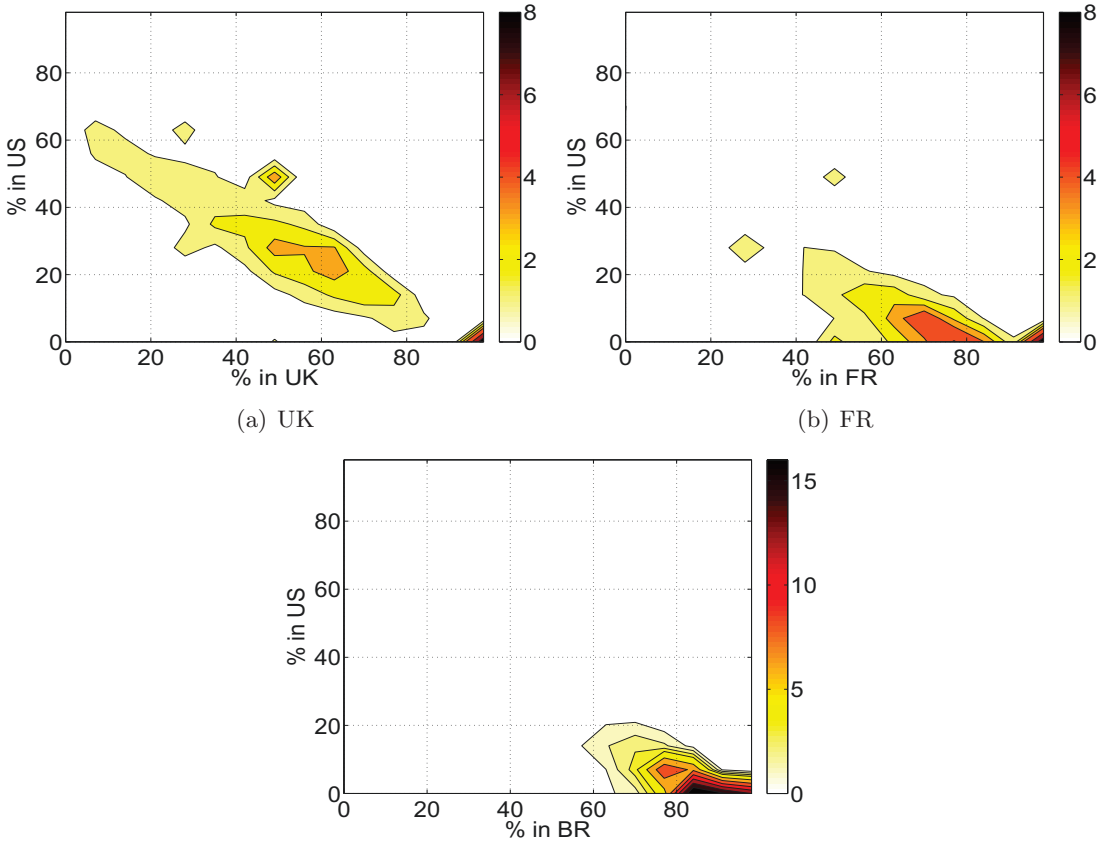
US: due to its predominant role, it has to be considered as a separated profile. It keeps more than 70% *friend→follower* relationships local. This is consequence of first, the predominance of US users in Twitter and second, the strong local culture (*e.g.*, sports, music, TV, etc) of US.

Local profile: This is formed by a group of countries that keep local a higher number of links than those going to US or other countries. This is *Local* > *US* & *Local* > *Other* in Figure 4.40. This profile includes Brazil, The Netherlands, Indonesia, Germany and Spain. All these countries have an official language different than English and present a relatively high popularity for Twitter. Moreover, we found also some noticeable differences within the group. On the one hand, Brazil is the country showing the highest locality in our dataset with almost 80% of local links. This is because it is a big country with a strong local culture and the spoken language (Portuguese) is not very spread. Just other countries, not very representative in Twitter, such as Portugal use Portuguese. On the other hand, we have Spain whose local links are reduced to 41%, since now many relationships (> 20%) are established with Latin-America. Note that Spain shares a common language with most south and central American countries.

Shared profile: This group is formed by those countries that distribute their *friend→follower* links roughly equally among those that remain local, those that go to US and those that go to other countries. This profile includes France, Mexico, Italy an Japan that are those countries where Twitter is less popular among the studied ones.

English profile: This group is formed by all those countries from our dataset where English is the official or a co-official language (apart from US): UK, Canada, Australia and India. In addition, all these countries are members of the Commonwealth of Nations. Language becomes the major driver to define the geographical properties of the links originated in these countries. The demographic predominance of US (another English speaking country) produces that the major fraction of links originated in the countries within this group are destined to US (*e.g.*, 48% in the case of India and 47% in the case of Australia and Canada). We refer to this phenomenon as *External locality*. Furthermore, a lower but also important portion of links stay local (*e.g.*, 34% for UK and 31% for Canada) and the rest are shared mainly with other English speaking countries.

In summary, the results reveal that there are three main drivers that define the locality profile for the friend→follower relationships originated in a specific country namely, the language and culture of the country and the local popularity of Twitter. The combination of these factor highlights the presence of four different profiles.



