Information Resonance on Twitter: Watching Iran

Zicong Zhou
Electrical Engineering, UCLA
zzhou@ee.ucla.edu

Roja Bandari
Electrical Engineering, UCLA
roja@ucla.edu

Joseph Kong *
Electrical Engineering, UCLA
Northrop Grumman
jskong@ee.ucla.edu

Hai Qian
Electrical Engineering, UCLA
haiqian@ee.ucla.edu

Vwani Roychowdhury
Electrical Engineering, UCLA
vwani@ee.ucla.edu

ABSTRACT

Twitter has undoubtedly caught the attention of both the general public, and academia as a microblogging service worthy of study and attention. Twitter has several features that sets it apart from other social media/networking sites, including its 140 character limit on each user's message (tweet), and the unique combination of avenues via which information is shared: directed social network of friends and followers, where messages posted by a user is broadcast to all its followers, and the public timeline, which provides real time access to posts or tweets on specific topics for everyone. While the character limit plays a role in shaping the type of messages that are posted and shared, the dual mode of sharing information (public vs posts to one's followers) provides multiple pathways in which a posting can propagate through the user landscape via forwarding or *"Retweets"*, leading us to ask the following questions: How does a message resonate and spread widely among the users on Twitter, and are the resulting cascade dynamics different due to the unique features of Twitter? What role does content of a message play in its popularity? Realizing that tweet content would play a major role in the information propagation dynamics (as borne out by the empirical results reported in this paper), we focused on patterns of information propagation on Twitter by observing the sharing and reposting of messages around a specific topic, i.e. the Iranian election.

We know that during the 2009 post-election protests in Iran, Twitter and its large community of users played an important role in disseminating news, images, and videos worldwide and in documenting the events. We collected tweets of more than 20 million publicly accessible users on Twitter and analyzed over three million tweets related to the Iranian election posted by around 500K users during June and July of 2009. Our results provide several key insights into the dynamics of information propagation that are special to Twitter. For example, the tweet cascade size distribution is a power-law with exponent of -2.51 and more than 99% of the cascades have depth less than 3. The exponent is different from what one expects from a branching process (usually used to model information cascades) and so is the shallow depth, implying that the dynamics underlying the cascades are potentially different on Twitter. Similarly, we are able to show that while Twitter's Friends-Followers network structure plays an important role in information propagation through retweets (re-posting of another user's message), the search bar and trending topics on Twitter's front page offer other significant avenues for the spread of information outside the explicit Friends-Followers network. We found that at most 63.7% of all retweets in this case were reposts of someone the user was following directly. We also found that at least 7% of retweets are from the public posts, and potentially more than 30% of retweets are from the public timeline. In the end, we examined the context and content of the kinds of information that gained the attention of users and spread widely on Twitter. Our data indicates that the retweet probabilities are highly content dependent.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Science

General Terms

Experimentation, Measurement

Keywords

Twitter, Social Media, Information Propagation, Cascades, Iranian Election

1. INTRODUCTION

On June 12th 2009, Iran held its presidential election between incumbent Mahmoud Ahmadinejad and three other candidates, including a popular challenger named Mir Hossein Mousavi. The result, announced as a landslide for Ahmadinejad, led to charges of election rigging, and massive protests across Iran. With international news reporters purged from the country shortly after the election, Iranian citizen journalism became the only means of documenting the events and Twitter became a window for the world to

*The work was entirely performed when J. Kong was at UCLA.

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

witness the mass protest movement and its violent crackdown by the authorities.

Twitter is a microblogging service that allows each user to post tweets of a maximum 140 characters on their profile page. Each user can then follow a collection of other users of her or his choice in order to view their tweets aggregated in a home page. We will call those who follow a user, his or her followers, and we will call those whom the user follows, his or her friends. Since following someone's tweets does not automatically mean that they will follow you back, Twitter's Friends-Followers network is a directed graph.

Some conventions, without being required by Twitter, have been widely adopted by users. Using the "#" sign to tag a post according to its content (called a hashtag) is one such convention used in many tweets. The hashtag can be used as a search keyword to access a public listing (called the public timeline) of all the tweets that use that specific hashtag. When a keyword becomes very popular at any point in time, it appears as a trending topic on all users' home pages and on the twitter front page, giving all users direct access to all the tweets on that topic. Another convention that became a widely used standard (and recently implemented in the service as a proper function) was retweeting, where user2 would repeat user 1's tweet almost exactly, and adding "RT @user1" at the beginning of the tweet to give credit. For a more detailed guide to Twitter, please see [21].

We observe and analyze the dynamics of information propagation through the study of tweets about the Iranian election. Since different content can create different dynamics of information propagation, we focus our study on this very specific, yet large set of data. We study the characteristics of the Friends-Followers network (which we shall abbreviate the F-F network) in our dataset. We then visualize how information resonates and spreads widely among a large number of users, and study the mechanisms of information propagation both inside and outside the underlying F-F network.

