The World of Connections and Information Flow in Twitter

Meeyoung Cha, Fabrício Benevenuto,
Hamed Haddadi, and Krishna Gummadi

***Abstract*—Information propagation in online social networks like Twitter is unique in that word-of-mouth propagation and traditional media sources coexist. We collect a large amount of data from Twitter to compare the relative roles different types of users play in information flow. Using empirical data on the spread of news about major international headlines as well as minor topics, we investigate the relative roles of three types of information spreaders: 1) *mass media* sources like BBC; 2) *grassroots*, consisting of ordinary users; and 3) *evangelists*, consisting of opinion leaders, politicians, celebrities, and local businesses. Mass media sources play a vital role in reaching the majority of the audience in any major topics. Evangelists, however, introduce both major and minor topics to audiences who are further away from the core of the network and would otherwise be unreachable. Grassroots users are relatively passive in helping spread the news, although they account for the 98% of the network. Our results bring insights into what contributes to rapid information propagation at different levels of topic popularity, which we believe are useful to the designers of social search and recommendation engines.**

***Index Terms*—Computer mediated communication, social network services, twitter.**

I. Introduction

The impressive growth of social networking services has made personal contacts and relationships more visible and quantifiable than ever before. These services have also become important vehicles for news and channels of influence. In particular, Twitter has emerged as a popular medium for discussing noteworthy events that are happening around the world.

Twitter is not a typical social network; its topological characteristics make it more akin to a broadcast medium. Its striking popularity has attracted popular news sources and high-profile users to join the network, including traditional media (e.g., BBC, CNN), celebrities (e.g., Oprah Winfrey), politicians (e.g., Barack Obama), and influentials. Many ordinary users have also joined the network. These different players in Twitter are interconnected through bidirectional social links as well as unidirectional subscriber links, which they use to exchange information. Thus, Twitter presents a unique opportunity to answer longstanding and important social science questions about the interaction among different types of individuals with vastly differently popularity in the diffusion of information [6], [11], [12].

Studying the relative roles different users play on information flow helps us better understand why certain trends or news are adopted more widely than others [8]. Understanding these differences is critical not only for designing better search systems that facilitate the spread of up-and-coming topics while curtailing the storm of spam [1], but it is also a necessary step for viral marketing strategies that can impact stock marketing and political campaigns. Such a study, however, has been difficult because it does not lend itself to readily available quantification; essential components like human connections and information flow cannot be reproduced at a large scale within the confines of the lab.

This paper uses Twitter as a means to conduct research on longstanding social science research questions in a computational framework. We focus on the relative roles different users play on information flow in order to understand why certain trends or news are adopted more widely than others. For the study, we crawled the Twitter network and gathered all public tweets and follow links. In total, we found 2 billion follow relationships among 54 million users who produced a total of 1.7 billion tweets. To the best of our knowledge, this is the largest data gathered and analyzed from the Twitter network.

In order to quantitatively measure the role of users in spreading information, we examined how effective they are as information *spreaders* and measured the size of the audience they could reach in the network. Here, *audience* represents the distinct number of users who either posted or received one or more tweets about a specific event. We develop a computational framework that checks, for any given topic, how necessary and sufficient each user group is in reaching a wide audience.

By analyzing the structure of the connection network and the distribution of links, we found a broad division that yields three distinct user groups based on in-degree: the extremely well-connected users with more than 100 000 followers, the least connected masses with no more than 200 followers, and the remaining well-connected small group of users. Our division of users is based on the definition of different user roles from the theory on information flow [6]: *mass media*, who can reach a large audience, but do not follow others actively; *grassroots*, who are not followed by a large number of users, but have a huge presence in the network; and *evangelists*, who are socially connected and actively take part in information flow like opinion leaders.

For evaluation of our framework, we examine the spread of hundreds of topics, ranging from international headlines that reached up to tens of millions in audience all the way down to topics of local interest that reached only thousands or even smaller audiences. Our analysis reveals several interesting findings about the different roles users play in spreading popular and nonpopular news in Twitter. In the spread of international headlines, grassroots and evangelists accounted for an overwhelming majority of users generating and spreading tweets. However, mass media, despite being only 0.01% of the network, were both necessary and sufficient to reach the majority of Twitter audience. Furthermore, the remaining Twitter audience could be reached by evangelists almost entirely. While the reach of mass media is expected from both traditional theory [6] and anecdotal evidences [8], the reach of evangelists is unexpected and impressive.

Manuscript received March 4, 2011; revised August 1, 2011; accepted November 5, 2011. Date of publication February 22, 2012; date of current version June 13, 2012. Meeyoung Cha was supported under the framework of international cooperation program managed by National Research Foundation of Korea (2011-0030936). Fabrício Benevenuto was supported by the Research Foundation of the state of Minas Gerais (Fapemig). This paper was recommended by Associate Editor L. Fang.

M. Cha is with the Graduate School of Culture Technology, KAIST, Daejeon 305-701, Korea (e-mail: meeyoungcha@kaist.edu).

F. Benevenuto is with the Federal University of Ouro Preto, 35400-000 Ouro Preto-MG, Brazil (e-mail: fabricio@dcc.ufmg.br).

