Characterization of information propagation in Google+ and its Comparison with Twitter

Roberto González, Rubén Cuevas, Ángel Cuevas
Universidad Carlos III de Madrid
Leganes, Madrid, Spain
{rgonza1,rcuevas,acrumin}@it.uc3m.es

Reza Farahbakhsh
Institut Mines-Telecom, Telecom SudParis
Evry, Île-de-France, France
reza.farahbakhsh@itsudparis.eu

Reza Motamedi, Reza Rejaie
University of Oregon
Eugene, OR, US
{motamedi,reza}@cs.uoregon.edu

ABSTRACT

This paper characterizes the propagation of the information in one of the major OSNs, Google+ (G+). It extends previous works, which study information dissemination in other OSNs, based on two main contributions: (*i*) it is the first work that takes into account all publicly available information in a major OSN to perform such characterization rather than only use a limited sample of the network. (*ii*) To the best of our knowledge, this paper presents the most extensive head-to-head comparison of information propagation between two major OSNs such as Google+ and Twitter. We investigate the answer for two fundamental questions: First, how the information is propagated in G+ and how such propagation compares with the information dissemination in Twitter? Our results reveal that information is disseminated faster in Twitter, but it has more probabilities of being propagated and travel longer paths in G+. Second, are there influential users that play a key role in the information dissemination in G+? and, if so, who are them? Our results identify two type of influential users in G+ that we refer to as *Disseminators* and *Leaders*. The former are relatively popular users that publish a large number of posts that all together reach a large number of users. In addition, they use to publish entertainment content such as funny videos and animations, beautiful pictures and quotes with deep emotional messages. The latter are individual persons that publish few posts that engage a large number of users. Finally, it is interesting to notice that among the Top 50 *Leaders* we find 7 accounts associated to either Google employees or Google product.

Keywords

Online Social Networks, Information Propagation, Ripples, Measurements, Google+.

1. INTRODUCTION

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 evolve 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 [13] or social science [4]. 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) [18], Facebook [27] or Flickr [6, 7, 31]. However all these works just use a sample of the information in those OSNs to perform the analysis of the information propagation. In this paper we aim at characterizing the propagation of information in Google+ using a complete sample of the public available information in this system.

In Google+, similarly to other networks such as Facebook, the basic piece of information is the so called *post*. A post can attach different type 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 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 [29] that provides the reshare tree of each post that has been reshared at least once. In addition, the Ripple of a post 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. We will leverage this data to carefully characterise the main aspects of information propagation in G+. In particular, we divide this paper in two parts each of them addressing fundamental questions in the dissemination of information in G+.

In the first part of the paper 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 [18].

In the second part of this paper we address another interesting question, this is, which persons are influential in the propagation information process in G+? For this purpose we define meaningful metrics to analyze the presence of two types of influential users, namely *Disseminators* and *Leaders*. Disseminators are users that get an overall large number of reshares across all their posts whereas Leaders are users that, regardless of their activity, obtain a large number of reshares for each post they publish in the system. Our analysis reveals the presence of both types of influential users in G+. Then, to conclude this part of the paper we perform a systematic manual investigation to unveil the identity and behaviour of the Top-Disseminators and Top-Leaders in G+.

This paper presents three main contributions that extend the existing work: (*i*) We present the first characterization of information propagation in Google+ (G+). (*ii*) This is the first study that characterizes the information propagation in an OSN making use of all publicly available information. All previous studies have just used a sample of the overall information. (*iii*) To the best of our knowledge, this paper presents the most extensive head-to-head comparison of information propagation between two major OSNs such as Google+ and Twitter.

Finally, the main findings of this paper 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 use to happen in other systems like P2P networks.
- *Disseminators* are popular users that post frequently in the system. Furthermore, they mainly post entertainment content: funny photos or animations, beautiful photos or quotes with deep emotional messages.
- *Leaders* are usually individual persons with low activity that publish personal, professional or funny content. Surprisingly we found 7 accounts directly linked to Google among the Top 50 Leaders. This indicates that G+ users are interested in information related to Google.

The remainder of the paper is structured as follows. In Section 2 we present the methodology and measurements utilized to perform the characterization of the information in G+. We characterize the basic properties for the information propagation in G+ in Section 3. Following, we discuss the presence and characteristics of influential users in Section 4. Section 5 presents related works in the literature. Finally, we provide concluding remarks in Section 6.