(a) UK (b) FR

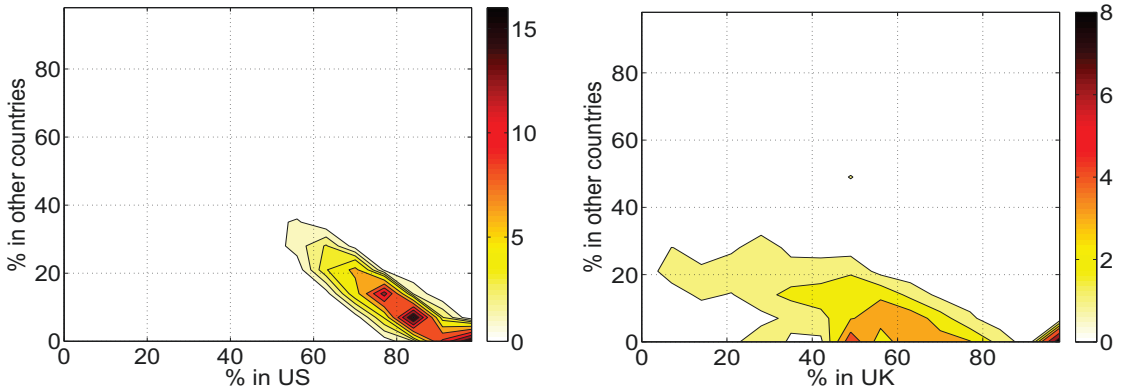
(c) BR

Figure 4.41: Percentage of *friend→follower* relationships that remain local vs. those that go to US for each individual user within the following countries: US, UK, France and Brazil

4.4.2.4 User-level Analysis

Again for this analysis we group the users per country and consider the Top 14 countries. For every friend in a specific country we calculate the fraction of *friend→follower* links that stay local within the country, go to US and go to another country different than US. Due to space limitations, in this section we present results for one representative country per each defined profile above. Specifically, we consider the country with the largest number of users from each profile. These countries are: Brazil for the local profile, France for the shared profile, UK for the English profile and US since it represents a unique profile. Note that the described experiments have been conducted for every country within each profile and the obtained results lead to similar conclusions to those presented in this thesis.

Figures 4.41 and 4.42 depict density diagrams in which the x-axis represents the percentage of *friend→follower* links that remain local and the y-axis represent the percentage of *friend→follower* links that go to either US (Subfigures 4.41(a,b,c)) or another country



(a) US (b) UK

(c) FR (d) BR

Figure 4.42: Percentage of *friend→follower* relationships that remain local vs. those that go to a country different than the US for each individual user within the following countries: US, UK, France and Brazil

(Subfigures 4.41(d,e,f,g)) for each individual user within each analyzed country.

The results show clear differences across the studied countries. First, we can observe that the intra-country locality grows in the following order: BR (locality Profile) > US > FR (Shared Profile) > UK (English Profile, presenting an external locality phenomenon). Specifically, most of the Brazilian users have between 80% and 100% of internal followers, whereas in US we observe a slightly lower intra-country locality where users present a percentage of local followers between 70% and 90%. Looking at the European countries, we observe a higher level of localization in France where the vast majority of users show between 40% and 80% of local followers, whereas the UK presents a less concentrated diagram where the percentage of local followers per user ranges from 20% to 80%. Moreover, we observe how the remote followers of UK are more concentrated in US whereas French users tend to have a balanced presence of followers in US compared to other countries.

		User Level		Link Level	
Country	Distance Limit (Km)	α	x_{min}	α	x_{min}
US	All	27.45	15K51	26.86	16K09
UK	$\leq$ 5K	1.89	221.72	2.32	659.72
UK	$>$ 5K	7.32	7K04	4.94	6K33
FR	$\leq$ 5.5K	2.01	295.91	2.36	540.70
FR	$>$ 5.5K	6.93	6K98	16.58	8K2
BR	$\leq$ 6K	4.00	1K16	4.20	1K79
BR	$>$ 6K	7.02	7K22	5.38	8K42

Table 4.11: Power law parameters for the distribution of user- and link-level distances for US, UK, France and Brazil. For those distribution having two differentiated parts we present specific parameters for each part.