H. Haddadi is with the Queen Mary University of London, E1 4NS London, U.K. (e-mail: hhaddadi@rvc.ac.uk).

K. Gummadi is with the Max Planck Institute for Software Systems, 66123 Saarbruecken, Germany (e-mail: gummadi@mpi-sws.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2012.2183359

1083-4427/$31.00 © 2012 IEEE

The democratization of technology like Twitter is fundamentally changing the way people interact with one another, as well as with local opinion leaders, small businesses, and mass media. While there are many skeptics who doubt that social networks will generate profits that match expectations [4], our study could serve as early evidence for the true "social" opportunity that lies on Twitter in the role of evangelist group reaching out to audience with both popular and long-tail content.

II. Methodology

In this section, we describe how we collected the Twitter data and describe the key topological characteristics of the Twitter network.

A. Measurement Methodology

We asked Twitter administrators to allow us to gather data from their site at scale. They graciously white-listed the IP address range containing 58 of our servers, which allowed us to gather large amounts of data. We used the Twitter API to gather two pieces of information for each Twitter user: 1) profile data including information about the user's social links, i.e., other Twitter users she is following; and 2) all tweets ever posted by the user including the time when tweets were posted.

Twitter assigns each user a numeric ID which uniquely identifies the user's profile. We launched our crawler twice in August 2009 to collect all user IDs ranging from 0 to 80 million. We did not look beyond 80 million, because no single user in the collected data had a link to a user whose ID is greater than that value. Out of 80 million IDs, we found 54 981 152 accounts in use, which were connected to each other by 1 963 263 821 social links. Out of all users, nearly 8% of the accounts were set private, so that only their friends could view their tweets. We ignore these users in our analysis. The social link information is based on the final snapshot of the network topology at the time of crawling since we do not know when the links were formed. In total, we gathered 1 755 925 520 tweets which added up to more than 1 Terabyte of data. Gathering such a large amount of information took more than 1 month using 58 machines. A limitation of this data set is that the social link information is based on the final snapshot of the network topology at the time of crawling and we do not know when the links were formed.

The network of Twitter users comprises a single disproportionately large connected component (containing 94.8% of users), singletons (5%), and smaller components (0.2%). The largest component contained 99% of all links and tweets. Because our goal is to explore how different types of users influence each other in spreading information, we focus on the largest component of the network, which is conceptually a single interaction domain for users.

This data set is perfect for the purpose of our study as it contains near-complete data from Twitter instead of small and potentially biased samples. We have used this data set in a number of recent efforts, such as to study user influence [2], information propagation [10], and spam [11].

B. Twitter Network Characteristics

The Twitter network exhibits a number of characteristics that distinguish it from other social networks. One prominent way we find this is the user degree. Fig. 1 shows the fraction of users in the network with given in- and out-degrees, where a node's *out-degree* refers to the number of users whose tweets the node follows, while a node's *in-degree* refers to the number of users following the node. The two distributions are similar, except for the two anomalous drops in the out-degree distribution around 20 and 2000. The first glitch is due to the "suggested users" feature on Twitter, where all users are presented with a list of 20 popular users to follow upon registration. Unless a user specifies not to follow them, those suggested users are automatically added to the user's out-degree list. The second glitch occurs because Twitter previously limited the total number of individuals a user can follow.

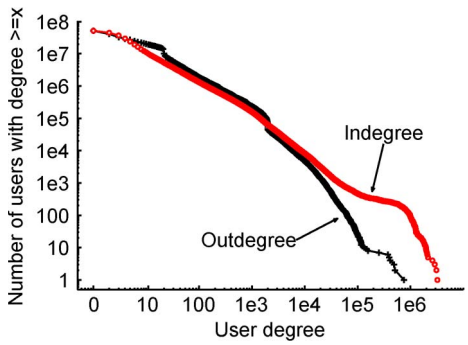

Fig. 1. Degree distribution of the Twitter users.

The distributions for both in- and out-degree are heavy tailed. A majority of the users have small degree, but there are a few users with a large number of neighbors; 99% of users have fewer than 100 in- or out-neighbors. Such a skewed degree distribution indicates that the network contains nodes that connect to a large number of other nodes. In fact, users with extremely large in-degrees exist in the network, at a scale that is unprecedented. The maximum in-degree observed in other online social networks, such as Orkut and Flickr, is limited to a few thousand. On Twitter, in contrast, we find users who have millions of neighbors. Interestingly, the data points beyond 100 000 in the in-degree distribution in Fig. 1 represent users who have many more followers than the overall heavy-tail distribution predicts. Our manual inspection identifies this region as consisting of public figures like Ashton Kutcher and Oprah Winfrey and traditional media sources like BBC.

In summary, the Twitter network exhibits topological features that distinguish it from other social networks; it stands out as a broadcasting system encompassing users of vastly different abilities to propagate and receive information.

III. Types of Users: The Grassroots, Evangelists, and Mass Media

Having established that the number of links per user varies tremendously on Twitter, from hundreds of thousands of followers for popular media sources and public figures to a handful of friends for ordinary users, we first seek to find a robust division of user groups. We present a data-driven approach for categorizing user groups and describe the characteristics of these distinct user groups we find.