2. METHODOLOGY AND DATASETS

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*. 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+.

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.

2.1 Measurement Methodology

Our aim is to collect all posts in G+ and its associated reshare trees. To this end, we follow a methodology divided in 3 phases.

In the first phase we leverage the technique described in our previous work [12] to capture the ids of all users connected to the Largest Connected Component (LCC) of G+ in July 2013.

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

Using as input the collected list of user ids, in the second phase, we leverage the G+ API to collect all public posts of each user in the LCC of G+. Since the G+ API limits the number of queries that can be done from a single IP address to 10K per day, we use a distributed architecture of proxies installed in PlanetLab nodes and more than 600 G+ accounts in order to speed up our data collection process. In particular, using our tool we collected every single (public) post published by every user within the LCC of G+ from the release date of G+ (June 28th 2011) until the date in which we started this second phase (July 3rd 2013). Our crawler needed 72 days to collect 540M (public) posts. For each post the crawler retrieved the following information of interest for this paper: total number of reshares (including both public and private reshares) and type of post (e.g., photo, video, text, etc).

Finally, in the third phase, we leverage a public feature of G+ named *Ripples* [29]. 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 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*.

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.

2.2 Dataset filtering

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`.)

The goal of our paper 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.

2.3 Other datasets

In order to compare the main characteristics of the information propagation in G+ and TW, we have collected a dataset including the number of retweets for 2.3M tweets collected from more than 17K randomly selected users. We refer to this dataset as *TW-retweets*. This dataset was collected between March 28th 2013 and April 2nd 2013. Furthermore, part of our comparison analysis with TW will refer to the results obtained by Kwak et al. [18] using a dataset collected in 2009 (3 years after the relase of TW).

3. 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 [18] or obtained from our *TW-reshares* dataset.

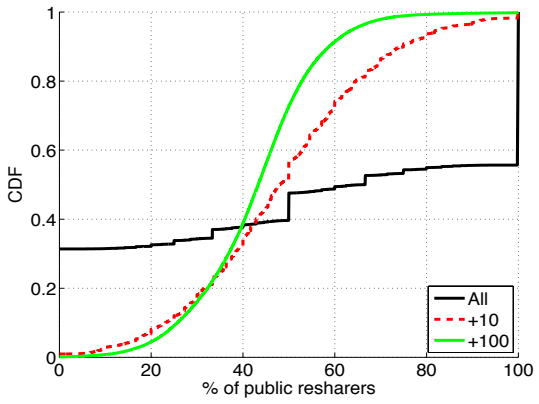


Figure 1: CDF of percentage of public resharers per post. We plot the results for three set of posts grouped according to the number of resahres they attract. (*i*) All posts (*All*), (*ii*) All posts with 10 or more reshares (*+10*), and (*iii*) All posts with 100 or more reshares (*+100*)

3.1 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+ [22] and then TW presents a much longer tail of non-propagated tweets that leads to the reported result.

3.2 Public vs. Private information propagation

In Twitter most of the available information is public due to its broadcasting nature [18]. 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 [16].

We can accurately compute the percentage of public reshares for each post within our *G+ reshares* dataset. As indicated in Section 2, 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

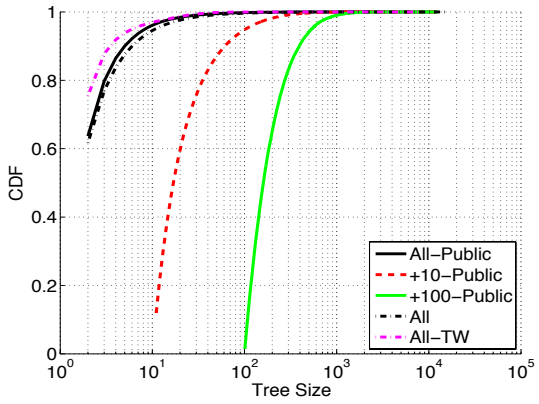


Figure 2: CDF of the tree size per post for different groups of posts within our *G+ reshares* and *TW retweets* datasets.