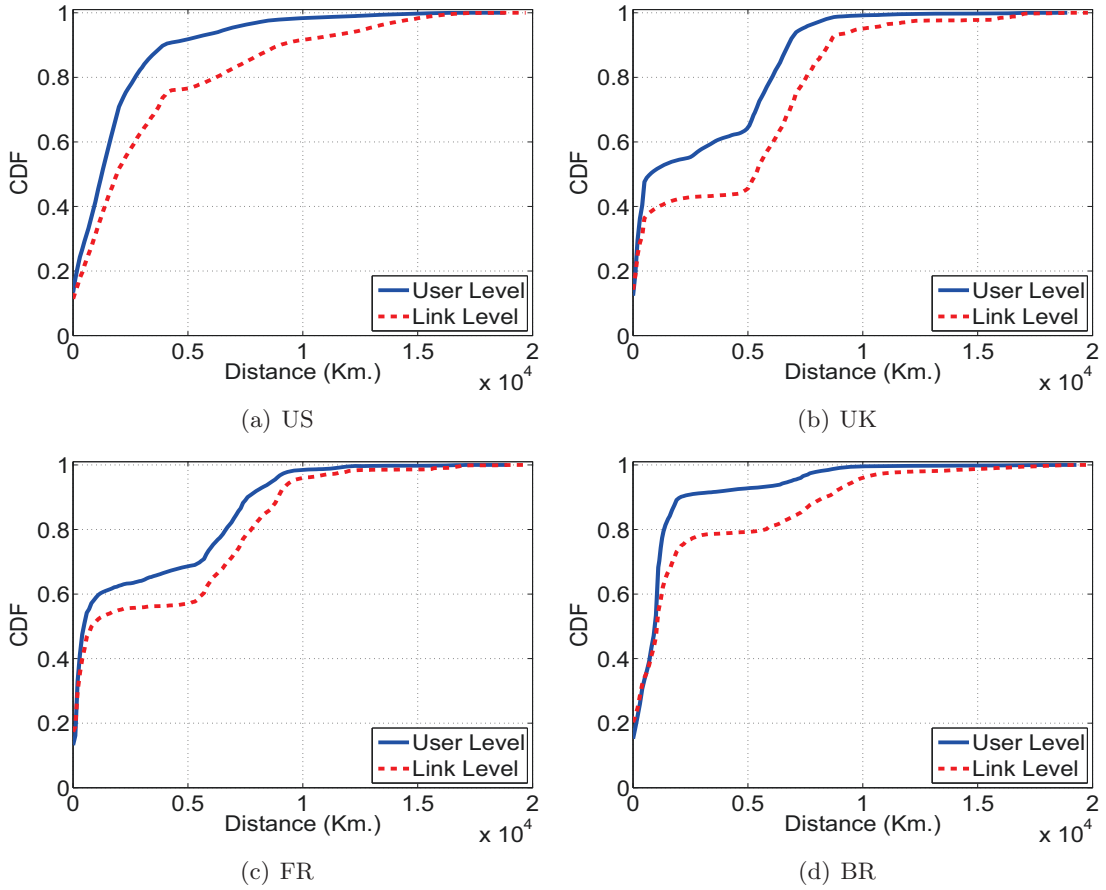
4.4.2.5 Distance-based Analysis

The previous subsection has demonstrated the presence of clearly differentiated profiles across the studied countries. In this subsection we use geographical distance associated to *friend→follower* links instead of GPC communities information (*i.e.*, user's country) in order to validate and extend our previous observations. Due to space limitations, we again present results for one representative country per profile that are Brazil, France, UK and US. We have repeated the experiments for the rest of Top 14 countries and we conclude that the overall observations presented in this thesis are generally valid.

As in the previous subsection, we perform a twofold analysis: link- and user-level analyses. The link level analysis considers separately each individual link originated in a specific country. As mentioned before this makes that popular users have a major impact in the observed results than unpopular users since the former contribute more links. In order to perform the user level analysis we have to calculate a distance metric that characterizes the typical distance from a friend to its followers. To this end, we compute the user-level distance as the median of all the *friend→follower* distances associated to a friend.

Figure 4.43 presents the distribution of *link-level* and *user-level* distances for each one of the analyzed countries. In addition, Table 4.11 shows the analytical distribution that best fit the empirical link- and user-level distribution for each country. In particular, we have used a power-law fitting technique [138], and in those cases where the distribution has two differentiated parts (*i.e.*, UK, FR and BR) we have applied the fitting technique separately for each part. Finally, we have computed a Kolgomorov-Smirnov test [139] for each empirical/analytical distributions pair and confirmed the accuracy of all the presented analytical distributions.

We observe that around 90% of US users have a typical user-level distance to its followers $\leq$ 4000km that defines the intra-country boundary for most relationships originated in US. This intra-country locality effect is even more impressive in Brazil where 90% of



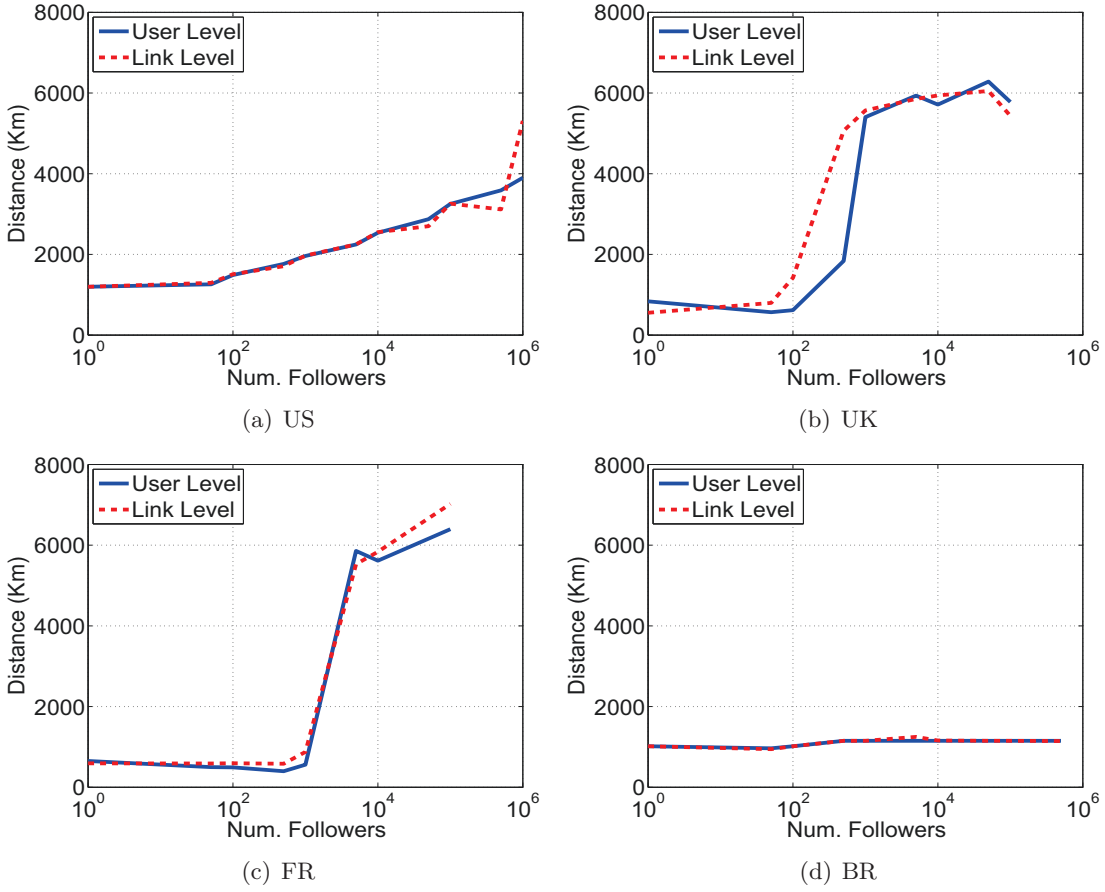
(a) US (b) UK

(c) FR (d) BR

Figure 4.43: Distribution of user- and link-level distances for US, UK, France and Brazil

the users have a user-level distance $\leq$ 2000km, when the limit for most intra-country relationships is also about 4000km. If we analyze UK, it shows, a clear bi-polar distribution that validates the observation done by our geopolitical analysis. Around 60% of links have an associated link-level distance over 5000km that correspond to cross-continental links from which a major portion goes to US. Furthermore, around 40% of links have an associated link-level distance of few hundreds km that correspond to local relationships. If we focus now on France, 60% of its links have an associated link-level distance shorter than 1000km. Several neighbor countries such as Belgium, Switzerland[14], The Netherlands, Italy and Germany are located within this distance range. Hence, this 60% of links is divided into intra-country relationships and inter-country relationships with followers located in neighbor countries. In addition, around 1/3 of the French users present a user-level distance to its followers between 5500 and 9500Km, which mostly represents the followers population in US. Therefore, our distance-based analysis validates the observations done during our geopolitical-based analysis and the presence of four different profiles.