A. Categorizing User Types

We wish to find without prior knowledge about the groups of a broad grouping of users that is also meaningful in the context of existing theory. While multiple different ways to categorize users coexist in the theory of information diffusion, our goal is not to explore each of those methods and compare them. Instead, we take a data-driven approach and investigate if the connection pattern entails any generic division of users.

In a broadcast medium like Twitter, the direction of links determines the flow of information. Users with large in-degrees (i.e., having many

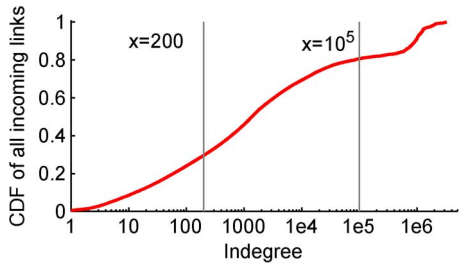

Fig. 2. Distribution of links in the network.

TABLE I
OUT-DEGREE TO IN-DEGREE RATIO PER USER GROUP

Out-degree/In-degree	Grassroots	Evangelists	Mass media
Avg	2.6	1.1	0.5
Med	1.6	0.8	0.001

TABLE II
OBSERVED LINK RECIPROCITY OF THE THREE USER TYPES

Reciprocity	Grassroots	Evangelists	Mass media
Returning	39.7 %	46.4 %	1.84 %
Establishing	16.2%	53.4%	88.6%

followers) can effectively spread information to a large number of nodes [9]. To examine the proportion of the network that could reach a wide audience, we plot the distribution of the links in Fig. 2. The y-axis represents the cumulative fraction of all links pointing to a given user as a function of the in-degree of users on the x-axis, which is calculated by summing the links of all nodes breaking an in-degree of x or less.

The follow links are well spread out across varying in-degrees. As we scan users in increasing order of in-degree and see how many links they receive, the total number of links pointing to the users increases gradually. However, there are points when the addition of links speeds up or slows down. The plot shows the changes in slopes near in-degree of 200–300 and after 100 000. Changes in slopes indicate that the inherent characteristic of the network alter at that point (i.e., phase transition). Nearly 30% of all links are directed toward the least connected users, 20% of links point to the most connected, and 50% of links point to the relatively well-connected users. This implies that, if these points are natural divisions of users, all three groups are important and hold the potential to play a significant role in disseminating information.

Despite all three groups receiving significant amount of links, they have vastly different percentages of users that belong to each group. Over 51 million (98.6%) users were in the least connected group. Nearly 700 000 (1.4%) users belonged to the second group, and 8000 ($< 0.01\%$) users were in the most connected group.

When compared to the two-step flow theory [6], the division seen in Fig. 2 interestingly fits well to the broad category of **mass media**, **grassroots**, and the remaining socially well-connected users whom we will call **evangelists** in this paper. Evangelists are also called by the names of influentials, opinion leaders, hubs, or connectors. Some researchers even subdivide this group into media elite, cultural elite, and experts.

While the division separating mass media from the rest is rather arbitrary, as we chose it to be 100 000 in this work, our goal in this paper is not to find an optimal division point. Rather, our goal is to separate out the effect of the extremely well-connected few from the rest of the network.

The division that separates grassroots from the rest is around 200–300. A similar transition has been widely known as the Dunbar's number [3]. The number of people with whom one can hold personal relationships is limited to about 150 individuals.

B. High-Level Properties of the User Types

Before we understand the roles of the three user groups in information flow, we first studied the properties of the user groups that are important in information flow. Here, we present two such properties. The first property is the ability of a user to receive and aggregate information, seen from the link reciprocity pattern. The second property is the tweet posting pattern.

1) Ability to Aggregate Information: The ability to aggregate information in the network varies according to a user's popularity as shown in Table I, where the out-degree to in-degree ratio is displayed for each user group. The out- to in-degree ratio decreases in general as a user has more followers, indicating that the less popular a user is, the more actively she follows others. In fact, users with in-degree less than 10 follow others 8.56 times more than they are followed! The median trend shows a sharper drop for the mass media group. This is in line with our intuition that the mass media group does not play a role as an aggregator in the network, because they are an authority or institution—like entity that could potentially hire information sources independently outside the network.

The same trend could be seen per group. Table II displays the probability that, upon receiving a follower, a user in a given group will reciprocate (i.e., follow her follower back). Grassroots and evangelists tend to reciprocate most of their followers, but the mass media do not. This matches our intuition that grassroots users and evangelists use Twitter to maintain and develop their social relationships, while most mass media nodes mainly use the network as a broadcast service. Some users in the mass media type do reciprocate—for instance, Barack Obama on Twitter follows more than 600 000 users. In general, however, the mass media type rarely reciprocated links (1.84%).

Table II also displays the probability that a link initiated by a user in one group would be reciprocated (i.e., reciprocity upon link establishment). When a grassroots user forms a link, that link is reciprocated only 16.2% of the time. In contrast, when a mass media user forms a link, a reverse link will exist with a high probability (88.6%)! This trend emphasizes the relative importance of the different types of nodes. Users have higher incentive to reciprocate links from highly connected nodes with larger numbers of followers than those from less connected nodes. The striking difference in the returning and establishing reciprocity strongly indicates that mass media are not social in their link reciprocity and that grassroots and evangelists are more social in their relationship with others.