Section 3 describes our methodology of data collection, and cleaning and filtering the data to collect tweets related to Iran's election. In section 4 we perform some basic analysis of the directed social network, deriving the distribution of tweets and retweets, and the relation between number of tweets and number of followers. Section 5 focuses on visualization of information propagation as retweet cascades and study of their shapes. We observe that cascades tend to be wide rather than deep, commonly with a central hub and that a message can reach a large audience on the network even when the number of users who retweet a message is not very large. In Section 6 we study the mechanism of information propagation on Twitter and different ways that information is found and shared on this network. Due to a design that facilitates access to real-time tweets of all users with public profiles (**the public timeline**) and allows anyone to search this timeline for a keyword, Twitter users have access to tweets that are not posted by their direct friends in the F-F network. We study the ways through which information is found and retweeted through the F-F network as well as through the public timeline. We find that 63.7% of tweets are propagated through direct links in the F-F network, confirming that the underlying social network plays an important role in information propagation. We then study the propagation of information outside the F-F network links and find that a notable number of retweets are by users retweeting someone whom they are not directly following. Larger cascades included more retweets of this nature. We define retweet rate for different cascades and using this metric, we show that the content of information plays a key role in determining the popularity of tweets. Furthermore, we find that the retweet rate decays exponentially as the cascade spreads away from the source. Therefore it might not be possible to ignore the modeling of content-based and depth-based properties when it comes to studying information propagation of socially significant information. Finally, in Section 7 we present a brief taxonomy of cascade content and the users who posted the tweets leading to large cascades.

2. RELATED WORK

Online social network systems have emerged recently as the most popular forums for user participation, social intercourse, and content generation. Research work conducted on modeling and analyzing various aspects of social networks have identified many recurring patterns, such as power law degree distributions, small world, local clustering and communities structures [4, 9, 12, 19, 20], in the underlying friendship or contact networks. Moreover, microscopic network evolution models have been proposed [2, 15].

One of the distinguishing features of online social networks and social media is their potential for information propagation. It has been studied both empirically and theoretically for many years by sociologists concerned with diffusion of innovation [22]. Watts [23] theoretically analyzes cascades on random graphs using a threshold model. Wu et al. [24] present an epidemic model to study global properties of the spread of email messages. Leskovec et al. [18] empirically analyze the topological patterns of cascades in the context of a large product recommendation network and study efficacy of viral product recommendation strategies[14]. Leskovec et al. examine information propagation structure [17] on blogosphere and propose algorithms for identifying influential nodes [16]. Bakshy et al. [3] trace the spread of influence in a multi-player online games and found patterns similar to our findings with social news dynamics on Twitter. However, in these previous studies, the underlying network is defined by message passing among the users and agents, i.e., an edge connects two nodes, A and B if the message or link posted by A is copied or published by node B, and thus the cascades studied are analogous to the Tweet Networks studied in this paper. On Twitter, however, we have visibility of the Friends-Followers (F-F) network as well, and it provides us with a unique opportunity to study the role played by the F-F network vs the role played by the public timeline (or the posted messages) and the links to trending topics accessible to all viewers. For example, it provides us with an opportunity of finding what types of content led to cascades of significant size, and that the tweet infection rate is a function of the content type and hence a notion of fitness has to be introduced.

Twitter has attracted much attention from researchers since it became an important social network as well as social media. Java et al. [11] study the topological and geographical properties of Twitter's social network, and show how users with similar intentions connect with each other. Huberman et al. [10] point out that the use of @user is a form of conversation, which indicates the hidden network of connections underlying the "declared" set of friends and

followers. Boyd et al. [6] present various conventions and styles of retweeting prevalent today and examine the emergence of retweeting as a conversational practice. Kwak et al. [13] crawl the entire Twittersphere to study its topological characteristics and retweet trees between different users. However, we find that the cascades and information mechanisms for tweets are highly topic and content dependent, and hence, we chose to study a particular event that comprises a medium size network, and provides a window into various subtler aspects of information propagation on Twitter. For example it allows us to study the role played by the public timeline vs the F-F network in propagating information.

3. MEASUREMENT METHODOLOGY

We used the Twitter API [1] to crawl the social network and download a large number of public user pages on Twitter. Since our goal here is to study the topological characteristics of information propagation regarding the Iranian Election, our data sampling process is highly biased toward users who have tweeted about this topic.

3.1 Data Collection

We began with a list of 100 most active users on the topic of Iranian Election as reported by the Web Ecology Project [5]. Using these users as seeds, we traversed their directed F-F network (friends and followers) and reached about 126K valid users who were one step away from the seed users, which we will call depth-1 users. We continued to traverse the F-F network of these depth-1 users, and this gave us 23 million distinct depth-2 users. We then crawled the F-F network of these 23 million users and finally collected about 20 million users' F-F network (Some of the target users were invalid or had protected profile, so we were not able to download their F-F network).