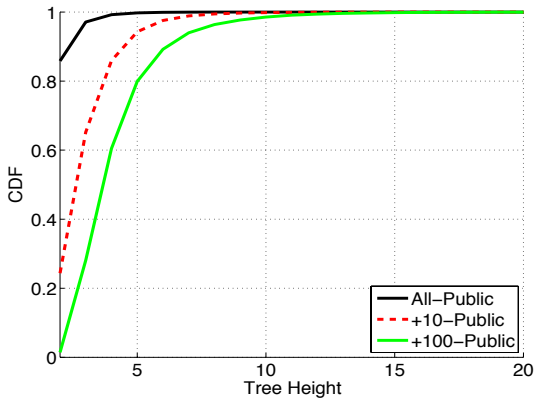


Figure 3: CDF of Tree Height for different groups of posts within our *G+ reshares* dataset.

number of total reshares to obtain the fraction of public reshares for each post in our dataset.

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 1 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 the paper 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 [1]. 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.

3.3 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.

3.3.1 Tree Size Analysis

Let us start by analyzing the distribution of the size of propagation trees in G+. To this end Figure 2 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 $\leq$ 5 and $\leq$ 6 for *All-Public* and *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.

3.3.2 Tree Height Analysis

Now we focus in analyzing the height for the propagation trees in G+. Figure 3 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-*

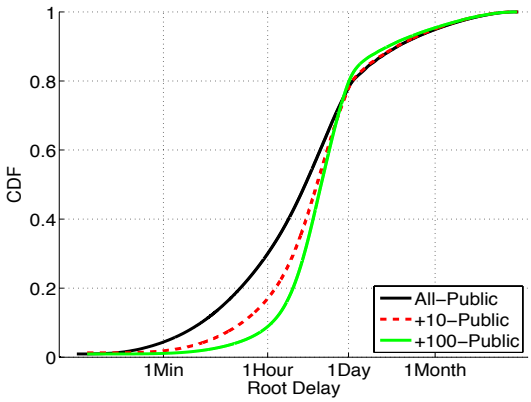
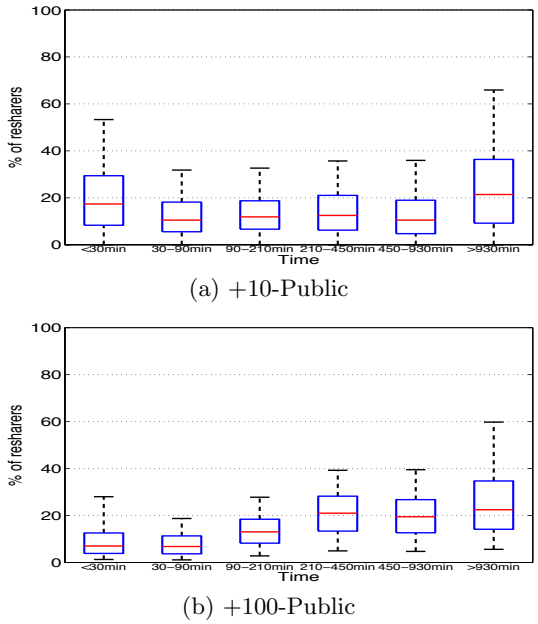


Figure 4: CDF of root delay. We plot the results for three sets of public posts grouped according to the number of public reshares they attract: (*i*) All public posts (*All-public*), (*ii*) posts with 10 or more public reshares (*+10-public*), and (*iii*) posts with 100 or more public reshares (*+100-public)*



(a) +10-Public

(b) +100-Public

Figure 5: Boxplot of the percentage of reshares per tree in different delay time windows.

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. [18] for the comparison with TW.

We observe that 6.8% of *All-Public* trees have a height $\geq$ 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. [18] report a maximum height equal to 11 for Twitter[2].

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

3.4 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+:

- *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. [18] for the comparison between G+ and TW, as we did for the discussion of trees' height.

[2]We would like to remind that the dataset of [18] was collected in 2009, three years after the release of Twitter. Our dataset has been collected two years after the release of G+.

3.4.1 Root Delay Analysis

We start our analysis of the temporal properties by looking at the root delay. Figure 4 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 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 5 shows the obtained results for *+10-Public* and *+100-Public*. In particular, the boxplot[3] 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 5(a)) or *+100-Public* (Figure 5(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.

[3]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 paper follow this definition.