2) Participation in Information Spreading: While Twitter users lavish disproportionate attention on the small fraction of mass media and evangelists in terms of the number of links and link reciprocity, this does not necessarily mean that mass media and evangelists actively spread news. To see which type of users is the most active in terms of news spreading, we examined the number of tweets posted by each group of users.

From a weekly volume perspective, we observe that mass media and evangelists tweet disproportionately more than other types of users. In average, mass media users posts 78.4 tweets per week, which is nearly twice as many as the evangelists (39.7), and orders of magnitude higher than grassroots (0.5). However, in terms of total volume, over 36% of tweets are sent by evangelists, as opposed to 62% posted by all grassroots and less than 1% by the mass media. It is the high degree connectivity of the mass media and evangelists which makes them so vital in the news spreading chain.

TABLE III
SUMMARY INFORMATION OF THE SIX MAJOR TOPICS EVENTS STUDIED

Topic	Period	Keywords	Description	Spreaders	Tweets	Audience
Iran	Jun 11—Aug 10	#iranelection, politicians' names	2009 Iran election	302,130	1,482,038	22,177,836
Moldova	Apr 6—Jun 5	#pman	Moldova civil unrest in 2009	6,781	46,575	3,103,466
AirFrance	Jun 2—Aug 1	flight 449, airfrance	A plane crash of AirFrance	9,233	12,595	6,317,124
Swine	May 3—July 2	Mexico flu, H1N1, swine	Outbreak of H1N1 influenza	239,329	495,825	20,977,793
Boyle	Apr 21—Jun 20	Susan Boyle	Appearance of Susan Boyle	40,665	62,304	12,788,511
Jackson	Jun 25—Aug 24	Michael Jackson, #mj	The death of Michael Jackson	610,213	1,418,356	23,550,211

IV. RELATIVE ROLES PLAYED IN SPREADING MAJOR TOPICS

In order to quantitatively measure the role of the three user groups in spreading information, we examined how effective they are as information spreaders in the network and measure the size of the audience they could reach in the network. We develop a computational framework that checks, for any given topic, how necessary and sufficient each user group is in reaching a wide audience.

A. Finding Tweets Related to Major Headlines

We picked six major events that occurred in 2009 that were widely reported to have been covered by Twitter.[1] These six events, summarized in Table III, span political, health, and social topics. To extract tweets relevant to the six events, we first identified the set of keywords describing the topics by consulting news websites and informed individuals. Given our selected list of keywords, we identified the topics by searching for keywords in the tweet data set. We focused on a period of 60 days starting from one day prior to a key date; this either corresponds to the date when the event occurred or the date when the event was widely reported in the traditional mass media (TV and news papers). We limited the duration because popular keywords were typically hijacked by spammers after certain time.

Table III also displays the keywords and the total number of users and tweets for each topic. We refer to users who generated one or more tweets about an event as *spreaders* and users who either posted or received one or more tweets about an event as the *audience*. The number of spreaders varies tremendously across the different events. The most popular event has two orders of magnitude (100 times) more spreaders than the least popular event. Interestingly, however, the audience sizes show much less variation across the events. The most popular event reached an audience of 23 million, which is only a factor of 8 larger than the least popular event, which reached an audience of 3 million. All events reached an audience of several million, which is a significant fraction of all twitter users. Thus, it is possible to reach a large fraction of all network users even with a small number (few thousand) of spreaders.

B. Relative Roles of Each User Group

Below, we first use the examples of six major events to study the presence of the three different types of users in information flow. We then introduce the framework for checking the relative roles of users and highlight our findings on the reach of audience.

1) Grassroots and Evangelists Account for Most Spreaders: We first examine the presence of the three types of users across the events. To do that, we compute the fraction of grassroots, evangelists, and mass media among those users who tweeted about the event. Mass media uniformly account for only a small fraction of all spreaders across the different news events. Compared to the overall presence of mass media (0.01%) in Twitter, however, their presence on major news events is two to six times higher. While grassroots account for a majority of spreaders (more than half in all cases), it is the evangelists who account for a considerable fraction of all spreaders. This is surprising given that evangelists account for 1.4% of all Twitter population. The AirFrance event involved the highest fraction of evangelists, while the Jackson and Swine events showed the highest involvement of grassroots users.

2) Grassroots and Evangelists Account for Most Tweets: Focusing on the fraction of tweets produced by each user type, we find that grassroots in most cases dominate the number of tweets. This is partly due to the large number of users in this group and also partly due to the presence of spammers, advertisers, and accounts dedicated to particular topics in this category. Once again, we notice that the evangelists post more than their fair share of tweets. In fact, they account for the majority of tweets related to some news events like AirFrance, even when they are not the majority of all spreaders. This shows the eagerness of this group to voice their opinions, influence others, and ultimately attract followers and media attention. Mass media, in contrast, account for only a small fraction of all tweets. This is largely due to the fact that with such large audiences, the mass media are carefully followed and the quality and frequency of messages are essential in raising their public profiles.

In addition to the sheer number of tweets, another key metric for measuring the importance of a given user group is the audience size. We next investigate what fraction of Twitter users each of these types can reach.