[14]Note that French is co-official language in both Belgium and Switzerland.



(a) US (b) UK

(c) FR (d) BR

Figure 4.44: Median link- and user-level distance as function of the users' popularity for US, UK, France and Brazil

Finally, we observe that every country shows a higher locality (more skewed curve) at the user-level than at the link-level. This suggests that unpopular users tend to have a more localized followers population than popular users. In order to confirm this hypothesis we group the users by its popularity[15] (*i.e.*, number of followers) and for each group we calculate the median for user- and link-level distances. Figure 4.44 shows the obtained results. In general, we observe that our hypothesis is correct since more popular users typically present a larger user-level distance and their relationships show a higher link-level distance. However, we observe significant differences among the analyzed countries that are worth to discuss. US shows a quasi-linear correlation between popularity and locality. The higher the popularity is the longer are the user's *friend→follower* links. Contrary, Brazil users show a high intra-country locality (median distances around 1000km) that is almost independent of their popularity (*i.e.*, the curve is almost flat). Finally, we can observe a clearly denoted bi-polarity in UK and France. In UK those unpopular users

[15]We group the users in the following popularity buckets as function of the number of followers: [1-50],[51-100],[101-500],[501-1000],[1001-5000],[5001-10000], [10001-50000], [50001-100000], [100001-500000] and a last bucket including all those users having > 500K followers.

with less than 100 followers present a clearly marked intra-country locality, whereas the popular users show an *external* locality phenomenon with most of its followers in other continents (mainly US). In France we observe the same bi-polar phenomenon but the transition happens for 1000 rather than 100 followers.

4.4.2.6 Summary

The geopolitical- and distance-based analyses conducted in this section have revealed important insights on the geographical properties of *friend→follower* relationships in Twitter. The combination of language, culture and Twitter popularity has a clear influence in the locality level of the users' relationships in different countries. Indeed, these factors produce the presence of four different country profiles that we have thoroughly discussed along the section. Furthermore, the conducted user- vs link-level distance analysis have demonstrated that locality and popularity are generally inversely proportional. However, the level of correlation varies across countries.

The insights revealed on this section demonstrate that the user's GCP community (*i.e.*, country) clearly impacts its relationships. Moreover, we have showed how the combination of factors such as language and local Twitter popularity produces interesting interactions between different GCP communities (*e.g.*, external-locality phenomenon).

4.4.3 Twitter Information Flows' Locality

The goal of this section is understanding the level of locality existing in the information flow in Twitter. For this purpose we use our *Tweets Dataset* that includes more than 250K Twitter conversations. We first compare the locality level observed in the conversations generated in different GCP communities. Again in this section, we use GCP communities formed by users within a country. Afterwards we study how the popularity of Twitter conversations influences their level of locality.

4.4.3.1 Locality of Twitter Conversations in Different Countries

Figure 4.45 shows the CDF of the percentage of retweets done from a different country than that one where the conversation was originated. The figure shows results for all the conversations in our datasets (All) as well as conversations originated in US, Brazil, France and UK (representative countries of each profile defined in Section 4.4.2). Let us first analyze the aggregate behaviour by looking at the curve associated to "All" conversations. We observe, that in general Twitter conversations show a low locality. Specifically, just 10% of the conversation remain local within a country whereas more than 20% of the conversations have all the retweets in different countries than the country

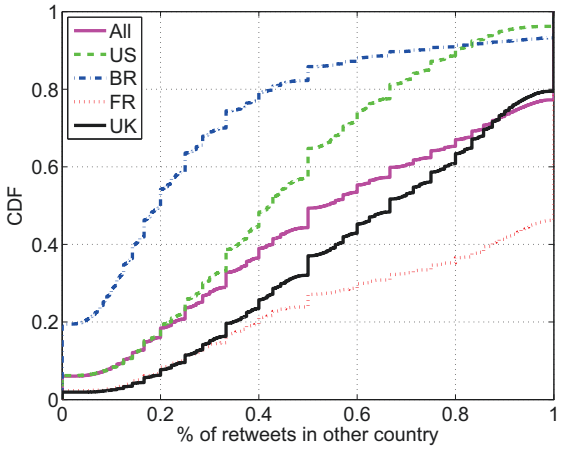


Figure 4.45: CDF of the percentage of retweets posted from a different country than the original tweet for "All" conversations and conversations originated in US, UK, France and Brazil

associated to the original tweet. If we focus now on different countries, as expected, we observe very different behaviours. On the one hand, US and Brazil show a higher locality level compared to the aggregate trend represented by "All". Specifically, the conversations originated in Brazil present the highest locality level (70% of the conversation present at least 70% of local retweets) clearly above the level shown by conversation generated in US. On the other hand, the conversations generated in UK and France show a locality level below than the aggregate trend. In the case of UK, the fact that English is a widespread language and the predominance of US in the number of users ease that conversations originated in UK rapidly move outside the country. France, shows a surprisingly low locality since more than half of the conversation originated in France have all its retweets outside France. This seems to be a consequence of the low popularity of Twitter in the country.