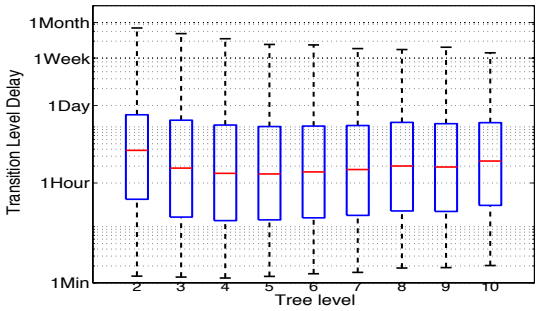


Figure 6: Transition Delay at different levels of the reshare tree

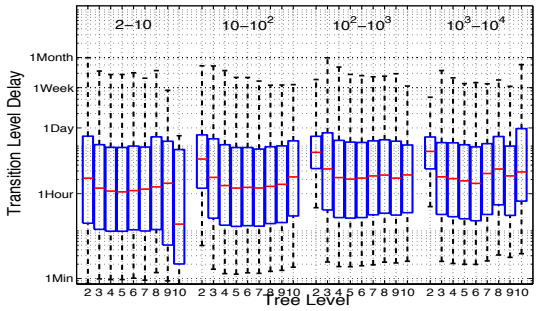


Figure 7: Transition Level Delay at different levels of the reshare tree for four sets of posts grouped by their popularity (i.e., number of public rehsares they attract)

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.

3.4.2 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 6 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 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 7 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 5.

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 6[4].

3.5 Favouritism in Information Reshare

In this subsection we conduct an analysis in order to investigate whether the reshare events made by the followers[5] 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[6] or contrary have a favouritism for few of them.

To address this issue we follow the methodology proposed by Kwak et. all [18] and analyze the disparity [2] 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:

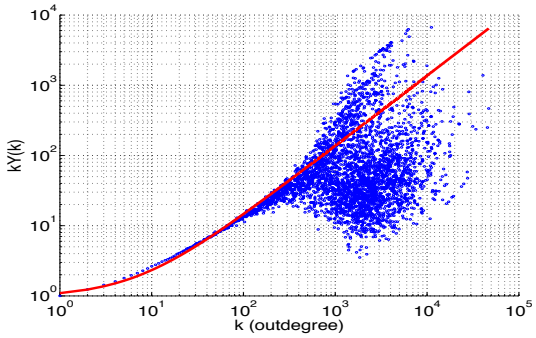
$$Y(k,i) = \sum_{j=1}^{k} \left(\frac{|r(i,j)|}{\sum_{l=1}^{k} |r(i,l)|} \right)^2 \quad (1)$$

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 8 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, the

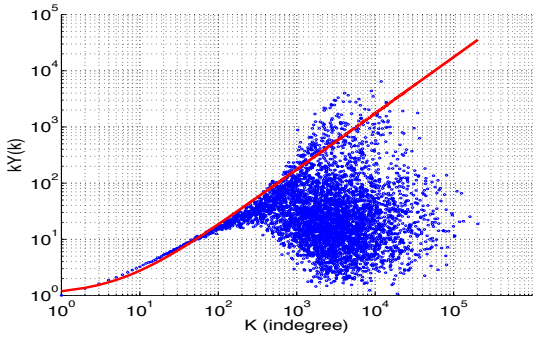
[4]There are few exceptions such as: (i) in the popularity bucket "2-10" the transition delay for level 10 is significantly smaller than for other levels; (ii) in the popularity bucket "10^3-10^4" Levels 6 and 7 breaks the convexity of the curve.

[5]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.

[6]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 8: Disparity in reshare trees

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) ∼ 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.*

3.6 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[7] 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 [22] 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 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.

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

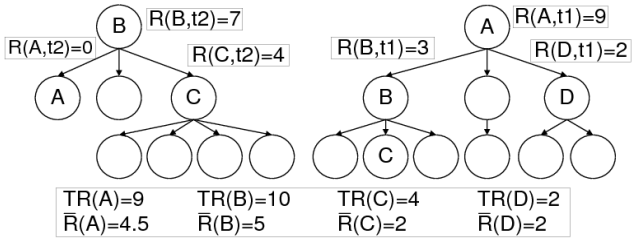


Figure 9: Graphical example to explain the metrics Reach, Total Reach (TR), Avg. Reach ($\overline{R}$) using two propagation trees.

4. ANALYSIS OF USERS' INFLUENCE IN THE INFORMATION PROPAGATION IN G+

The contribution of different users to the information propagation in major OSNs such as Twitter has been reported to be skewed [5]. Indeed, the capacity of a user to disseminate information is dictated by its *influence*. In this section we perform a twofold analysis: First, we study the skewness in the contribution of users to the information propagation in G+. This analysis will reveal the presence (or absence) of influential users. Second, we apply a systematic manual methodology to investigate the identity of the most influential users in G+. Furthermore, we classify them into different groups based on their behaviour and profile information.