3) Mass Media is Necessary and Sufficient to Reach a Majority of Twitter Audience: To infer the significance of each group in reaching the audience, we devised the following test. We first sort the spreaders based on their in-degree and examine what fraction of the audience top spreaders can reach. The cumulative size of the audience would represent whether top spreaders alone can reach such audience (i.e., sufficiency). Next, out of all users in the audience of an event, we gradually remove the top spreaders and examine what fraction of the audience can still be reached. This second test lets us infer the necessity of the top spreaders in reaching such a wide audience.

Fig. 3 displays the necessary and sufficient conditions for reaching the audience for the six events. On each figure, the red line indicates how many spreaders are needed to reach the given audience size. The black line indicates the size of the audience that could be reached after removing the top-k spreaders, where k is varied from 0 to the total number of spreaders. In case two or more users have the same in-degree, we broke the tie based on the numeric user IDs so that the rank of every user is different. The x-axis represents the rank of the spreader based on in-degree, from the most followed (appearing on the left-hand side) to the least followed (appearing on the right hand side).

The two vertical lines in the figure mark the boundaries between mass media (denoted by M), evangelists (E), and grassroots (G) regions. In all cases, we observe that the mass media presence is sufficient to reach a significant majority, but not the entire audience. Apart from the Moldova case where the mass media did not play a key role, more than 70% of users can usually be reached by the mass media.

Not only are mass media sufficient to reach a significant fraction of the audience, they are necessary to reach an audience of such a size. Focusing now on the black line, we see that the size of the audience that is reached decreases rapidly as top spreaders are removed. Without

[1]Top Twitter trends http://tinyurl.com/yb4965e.

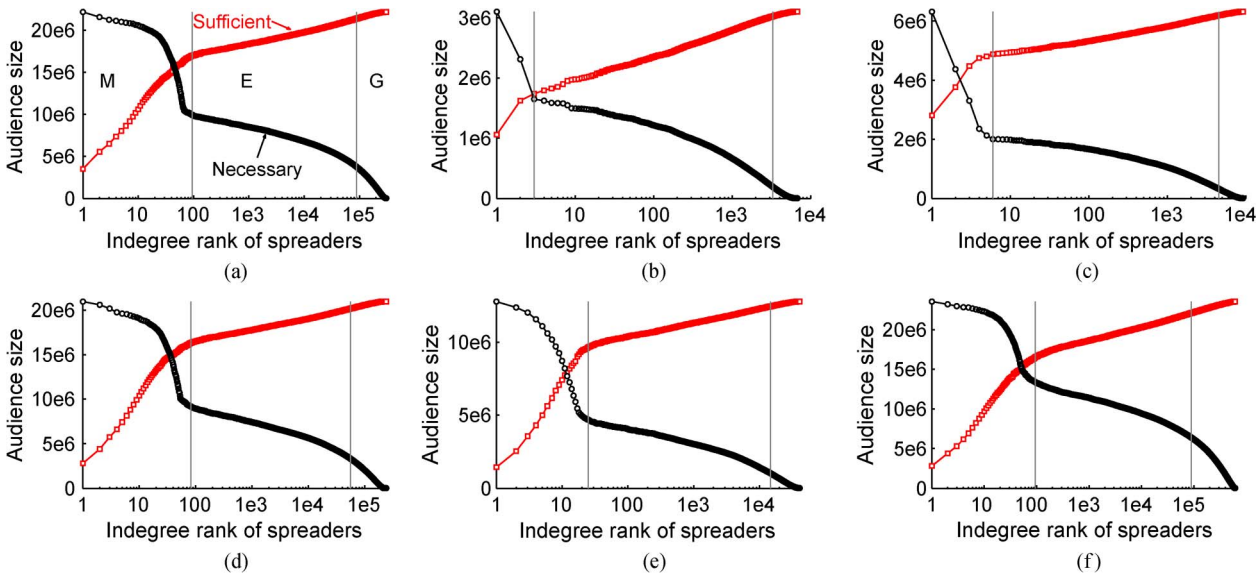

Fig. 3. Test of sufficiency and necessity conditions in reaching an audience of a given size for international headlines. (a) Iran; (b) Moldova; (c) AirFrance; (d) Swine; (e) Boyle; (f) Jackson.

mass media, we lose the majority of the audience in all cases. Due to their high in-degree, the mass media are able to directly cover a large fraction of the audience, even posting the fewest number of tweets. This effect is less severe in the Moldova, Jackson, and AirFrance cases, where about 50% of the nodes are still reachable even after removing the mass media. In the Moldova case, however, it is noticeable that only two nodes are enough to reach over half the audience population on Twitter. The two mass media users were a technology news website (TechCrunch) and an Internet analyst.

4) Evangelists Extend the Reach of Mass Media Considerably: Deviating our focus from mass media, here we investigate the importance of evangelists. On all graphs in Fig. 3, we observe a sharp drop in the sufficiency line after the mass media hand over to the evangelists. In most cases, however, at least an additional 25% of the audience can be reached through evangelists. This is a considerable fraction, particularly considering that evangelists make up only a small fraction of total users (1.4%) in Twitter.