4.4.3.2 Influence of Popularity in the Locality of Twitter Conversations

We have divided the conversations in the four following groups based on their number of retweets (r): $r < 10$, $10 \leq r < 50$, $50 \leq r < 100$ and $r \geq 100$.

Figure 4.46 shows the CDF of the percentage of retweets done from a different country than that one where the conversation was originated for the defined popularity groups. Furthermore, we add the curve including all the conversations (All) for reference. We observe that the different distributions are relatively close to each other. This suggest that the influence of the popularity of conversations in their locality is small. Only those conversations with > 100 retweets present a relatively significant lower locality than the other groups what is an expected result.

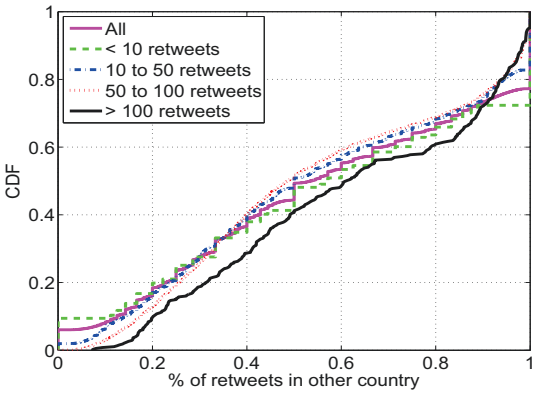


Figure 4.46: CDF of the percentage of retweets posted from a different country than the original tweet for different popularity levels of the conversation.

4.4.3.3 Summary

In this section we have studied the level of locality of more than 250K Twitter conversations. First, we have observed that Twitter conversations present a rather low locality since just 10% of them remain fully local within a country. Furthermore, our results reveal that the origin GCP community (*i.e.*, country) of the conversation have a much higher impact on the locality level than the popularity of the conversation. Indeed, the low impact of the conversation popularity in its locality level is a surprising result, since as occurred in the case of relationships we expected that locality level of Twitter conversations were correlated with their popularity.

Finally, the analysis of individual countries shows that the locality levels associated to relationships and conversations (*i.e.*, information flow) are clearly correlated for a country. Therefore, we can conclude that drivers such as language or Twitter popularity determine the overall level of locality observed at both the relationship and the information flow level.

Chapter 5

Conclusions and Future Work

Understanding the Online Social Networks has a key importance in order to improve both the Internet service itself and the services constructed over these social overlays. In this thesis we, have first presented a measurement framework used to obtain representative data from OSNs. Then, we have address two fundamental problems. On the one hand, we have characterized the birth and evolution of a Second Generation OSN such as Google+ in order to understand how the dynamics of the Social Media Market can be affected by the irruption of new players. On the other hand, we have carefully analyzed the information propagation aspects of two major social networks as Google+ and Twitter. In particular, we have presented the first study analyzing the full dissemination of a piece of content through propagation forests instead of individual propagation trees for a major OSN such as Google+. Furthermore, we have analyzed the geographical properties of information propagation in Twitter. Finally, to complement our analysis of the information propagation we have modelled the tweets arrival process in Twitter.

Our analysis of the birth and growing of G+, a Second Generation Social Network, reveals the following insights:

(i) Contrary to some widespread opinion, G+ is not really a "ghost town". First, the number of interested users who connect to the LCC of the network, is growing at an increasing rate. However, this rate is lower than the one depicted by official reports that most likely include a large number of singletons. These users appear to be automatically registered in G+ after creating a Google account to use other popular Google services. Second, the overall rate of actions and reactions is steadily growing in G+ which is a positive indicator about the level of user engagement.
(ii) Despite the growth in user population and activity, the connectivity and activity features of G+ seem to have reached a statistically stable state after the first year.
(iii) In this seemingly mature status our detailed analyses of connectivity and activity features reveal that Google+ is used as a broadcast social media system in which a relative

small group of popular and very active users contribute most of the posts and attract most users reactions.

In the case of characterization of the information propagation in social networks the most important results obtained are:

(i) A standard post is disseminated quicker in Twitter, but it attracts more reshares and travels longer paths in G+. Furthermore, the probability of getting a post reshared is higher in G+ than in TW. In addition, we have demonstrated that the external content is usually posted for individual users rather than been reshared around influential users. Moreover, while the lifespan of a given post is very small, the lifespan of a content inside the social network is usually very large and depends a lot in the kind of content. Finally, the social graph does not have a key role in the content dissemination, on the contrary, external factors as the content popularity or internals as the G+ hot topics or the Communities are responsible of most of the content dissemination in the network.

(ii) different countries show different *Follower Locality* profiles mostly influenced by the language and cultural characteristics of the country. On the one corner, we have countries with an extremely high intra-country Locality such as Brazil where most of its users keep local 80 to 90% of the followers. On the other extreme, we have countries experiencing an external Locality phenomenon such as Australia where 50% of the *friend→follower* links goes to US while just 25% keeps local within the country. Furthermore, we have seen that US is the dominant country in Twitter responsible for around half of the friends, followers and links in our dataset. Moreover, we conclude that the information usually do not remain in the country of the tweet original publisher user for the case of popular tweets.

(iii) Our statistical analysis demonstrates that the aggregated tweet arrival process can be approximated (but not accurately modelled) by a Gaussian process, which may further permit statistical inference and forecasting.

The previous findings can provide other researcher with a solid base in order to design better services. In particular this results can be extended in order to:

(i) Improve the content distribution using the social data provided. In this way the knowledge of the social interactions can be used in two main ways. First, in order to improve the design of the system itself the existence of locality in the systems indicate the possibility of efficiently distribute the system while the clear day-night pattern suggest some moments where the data replication can be done. And second, to help the distribution of other services by predicting where the content is going to be requested (*i.e.*, a Youtube video shared in Twitter).
(ii) Understand the user behavior in order to improve the advertisement in Online Social

Networks or even how to use them for other marketing purposes. This understanding of the users can help also in the detection of spammers or relevant topics.