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 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:

$$TR(u) = \sum_{t=1}^{T} R(u,t) \tag{2}$$

-*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:

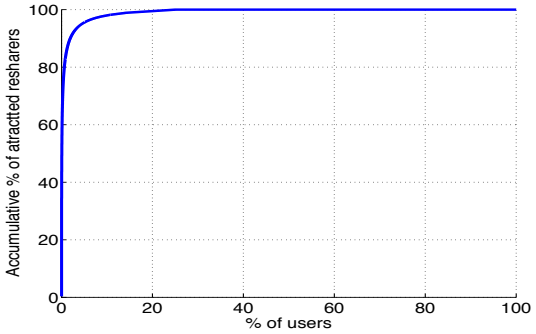


Figure 10: Skewness of the Total Reach across G+ users

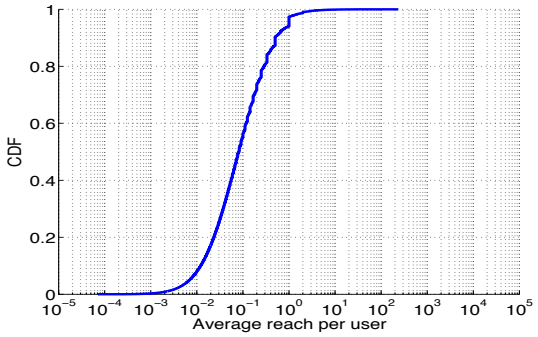


Figure 11: CDF of the Average Reach for G+ users

$$\overline{R} = \frac{1}{T}\sum_{t=1}^{T} R(u,t) \quad (3)$$

In Figure 9 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. Hence, we refer to users with a high TR as *Disseminators*. On the other hand, $\overline{R}$ captures the average number of users reached by each post and reshare of user u. Then, a high $\overline{R}$ defines a user as a *Leader* since each piece of information that he shares attracts the attention of a large number of people.

Note that studying both types of influence is important since they are of interest in different context. For instance, a marketing campaign for a new brand whose objective is to reach as many people as possible so that the new brand gets known would be interested on identifying *Disseminators*. However, *Leaders* seem to suit perfectly as opinion generators. Then a company with a failing product could be interested on exploiting *Leaders* to propagate positive messages about that product.

4.1 Are there influential users in G+?

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 10 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%). Therefore we confirm the presence of *Disseminators* in G+.

Figure 11 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. Hence, we also confirm the presence of *Leaders* in G+.

In short our results demonstrate the presence of both types of influential users, Disseminators and Leaders, in G+.

Metrics Pair	Rank Correlation
TR vs #followers	0.39
TR vs #activities	0.52
$\overline{R}$ vs #followers	-0.04
$\overline{R}$ vs #activities	-0.6

Table 1: Spearman (Rank) Correlation between Users' Influence (TR and $\overline{R}$) , Popularity (#followers) and Activity (#activities).

4.2 Overlapping between Disemminators and Leaders

The presence of two type of influential users in G+, *Disseminators* and *Leaders*, raises an interesting question: *are Disseminators and Leaders the same users?*. In order to answer it we have computed the Spearman Correlation Coefficient of TR with $\overline{R}$ for all users in our dataset. The Spearman Correlation Coefficient is 0.22. *Therefore, the obtained results indicate that there is a low overlapping between the group of* Disseminators *and* Leaders *in G+. In particular, we can only find 11 common users between the Top 100 Disseminators and Leaders.*

4.3 Impact of Popularity and Level of Activity in Users Influence

It is reasonable to think that *Disseminators* and *Leaders* may be popular users with a large number of followers and/or very active users that publish (reshare) a large number of posts in G+.

In this subsection we investigate this hypothesis. To this end, we study the correlation between the influence of a user (measured by its TR or $\overline{R}$) and its popularity (measured by its number of followers) or its level of activity (measured by the total number of activities, which include the posts plus the reshares of the user). In particular, we have computed the Spearman Correlation Coefficient of TR and $\overline{R}$ with the #followers and the #activities for all G+ users in our dataset. Table 1 presents the obtained results.