This finding highlights the importance of evangelists as information spreaders. The test of necessary condition further shows that when evangelists are removed from the network, only a small fraction of the audience can be reached by using grassroots only. This test result indicates that evangelists can extend the reach of audience by a considerable amount. We also notice that in a deeply involved and political event, such as the Moldova case, the evangelists are playing a critical role by covering around half the user population.

5) Grassroots Help Reach a Nonsignificant Fraction of the Audience: As opposed to the dominant reach of mass media and evangelists, grassroots reach out to only a small fraction of the audience. The sufficiency line on Fig. 3 shows that, despite the hype about grassroots-based spreading of information, grassroots account for a negligible fraction of all audiences across the different news events. This occurs despite the fact that grassroots account for nearly all Twitter users (98.6%) and a significant fraction of all tweets.

However, this fraction is not always negligible and is dependent on the type of news in the context. Their role is more important in cases where the mass media have less presence, such as the Michael Jackson case, as seen in Fig. 3(f). This is also the case with gossip-like events such as Swine Flu, Iran Elections, and Michael Jackson events. In contrast, when the mass media have a major role, such as in the AirFrance case, the grassroots play a much less significant role.

V. Relative Roles in the Spread of Less-Popular or Niche Topics

Having studied the role of the three user groups in spreading major news events, we now examine the reach of audience in the spread of less popular topics.

A. Finding Tweets Related to Minor Topics

In order to have a reasonable amount of selection and diversity in the set of less popular topics, we utilized hashtags (i.e., keywords that start with the "#" sign) and considered them as a clean piece of information that spread in the network. We examined the popularity of hashtags based on the number of tweets containing each hashtag and chose samples from different popularity levels. In doing so, we started with all tweets posted in the last 4-month period between May and August in 2009, which contained 1.1 billion tweets or nearly half of all tweets. There were 3 132 605 distinct hashtags contained in this set. The popularity of these hashtags followed a power-law distribution (also shown in the technical report).

Out of the 3 million hashtags, we selected a total of 100 samples at random from different scales of popularity. The sampled topics reached orders of magnitude smaller audience than the six major events we studied. While the major news reached audiences of up to tens of millions, the 100 minor topics reached audience of several millions down to fewer than 10. The content of these minor topics ranged from information that is of local interests such as regional carnivals, chatter from the Disney fan club, and Google's Lunar X Prize contest. We also observed several self-made terms such as "endangeredanimal" in the set of niche hashtags that reached audiences of up to 100.

Overall, only 8 out of 100 topics had one or more mass media spreaders involved. The remaining 92 topics were only spread by evangelists or grassroots users. Furthermore, 19 topics were solely spread by the grassroots users.

B. Reach of Audience by Each User Group

We repeated the reachability analysis and investigated the relative roles different user groups played in spreading the 100 topics. While the necessary and sufficient size of audience obtained by each group

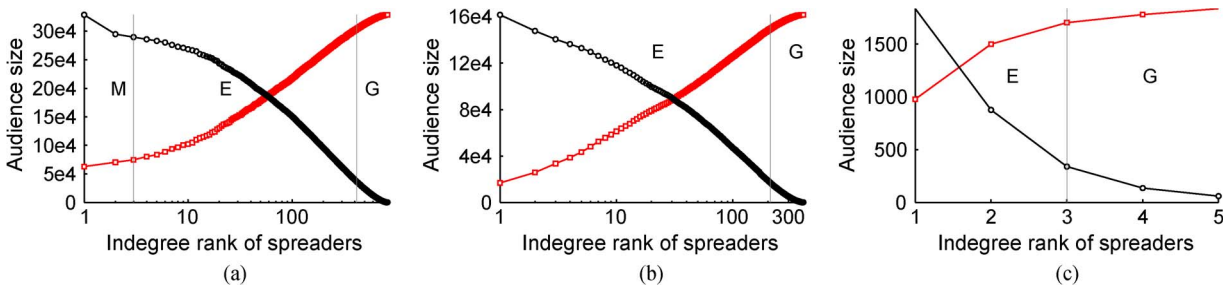

Fig. 4. Test of sufficiency and necessity conditions in reaching an audience of a given size for less popular topics. (a) Hashtag "#disneyland"; (b) Hashtag "#googleapps"; (c) Hashtag "#healthiest."

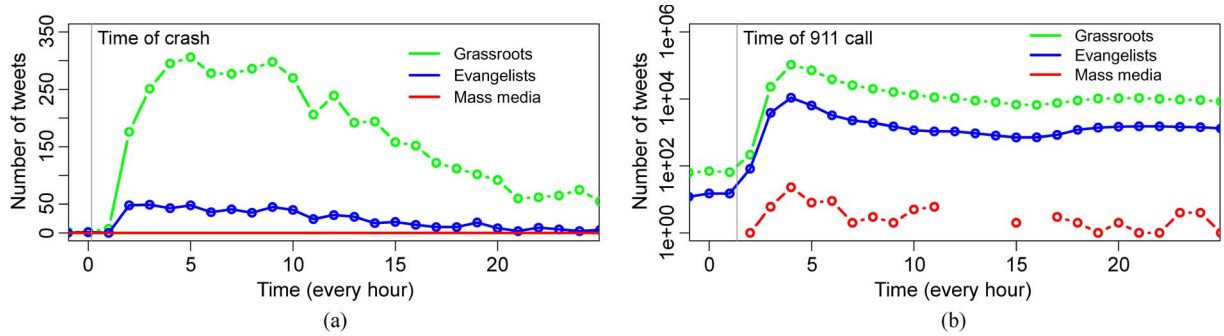

Fig. 5. Traffic volume per hour for the three user types. (a) AirFrance; (b) Jackson.

varied per topic, we saw common patterns of influence spreading. We picked three representative examples to summarize these patterns.