As future work we plan the use the results obtained in order to address this two specific topics.

References

[1] "Nielsen: The digital consumer report," 2014.

[2] "Facebook." `https://www.facebook.com/`. [Online; accessed June-2014].

[3] "Linkedin." `https://www.linkedin.com/`. [Online; accessed June-2014].

[4] "last.fm." `http://www.last.fm/`. [Online; accessed April-2013].

[5] "Twitter." `http://www.twitter.com`. [Online; accessed June-2014].

[6] "Sina weibo." `http://www.weibo.com/`. [Online; accessed November-2013].

[7] "Renren." `http://www.renren.com/`. [Online; accessed November-2013].

[8] "Tuenti." `http://www.tuenti.com/`. [Online; accessed June-2014].

[9] M. Cha, H. Haddadi, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proc. of AAAI ICWSM*, 2010.

[10] C. Honeycutt and S. C. Herring, "Beyond microblogging: Conversation and collaboration via Twitter," *Proc. of Hawaii International Conference on System Sciences*, 2009.

[11] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: understanding microblogging usage and communities," in *Proc. of WebKDD/SNA-KDD '07*, 2007.

[12] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," in *Proc. of WOSN'08*, 2008.

[13] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

[14] D. Zhao and M. B. Rosson, "How and why people Twitter: the role that micro-blogging plays in informal communication at work," in *Proc. of ACM GROUP '09*.

[15] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW*, 2007.

[16] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location-based online social network: Analysis and models," in *Proceedings of ACM Internet Measurement Conference (IMC 2012)*, 2012.

[17] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook friends: social capital and college students use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[18] C. Dwyer, S. R. Hiltz, and K. Passerini, "Trust and privacy concern within social networking sites: A comparison of facebook and myspace.," in *AMCIS*, p. 339, 2007.

[19] S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[20] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3, pp. 590–614, 2002.

[21] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *Proceedings of the 17th international conference on World Wide Web*, pp. 695–704, ACM, 2008.

[22] 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*, pp. 33–42, ACM, 2012.

[23] N. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song, "Evolution of attribute-augmented social networks: Measurements, modeling, and implications using google+," in *ACM IMC*, 2012.

[24] G. Magno, G. Comarela, D. Saez-Trumper, M. Cha, and V. Almeida, "New kid on the block: Exploring the google+ social graph," in *ACM IMC*, 2012.

[25] "Google official blog." `http://googleblog.blogspot.com.es/2013/10/google-hangouts-and-photos-save-some.html`. [Online; accessed November-2013].

[26] `http://online.wsj.com/article/SB10001424052970204653604577249341403742390.html?mod=WSJ_hp_LEFTTopStories`. [Online; accessed March-2012].

[27] B. S. Greenberg, "Person-to-person communication in the diffusion of news events," *Journalism & Mass Communication Quarterly*, vol. 41, no. 4, pp. 489–494, 1964.

[28] J. J. Brown and P. H. Reingen, "Social ties and word-of-mouth referral behavior," *Journal of Consumer Research*, pp. 350–362, 1987.

[29] E. Sun, I. Rosenn, C. Marlow, and T. M. Lento, "Gesundheit! modeling contagion through facebook news feed.," in *ICWSM*, 2009.

[30] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in flickr," in *Proceedings of the first workshop on Online social networks*, pp. 13–18, ACM, 2008.

[31] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*, pp. 721–730, ACM, 2009.

[32] B. Yu and H. Fei, "Modeling social cascade in the flickr social network," in *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, vol. 7, pp. 566–570, IEEE, 2009.

[33] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, 2003.

[34] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 199–208, ACM, 2009.

[35] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, (New York, NY, USA), pp. 7–15, ACM, 2008.

[36] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, (New York, NY, USA), pp. 241–250, ACM, 2010.

[37] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, (New York, NY, USA), pp. 1029–1038, ACM, 2010.

[38] M. R. Subramani and B. Rajagopalan, "Knowledge-sharing and influence in online social networks via viral marketing," *Commun. ACM*, vol. 46, pp. 300–307, Dec. 2003.

[39] J. Surma, M. Roszkiewcz, and J. Wojcik, "Towards understanding social influence in on-line social networks," in *Computational Science and Computational Intelligence (CSCI), 2014 International Conference on*, vol. 2, pp. 257–259, IEEE, 2014.

[40] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65–74, ACM, 2011.

[41] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren, "Google+ripples: A native visualization of information flow," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (Republic and Canton of Geneva, Switzerland), pp. 1389–1398, International World Wide Web Conferences Steering Committee, 2013.

[42] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," 1999.

[43] "Pingdom: Twitter growing pains cause lots of downtime in 2007." `http://royal.pingdom.com/2007/12/19/twitter-growing-pains-cause-lots-of-downtime-in-2007/`. [Online; accessed January-2011].

[44] "Examiner.com: Twitter users shocked to lose all their followers." http://www.examiner.com/sf-in-san-francisco/top-news-san-francisco-twitter-users-shocked-to-lose-all-their-followers.

[45] "CNET news: Twitter crippled by denial-of-service attack." `http://news.cnet.com/8301-13577_3-10304633-36.html`. [Online; accessed January-2011].

[46] "Twitter Blog: A Perfect Storm.....of Whales." `http://engineering.twitter.com/2010/06/perfect-stormof-whales.html`. [Online; accessed January-2011].

[47] S. Buchegger, D. Schiöberg, L.-H. Vu, and A. Datta, "Peerson: P2p social networking: early experiences and insights," in *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, SNS '09, (New York, NY, USA), pp. 46–52, ACM, 2009.

[48] A. Shakimov, A. Varshavsky, L. P. Cox, and R. Cáceres, "Privacy, cost, and availability tradeoffs in decentralized osns," in *Proceedings of the 2nd ACM workshop on Online social networks*, WOSN '09, (New York, NY, USA), pp. 13–18, ACM, 2009.