We observe a noticeable positive correlation between TR and both the popularity and activity of users. This suggests that in order to become a *Disseminator* the user needs to reach a certain level of popularity and activity. However,

Type of Influential Users		Disseminators	Leaders
Nature	Individual	26%	66%
	Non-Individual	74%	34%
Profile	Content Aggregator	64%	22%
	Company	2%	6%
	Celebrity	4%	10%
	Media	8%	0%
	Personal Page	18%	48%
	Google Related	4%	14%

Table 2: Division of Top 50 *Disseminators* and Top 50 *Leaders* according to their Nature and Account Profile.

this is a necessary but not a sufficient condition. Therefore, being very popular and/or very active do not guarantee a large Total Reach.

Furthermore, we observe that the popularity of a user is uncorrelated with its $\overline{R}$. More surprisingly, there is a significant negative correlation between $\overline{R}$ and the user activity. This indicates that *Leaders* are actually low-active users.

In summary, we conclude that neither a high popularity nor a high activity are sufficient elements to become an influential user in G+.

To complete our analysis of the users' influence, we devote the rest of the section to identify who are the *Top-Disseminators* and *Top-Leaders* in G+.

4.4 Characterizing Leaders and Disseminators

We have demonstrated that there is a reduced pool of *Disseminators* and *Leaders* in Google+. In this subsection, our goal is to provide a comprehensive view of who are these users and characterize their activity in the network. To this end, we have identified the Top 50 *Disseminators* and *Leaders* and we have performed a systematic manual inspection of their profiles and activity in G+. Then, our first step has been categorizing each user based on two criteria that help us on building a more meaningful discussion:

Account Nature: In G+ we can find accounts that represent individual users (e.g., regular users, celebrities, politicians, etc) and other accounts that represent something different than individual persons (e.g., industry, services, organizations, public bodies, products, etc). In the rest of the section we refer to the former as *Individual* users and the latter as *Non-Individual* users.

Account Profile: Our manual inspection has revealed 6 different profiles across Top influential users in G+.

- *Content Aggregators*: These users post/reshare a large amount of activities that are usually retrieved from third party sources. Usually the posts of these users contain what we call entertainment content: funny photos or animations, beautiful photos or quotes with deep messages (e.g., "Be selective in your battles, sometimes peace is better than being right"). A good example of *Content Aggregators* is the G+ user "Funny Pictures & Videos".
- *Companies:* The second profile is formed by G+ accounts linked to companies.
- *Celebrities:* This profile includes users that have acquired a relevant fame because their profession expose them in mass media (e.g., actor, singer, artist, politician, etc).
- *Media*: It contains users related with Internet and traditional news and media communication. One example of this profile among Top influential users is the G+ account "Mashable".
- *Personal Page*: This group is formed by individual users that seem to use G+ without any obvious promotion purpose further than express their opinion, share pictures, etc. It includes both, regular (i.e., anonymous) users or relevant persons like the former president of My Space, Tom Anderson.
- *Google Related*: We have found quite a few accounts among Top influential users that are close related to Google. Some of them are G+ accounts for Google employers and some others refer to Google products or services.

Table 2 shows the portion of *Disseminators* and *Leaders* associated to each category for the two employed criteria (Nature and Profile). Next, we characterize separately Top *Disseminators* and *Leaders*.

4.4.1 Top Disseminators

The results reveal that 2/3 of the Top 50 *Disseminators* are actually Content Aggregators that, as we have mentioned above, mainly offer entertainment content to their followers. G+ users seem to appreciate that type of content since it achieves a large number of reshares.

Furthermore, in Section 4.3 we conclude that in order to become a *Disseminators* a G+ user needs to reach a certain level of popularity and activity. Next, we demonstrate that this is a necessary but not a sufficient condition.

On the one hand, in terms of popularity, *Content Aggregators* are relatively popular users with 83K followers in median, while the remaining 18 users within the Top 50 *Disseminators* present a much larger popularity with a median of 2.4M followers. However, we can find 3224 users in G+ users with more than 83K followers, thus more popular than a standard Top *Content Aggregators*, but they do not manage to propagate their posts as *Content Aggregators* do.

On the other hand, all users in the Top 50 *Disseminators* group present a high number of activities ranging between 422 and 13K. Hence, all of them fulfil the requirement of being active. However, we can find more than 3K users in G+ with more than 5K activities, and we only find 10 users with more than 5K activities among the Top 50 *Disseminators*. Therefore, being a very active users does not ensure to engage users to disseminate your information.