Fig. 4(a) shows the case of the hashtag "#disneyland" that reached an audience of hundreds of thousands. In the spread of this hashtag, two mass media, 327 evangelists, and 497 grassroots users were involved. The two mass media users were sufficient to reach 20% of the audience (i.e., red line) and necessary to reach 10% of the audience (i.e., black line). Evangelists, then, were in charge of reaching the majority of the audience. It is impressive to observe that even when the topic is no longer a top headline, grassroots do not play a major role in reaching out to the audience.

Fig. 4(b) and (c) shows the representative examples of when no single mass media spreader participated in the spread of a hashtag. Only evangelists and grassroots users participated in the conversation for these topics. In the case of hashtags "#googleapps," a similar number of evangelists and grassroots users (150 each) participated in the conversation. We again find that the role of evangelists group in reaching the audience is predominant. Over 90% of the audience was both sufficiently and necessarily reached by evangelists! Finally, the hashtag "#healthiest" only had five distinct spreaders among whom two were evangelists and three were grassroots. While the topic only spread to an audience of about 1700, the two evangelists still played a major role in reaching the majority of the audience.

One consistent observation we could make from the manual inspection of individual topic spreading was that a significant fraction of the audience were reached by popular users like evangelists and mass media, rather than by grassroots. Given that our samples include topics of medium to niche popularity, this finding was surprising to us.

VI. Dynamics in Information Flow

In the previous section, mass media and evangelists, despite being a small fraction of all users, turned out to be crucial in reaching a large audience. What we have ignored in the previous section is the exact times when different types of users engage in the spread of news events. The sufficiency and necessity conditions assume two extreme temporal orders: when the spreading takes place from the highest degree node to the lowest degree node and vice versa. In both cases, we have confirmed the significant role of mass media and evangelists.

Based on our understanding of the relative roles that different types of users play, in this section, we consider the exact times of topic adoption by different user types and analyze how these users interact in the actual spreading of messages. Given the high presence of influential users within the news event network, we are curious to see if the flow of information in Twitter obeys the traditional top-to-bottom broadcast pattern, like the two-step flow of communication hypothesis [5] and those emphasizing the role of influentials [8]. However, as we will demonstrate, the information flow in Twitter does not follow the traditional top-to-bottom broadcast pattern where news content usually spreads from mass media down to grassroots users.

A. Temporal Dynamics in User Participation

As a start, we examined the exact times at which the three user types generated tweet posts. Fig. 5 shows the traffic volume in terms of the number of tweets per hours by the three user types on AirFrance and Jackson, for the first 24 h of each event. Unlike the other news events, these two news events had urgent breakouts in their nature (i.e., crash of an airplane and sudden death of a celebrity figure). Upon the breakout of these events, these news topics spread like wildfire within Twitter, in particular among evangelists and grassroots users. Within a few hours, the two news topics generated hundreds of tweets to tens of thousands of tweets per hour alone by the grassroots group.

The figures interestingly demonstrate that mass media presence is not always immediate in the spread of urgent news. Mass media was almost nonpresent at the beginning of the AirFrance event.[2] In the case of Jackson, mass media sources were silent for several hours in producing news on the topic, while evangelists and grassroots users continued to show interest on the topic.

[2]Note that while popular news media outlets covered headlines on AirFrance crash, the news on them on Twitter appeared only later in time.

TABLE IV
DESCRIPTION OF THE TOP URLS

Event	URL Description	Audience
Iran	Promoting to use a green avatar	12,832,998
Moldova	Article on Twitter revolution	735,223
AirFrance	Article on hoax AirFrance crash images	11,111,145
Swine	A comic website about swine flu	1,685,277
Boyle	YouTube video of the singer	626,042
Jackson	Petition to nominate MJ for Nobel prize	125,346

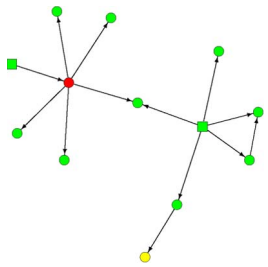

Fig. 6. Spreading example from Moldova. Colors denote different user types: red (mass media), green (evangelists), and yellow (grassroots).

In contrast, traditional broadcast models [5], [7] describe that information spreads in two separate paths. The first step is from mass media to a small set of *opinion leaders* or *influentials* (similar to the evangelists in our definition), and the second step is from opinion leaders to the masses (or grassroots users). Unfortunately, it is hard to test these theories and decide on the direction of the exact information flow from our data, because tweets on the same news event could include multiple threads of information flows. In order to investigate whether news topics spread from mass media sources to evangelists and then to grassroots users, we next focus on the flow of individual pieces of information.