[49] T. Xu, Y. Chen, J. Zhao, and X. Fu, "Cuckoo: towards decentralized, socio-aware online microblogging services and data measurements," in *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement*, HotPlanet '10, (New York, NY, USA), pp. 4:1–4:6, ACM, 2010.

[50] D. R. Choffnes and F. E. Bustamante, "Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems," in *Proc. of ACM SIGCOMM '08*.

[51] R. Cuevas, N. Laoutaris, X. Yang, G. Siganos, and P. Rodríguez, "Deep diving into BitTorrent locality," in *INFOCOM 2011. 30th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 10-15 April 2011, Shanghai, China*, IEEE, 2011.

[52] F. Picconi and L. Massoulié, "ISP friend or foe? making P2P live streaming ISP-aware," in *Proc. of IEEE ICDCS'09*, 2009.

[53] H. Xie, Y. R. Yang, A. Krishnamurthy, Y. Liu, and A. Silberschatz, "P4P: Provider portal for applications," in *Proc. of ACM SIGCOMM'08*.

[54] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks," in *Proc. of ACM CoNEXT '10*, 2010.

[55] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The little engine(s) that could: scaling online social networks," in *Proc of. ACM SIGCOMM'10*, 2010.

[56] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, 2007.

[57] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of Youtube network traffic at a campus network - measurements, models, and implications," *Comput. Netw.*, vol. 53, 2009.

[58] C. Wu, B. Li, and S. Zhao, "Characterizing peer-to-peer streaming flows," *IEEE J. Select. Areas Commun*, 2007.

[59] Y. Huang, T. Z. Fu, D.-M. Chiu, J. C. Lui, and C. Huang, "Challenges, design and analysis of a large-scale P2P-VOD system," SIGCOMM '08, ACM, 2008.

[60] T. Silverston, O. Fourmaux, A. Botta, A. Dainotti, A. Pescapé, G. Ventre, and K. Salamatian, "Traffic analysis of peer-to-peer IPTV communities," *Comput. Netw.*, March 2009.

[61] J. S. Otto, M. A. Sánchez, D. R. Choffnes, F. E. Bustamante, and G. Siganos, "On blind mice and the elephant: understanding the network impact of a large distributed system," SIGCOMM '11, ACM, 2011.

[62] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding online social network usage from a network perspective," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, (New York, NY, USA), pp. 35–48, ACM, 2009.

[63] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks," Co-NEXT '10, ACM, 2010.

[64] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, and S. Lopez-Buedo, "Characterization of the busy-hour traffic of IP networks based on their intrinsic features," *Comput. Netw.*, vol. 55, June 2011.

[65] K. Cho, K. Fukuda, H. Esaki, and A. Kato, "Observing slow crustal movement in residential user traffic," CoNEXT '08, ACM, 2008.

[66] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," SIGCOMM '10, ACM, 2010.

[67] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[68] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?," in *WWW*, 2010.

[69] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *ACM IMC*, 2007.

[70] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 49–62, ACM, 2009.

[71] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *ACM SIGIR*, 2012.

[72] L. Gyarmati and T. Trinh, "Measuring user behavior in online social networks," *Network, IEEE*, vol. 24, no. 5, pp. 26–31, 2010.

[73] C. Ding, Y. Chen, and X. Fu, "Crowd crawling: Towards collaborative data collection for large-scale online social networks," in *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, (New York, NY, USA), pp. 183–188, ACM, 2013.

[74] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *ACM WSDM*, 2010.

[75] R. Kumar, J. Novak, and A. Andtomkins, "Structure and evolution of online social networks," in *ACM KDD*, 2006.

[76] A. Mislove, H. Koppula, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *WOSN*, 2008.

[77] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Zhao, "Multi-scale dynamics in a massive online social network," in *ACM IMC*, 2012.

[78] S. Gaito, M. Zignani, G. Rossi, A. Sala, X. Wang, H. Zheng, and B. Zhao, "On the bursty evolution of online social networks," in *ACM KDD HotSocial Workshop*, 2012.

[79] S. Garg, T. Gupta, N. Carlsson, and A. Mahanti, "Evolution of an online social aggregation network: an empirical study," in *ACM IMC*, 2009.

[80] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger, "Sizing Up Online Social Networks," *IEEE Network*, vol. 24, pp. 32 –37, Sept-Oct 2010.

[81] H. Liu, A. Nazir, J. Joung, and C.-N. Chuah, "Modeling/predicting the evolution trend of osn-based applications," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (Republic and Canton of Geneva, Switzerland), pp. 771–780, International World Wide Web Conferences Steering Committee, 2013.

[82] J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai, and B. Y. Zhao, "Understanding latent interactions in online social networks," *ACM Trans. Web*, vol. 7, pp. 18:1–18:39, Nov. 2013.

[83] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini, "The role of information diffusion in the evolution of social networks," in *Proc. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[84] A. Boutet, A. Kermarrec, E. Le Merrer, and A. Van Kempen, "On the impact of users availability in osns," in *ACM SNS*, 2012.

[85] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Computational Intelligence*, September 2013.

[86] D. Schiöberg, S. Schmid, F. Schneider, S. Uhlig, H. Schiöberg, and A. Feldmann, "Tracing the birth of an osn: Social graph and profile analysis in google+," in *Proceedings of the 3rd Annual ACM Web Science Conference*, pp. 265–274, ACM, 2012.

[87] S. Kairam, M. Brzozowski, D. Huffaker, and E. Chi, "Talking in circles: selective sharing in google+," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1065–1074, ACM, 2012.

[88] Y. Xu, C. Yu, J. Li, and Y. Liu, "Video telephony for end-consumers: Measurement study of google+, ichat, and skype," in *ACM IMC*, 2012.