Finally, we want to highlight that only 1/4 of the Top *Disseminators* correspond to individual users.

4.4.2 Top Leaders

In this case, we find that 2/3 of Top 50 *Leaders* are individual users, the opposite of what we observe for *Disseminators*. Most of these users present a low-active profile, which confirms that *Leaders* usually present a small activity level as stated in Section 4.3. These individual users are mainly divided into 3 categories: *Personal Page*, *Google related* and *Celebrities*. Following, we describe the main outcomes out of the analysis of each one of these categories:

- *Leaders* within the *Personal Page* category are mostly regular users that are 1 order of magnitude less popular (14.5K followers in median) than *Disseminators*. Furthermore, they use to publish personal, professional and entertainment content.
- It is interesting to notice that among the Top 50 *Leaders* there are 7 accounts related to Google. In particular, we find 5 employers of Google and 2 products such as Nexus and Google Glass. Furthermore, if we extend our range to Top 100 *Leaders*, we would discover more Google related ac-

counts such as Youtube and Android. The presence of quite a lot Google related accounts among Top Leaders suggests that G+ users are very interested in the opinion of Google employers as well as in the information related to Google products and services.
- Apart from Google accounts, we can find two *Leaders* that are well-known professionals in the area of technology: Tom Anderson (former president of MySpace) and Linus Torvalds (the creator of Linux). Therefore, this suggests that G+ users are interested in technology further than the clear interest they show around Google.
- There are only 4 *Celebrities* among the Top 50 *Leaders*. These are popular users with 188K followers in median. Three of these celebrities, Selena Gomez and two components of the Korean band "Infinite", use G+ for promotional purposes. However, the remaining celebrity, the actor George Takei, shares personal content and funny posts as users in *Personal Page* category do.

Finally, we want to notice that *Content Aggregators* have an important role as *Leaders* as well since they represent 22% of the users among Top 50 *Leaders*.

4.5 Summary

We observe a very skewed distribution for the Total Reach and Average Reach among G+ users that confirm the existence of *Leaders* and *Disseminators*. These two type of influential users respond to very different profiles.

Disseminators are typically *Content Aggregators*. These users are mostly *Non-Individual* users, with a relatively high popularity, that have published many posts in the system. However, none of the previous conditions are sufficient for a user to become a *Disseminator*. In fact, the user needs to be able to engage people that propagates (i.e reshares) his information

Leaders are usually individual users that publish a reduce number of posts. Most of these posts attract a large number of reshares. In addition, we have revealed that 14% of the Top 50 *Leaders* are Google related accounts referring to Google employers and products. This demonstrates that G+ users are highly interested on any information coming from the Google environment.

5. RELATED WORK

OSN characterization. The successful irruption of social networks in our daily life has attracted the attention of the research community in the last years. We can find several works that characterize the main properties of the most popular OSNs such as Facebook [3, 28] or Twitter [15, 18]. More interesting for our research, there exist a number of efforts that are focused on the characterization of Google+. The first works in this area studied G+ graph properties [26, 21]. More recent research has made a step further and has investigated the evolution of the graph and the activity of users in G+ [12, 11]. Finally, there are some studies that have focused on concrete aspects like the new circle feature introduced in Google+ [16, 9] or the study of collaborative privacy management solutions [14].

Information propagation characterization. The propagation of information as well as the factors that influence it have been historically studied in areas like social science [4] or traditional media communication [13]. However, these studies were usually limited to a small population. The irruption of technologies like Web 2.0 and OSNs, which allow hundred of millions of users interact among them every day, have allowed to extend the information propagation studies to a much larger population. Therefore, in the recent years, we have experienced an increasing proliferation of works that address the information propagation in areas like viral marketing [17, 20], Internet blogs [10] or systems like Arxiv [8]. In addition, we can also find several studies that analyze the propagation of the information in social networks like Flickr [6, 7, 31], Twitter [18, 5, 30], Digg [19] or Facebook [27]. Finally, some works exploit the knowledge extracted from previous studies and propose novel solutions that, for instance, improve the performance of OSNs [25, 24] or define a system to quickly detect natural disasters [23].

Our paper present three main contributions in the area of information propagation in OSNs: (*i*) This is the first effort to characterize information propagation in G+. (*ii*) Previous studies have just analyzed the information propagation in a sample of the OSN. In contrast, our work considers all the information that is publicly propagated in G+. (*iii*) Finally, to the best of our knowledge, this paper is the the most extensive Head to Head comparison in the information propagation between two major OSNs, G+ and Twitter.