B. Interaction Among Different User Types

As a unit of information that spread in Twitter, we chose to look at the spread of URLs. For each of the six news events, we selected the most popular URL from our data set and examined how these six URLs were exchanged among Twitter users. Again, we call users who posted the URL as spreaders and users who posted or received at least one message containing the URL as audience. We then examined the topological structure of the audience network to understand the flow of information among different user types. To avoid picking spam URLs, we ranked URLs based on the size of the total audience it reached, rather than the number of times it was tweeted.

Table IV displays information about the final six URLs on each news event, along with the description and their audience size. Utilizing the information about the direction of Twitter follow links and timing of tweet posts, we could determine from whom each audience member received information about the same URL. For each audience member, we call the user who potentially delivered information as *sources*. Given two Twitter users A and B, we say information flew from A to B if and only if 1) both users posted the same URL, 2) B is following A, and 3) A posted the URL prior to B. In this case, we say A is the source of B. In case there are multiple candidate sources, we pick the user who most recently posted the same URL as the source.

An example propagation pattern is shown in Fig. 6, which shows the largest connected component of the Moldova spreading network. We only show the spreaders in this graph. A total of 13 users collaborated in sharing the URL, where two users (indicated in squares) have independently shared the URLs (i.e., these users do not have any source), while all other users had a source. The color of the nodes represents different user types: red nodes are mass media, green nodes are evangelists, and yellow nodes are grassroots. Interestingly, in this example, mass media user (colored in red) receives URL from an evangelist user, then spreads it to many other grassroots users.

TABLE V
SOURCE AND AUDIENCE TYPES

Audience type	Source type		
	Grassroots	Evangelists	Mass media
Grassroots	0.42%	24.86%	74.72%
Evangelists	2.43%	71.9%	25.67%
Mass media	5.91%	89.93%	4.16%

Table V shows the pattern of information flow among different user types. The results are aggregated over the six URLs. As expected, grassroots users receive most of their URL links from mass media and then evangelists. However, this natural order is not homogeneous across other groups. Both evangelists and mass media receive a large majority of their news from evangelists. In some cases, they also receive information from grassroots users. Mass media outlets in fact receive URLs at similar fractions from grassroots as well as from other mass media outlets.

Our findings indicate that, unlike the traditional models of communication, the flow of information is not always directed from high-degree nodes to low-degree nodes. The new role of grassroots—influencing and spreading information to much higher degree nodes—was unimaginable from traditional theories [5], [7]. In Twitter, however, the direction of follow links is not restricted by node degree, allowing information to flow in any direction. Grassroots users can also trigger a large flow of information, by getting the message to their neighboring evangelists and mass media nodes.

VII. CONCLUSION

We have presented, to the best of our knowledge, the first extensive analysis of a near complete data set obtained from the microblogging service Twitter. Acquisition of such a rich data set enabled us to identify the relationship among distinct groups of users—mass media, evangelists, and grassroots—and the roles that they play in viral spreading of political and social news messages. The connectivity trends between users differentiate Twitter, away from conventional social networks, toward a collaborative gossip and news publishing tool. This makes Twitter an ideal medium for studying the relative roles these distinct user groups play. Our analyses show that Twitter network exhibits topological features that distinguish it from other social networks; it stands out as a broadcasting system encompassing users of vastly different abilities to propagate and receive information.

We found that Twitter brings a playing field together for all three voices: the mass media, evangelists, and grassroots. On one hand, the mass media play a dominant role in the network. They excel at all aspects of news spreading; they have many followers, their links are well reciprocated, and they have topological advantages to collect diverse opinion of other users. Their tweets also reach a large portion of the audience directly, without the involvement of other influential users. On the other hand, the mass media in Twitter, unlike the traditional media networks, are not necessarily the first to report events. In some cases, in fact, it is the small, less connected grassroots or evangelists that trigger the spreading of news or gossip, even without the mass media's coverage of such topic. Evangelists, overall, played a leading role in the spread of news in terms of the contribution of the number of messages and in bridging grassroots who otherwise are not connected.

REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Proc. Annu. CEAS*, Redmond, WA, 2010.

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. AAAI ICWSM*, Washington, DC, 2010.

[3] R. I. M. Dunbar, "Coevolution of neocortical size, group size and language in humans," *Behav. Brain Sci.*, vol. 16, no. 4, pp. 681–735, 1993.

[4] A Special Report on Social Networking, The Economist, Jan. 2010, Accessed in Jul. 2011. [Online]. Available: http://tinyurl.com/ylxtsek

[5] E. Katz, "The two-step flow of communication: An up-to-date report on a hypothesis," *Public Opinion Quart.*, vol. 21, no. 1, pp. 61–78, 1957.

[6] E. Katz and P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*. New York: Free Press, 1955.

[7] E. Katz and P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*. New Brunswick, NJ: Trans. Publ., 1955.

[8] P. Lazarsfeld, B. Berelson, and H. Gaudet, *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Duell, Sloan, and Pearce, 1944.

[9] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001.

[10] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the web," in *Proc. ACM SIGCOMM IMC*, 2009, pp. 381–393.

[11] E. M. Rogers, *Diffusion of Innovations*. New York: Free Press, 1962.

[12] D. Watts and P. Dodds, "Influentials, networks, and public opinion formation," *J. Consum. Res.*, vol. 34, no. 4, pp. 441–458, Dec. 2007.