[89] L. Fang, A. Fabrikant, and K. LeFevre, "Look who i found: Understanding the effects of sharing curated friend groups," in *ACM WebSci*, 2012.

[90] H. Hu, G.-J. Ahn, and J. Jorgensen, "Enabling collaborative data sharing in google+," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, pp. 720–725, IEEE, 2012.

[91] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services.," *ICWSM*, vol. 2011, pp. 81–88, 2011.

[92] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks.," in *ICWSM*, 2012.

[93] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare.," *ICWSM*, vol. 11, pp. 70–573, 2011.

[94] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *SIGMOD Rec.*, vol. 42, pp. 17–28, July 2013.

[95] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?," *ICWSM*, vol. 10, pp. 34–41, 2010.

[96] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[97] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, (New York, NY, USA), pp. 497–506, ACM, 2009.

[98] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 7–15, ACM, 2008.

[99] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[100] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, p. 4, ACM, 2010.

[101] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270, ACM, 2010.

[102] S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris, and K. Papagiannaki, "Tailgate: handling long-tail content with a little help from friends," in *Proceedings of the 21st international conference on World Wide Web*, pp. 151–160, ACM, 2012.

[103] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, "Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades," in *Proceedings of the 20th international conference on World wide web*, pp. 457–466, ACM, 2011.

[104] H. Li, J. Liu, K. Xu, and S. Wen, "Understanding video propagation in online social networks," in *Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service*, IWQoS '12, (Piscataway, NJ, USA), pp. 21:1–21:9, IEEE Press, 2012.

[105] Q. Liu, X. Zhao, S. Smith, B. Y. Zhao, H. Zheng, and W. Willinger, "On the self-similarity of social network dynamics," *GSWC 2013*, p. 37, 2013.

[106] "Gnip.com." http://gnip.com/. [Online; accessed January-2013].

[107] "Datasift.com." `http://datasift.com/`. [Online; accessed November-2013].

[108] "Facebook for business." `https://www.facebook.com/business/products/ads`. [Online; accessed November-2013].

[109] "Google AdWords." http://www.google.com/adwords/how-it-works/target-your-ads.html.

[110] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose," *Proceedings of ICWSM*, 2013.

[111] "Twitter API Documentation." `http://dev.twitter.com/doc`. [Online; accessed January-2012].

[112] "Yahoo Geo Technologies." http://developer.yahoo.com/geo/.

[113] "Buzz shut down announcement." `https://support.google.com/mail/bin/answer.py?hl=en&answer=1698228`. [Online; accessed March-2012].

[114] *Google wave shut down announcement.* `http://support.google.com/bin/answer.py?hl=en&answer=1083134`.

[115] *Orkut Official Site.* `http://www.orkut.com`.

[116] *Orkut Statistics, Wikipedia.* `http://en.wikipedia.org/wiki/Orkut`.

[117] *Google+ Pages.* `http://www.google.com/+/business/`.

[118] *Google+ Pages announcement.* `http://googleblog.blogspot.com/2011/11/google-pages-connect-with-all-things.html`.

[119] *Google Registration. Arstechnica.* `http://arstechnica.com/gadgets/2012/01/google-doubles-plus-membership-with-brute-force-signup-process/`.

[120] *Google Registration. Blogspot.* `http://googlesystem.blogspot.com/2012/01/new-google-accounts-require-gmail-and.html`.

[121] "Flow chart for google+ sharing." `http://googleplushowto.com/2011/07/will-user-a-see-my-post-in-google-plus/`. [Online; accessed March-2012].

[122] "Google+ official learn more site." `http://www.google.com/intl/en/+/learnmore/`. [Online; accessed November-2012].

[123] "Google+ official support site." `http://support.google.com/plus/`. [Online; accessed June-2014].

[124] *Genealogy Data: Frequently Occurring Surnames from Census 2000. US Census Bureau.* `http://www.census.gov/genealogy/www/data/2000surnames/index.html`.

[125] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "Planetlab: an overlay testbed for broad-coverage services," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 3, pp. 3–12, 2003.

[126] `http://google-plus.com/category/statistics/`. [Online; accessed March-2014].

[127] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *ACM SIGKDD*, 2005.

[128] *Google Privacy Policy.* `http://www.google.com/policies/privacy/`.

[129] `http://en.wikipedia.org/wiki/Google+`. [Online; accessed June-2014].

[130] https://plus.google.com/107117483540235115863/posts/YUniwagZuKZ. [Online; accessed July-2012].

[131] "Announcement of new feature for google+." http://www.workinghomeguide.com/9918/google-new-features-hashtag-auto-complete-text-to-photos-and-video-status. [Online; accessed March-2014].

[132] "Announcement of new feature to add text to photos in google+." https://plus.google.com/107814003053721091970/posts/D7gfxe4bU7o. [Online; accessed March-2014].

[133] "Google+ redesign." http://mashable.com/2012/04/11/google-plus-redesign/. [Online; accessed May-2012].

[134] Z. Govindarajulu, "Rank correlation methods," *Technometrics*, vol. 34, no. 1, pp. 108–108, 1992.

[135] *Google+ (maximum number of contacts in your circles).* http://support.google.com/plus/bin/answer.py?hl=en&answer=1733011.

[136] R. Motamedi, R. Gonzalez, R. Farahbakhsh, A. Cuevas, R. Cuevas, and R. Rejaie, "What osn should i use? characterizing user engagement in major osns," technical report available at: http://www.it.uc3m.es/~rgonza1/pubs/whatOSN.pdf, Universidad Carlos III de Madrid, 2013.

[137] E. Almaas, B. Kovacs, T. Vicsek, Z. Oltvai, and A.-L. Barabási, "Global organization of metabolic fluxes in the bacterium escherichia coli," *Nature*, vol. 427, no. 6977, pp. 839–843, 2004.

[138] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[139] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. Vol. 46, No. 23, pp. 68-78, 1951.