6. CONCLUSION

This paper characterizes the propagation of the information in one of the major OSNs, Google+. To the best of our knowledge, our work is the first one that takes into account all publicly available information in an OSN to perform such characterization. In particular, our dataset includes 540M posts (i.e., basic piece of information in G+) and almost 30M propagation trees after filtering the posts that do not receive any reshare. Furthermore, the paper also presents the largest head to head comparison in terms of information propagation between two major OSNs, Twitter and G+. The comparison has revealed that 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 there are two different profiles of influential users in G+. (*i*) *Disseminators* that are mostly represented by relatively popular users that publish many posts that all together reach a large number of people. Most of these users publish entertainment content such as: funny animations, beautiful pictures or quotes including deep emotional messages. (*ii*) *Leaders* that are mainly individual persons that publish few posts, but all of them reach a large audience. Interestingly, we found 7 Google related accounts among the Top 50 *Leaders* in G+. These accounts belong to Google employers and products. This observation suggests a clear interest from G+ users in any information related to Google.

7. REFERENCES

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

[2] E. Almaas, B Kovacs, T Vicsek, ZN Oltvai, and A-L Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843, 2004.

[3] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

[4] Jacqueline Johnson Brown and Peter H Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, pages 350–362, 1987.

[5] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

[6] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18. ACM, 2008.

[7] Meeyoung Cha, Alan Mislove, and Krishna 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*, pages 721–730. ACM, 2009.

[8] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[9] Lujun Fang, Alex Fabrikant, and Kristen LeFevre. Look who i found: understanding the effects of sharing curated friend groups. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 95–104. ACM, 2012.

[10] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[11] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of attribute-augmented social networks: Measurements, modeling, and implications using google+. In *ACM IMC*, 2012.

[12] Roberto Gonzalez, Ruben Cuevas, Reza Motamedi, Reza Rejaie, and Angel Cuevas. Google+ or google-?: dissecting the evolution of the new osn in its first year. In *Proceedings of the 22nd international conference on World Wide Web*, pages 483–494. International World Wide Web Conferences Steering Committee, 2013.

[13] Bradley S Greenberg. Person-to-person communication in the diffusion of news events. *Journalism & Mass Communication Quarterly*, 41(4):489–494, 1964.

[14] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. Enabling collaborative data sharing in google+. In *Global Communications Conference (GLOBECOM), 2012 IEEE*, pages 720–725. IEEE, 2012.

[15] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*, 2008.

[16] Sanjay Kairam, Mike Brzozowski, David Huffaker, and Ed Chi. Talking in circles: selective sharing in google+. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1065–1074. ACM, 2012.

[17] David Kempe, Jon Kleinberg, and Éva 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*, pages 137–146. ACM, 2003.

[18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[19] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010.

[20] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[21] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. New kid on the block: Exploring the google+ social graph. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 159–170. ACM, 2012.

[22] Reza Motamedi, Roberto Gonzalez, Reza Farahbakhsh, Angel Cuevas, Ruben Cuevas, and Reza 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.

[23] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[24] Nishanth Sastry, Eiko Yoneki, and Jon Crowcroft. Buzztraq: predicting geographical access patterns of social cascades using social networks. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pages 39–45. ACM, 2009.

[25] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon 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*, pages 457–466. ACM, 2011.

[26] Doris Schiöberg, Stefan Schmid, Fabian Schneider, Steve Uhlig, Harald Schiöberg, and Anja Feldmann. Tracing the birth of an osn: Social graph and profile analysis in google+. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 265–274. ACM, 2012.

[27] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M Lento. Gesundheit! modeling contagion through facebook news feed. In *ICWSM*, 2009.

[28] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[29] Fernanda Viégas, Martin Wattenberg, Jack Hebert, Geoffrey Borggaard, Alison Cichowlas, Jonathan Feinberg, Jon Orwant, and Christopher Wren. Google+ ripples: a native visualization of information flow. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2013.

[30] Shaozhi Ye and S Felix Wu. Measuring message propagation and social influence on twitter. com. In *Social informatics*, pages 216–231. Springer, 2010.

[31] Bai Yu and Hong Fei. Modeling social cascade in the flickr social network. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, volume 7, pages 566–570. IEEE, 2009.