Since Twitter API only allows access to a maximum of 3,200 tweets per user, we collected as many tweets for these users as the API could provide. In total we collected the tweets as well as the F-F network for about 20 million users.

3.2 Coverage Estimation

We did not cover the entire connected component of Twitter but we had a qualitative coverage examination of our crawl. Since the IDs of user on Twitter are assigned sequentially, we uniformly selected 200K random IDs between the first ID and the last one. Among the IDs we tried to collected, there are 130K (65.0%) users with public profiles, 13K (6.5%) users with protected profiles and the remaining 57K (28.5%) IDs were invalid for different reasons. Based on these statistics, there should be around 55M valid users on Twitter by the end of September 2009 as maximum ID was 77M by then. Among the 130K users we downloaded, there were 1738 users who tweeted about the Iranian Election 1558 of which were included in our crawled dataset; there were 11108 tweets related to the Iranian Election and our dataset covers 10760 of them. Therefore, it appears that our dataset covers 89.6% of users and 96.9% of tweets relevant to the subject of Iranaŕs election on Twitter.

3.3 Data Cleaning

Before the analysis, we applied the following procedures to clean the data in order to better represent the structures of information propagation.

Only consider the tweets that have related keywords. We used most widely used keywords related to the Iranian Election [5] to filter the tweets first. As a result, we focused on a total of more than 3 million tweets posted by 500K users between June 1 2009 and August 1 2009.

Only consider the RT tag. In this paper we study information propagation as retweeting and only restrict the tweets that have form of 'RT @user'. On Twitter users may get similar news or messages from different sources, and it is possible for them to come up with the similar tweets without reference each other, which is not regarded as information propagation in our case.

Remove self retweet. Users sometimes retweet themselves in order to emphasize their message or increase the number of people who view their tweet, but self-retweets do not represent any information propagation.

Table 1: Iranian Election users' network statistics

Property	Statistics
Number of Nodes	470040
Average In-degree/Out-degree	87.10
In-degree Distribution α	-2.85
Out-degree Distribution α	-2.42
In-degree Distribution D	0.0167
Out-degree Distribution D	0.0087
Correlation of in-degree and out-degree	0.6936
Reciprocity	0.4813
Clustering Coefficient	0.1052
Assortativity	-0.2633

3.4 Link Inference for Tweet Networks

Although a retweet explicitly mentions the user who posted the original tweet (RT @user), there is no mention of or link to the specific tweet that is being retweeted. In order to build retweet cascades we need to find links between a tweet and its retweet. So when a retweet mentions a certain user (RT @user) we must search that user's messages and find the tweet that has similar textual content with the retweet. On blogsphere, text analysis technique is proposed to infer relationship among posts [1]. In this paper, we adopt the digests technique [8] to determine if two messages contain the same textual content.

4. FRIENDSHIP-FOLLOWER NETWORK

4.1 Network Structure

The F-F network we consider in this paper is clearly a subset of the complete Twitter F-F network. Our dataset contains tweets of 470,040 active users who posted at least one tweet about the Iranian Election between June 1 2009 and August 1 2009 and 40,938,802 edges between them. Figure 1(a) and Figure 1(b) show the in-degree and out-degree distributions of this network, both following a power law distribution. To test how well the degree distributions are modeled by a power-law, we calculated the best power-law fit using maximum likelihood [7]. Table 1 shows the estimated power-law coefficients, the corresponding Kolmogorov-Smirnov goodness-of-fit metrics D (K-S metrics D) [7] as well as other properties of network we studied.

[1] http://apiwiki.Twitter.com/

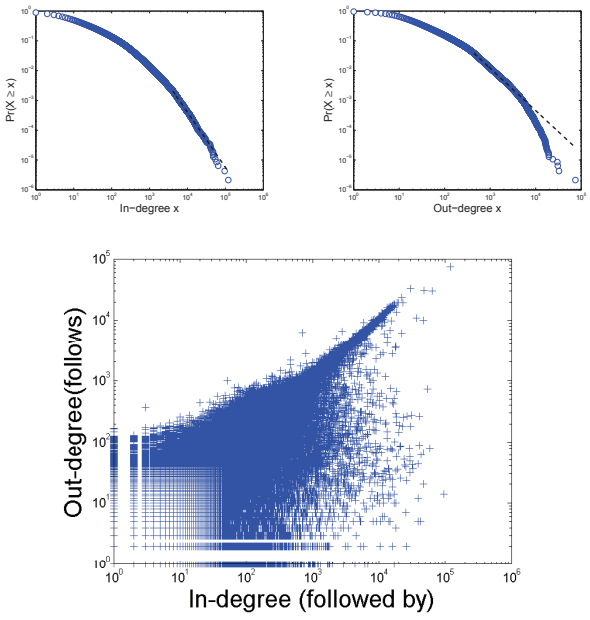


Figure 1: Cumulative distribution of Iranian Election users' in-degree and out-degree. And high degree correlation on scatter plot implies that there are a large number of mutual connections in Iranian Election users' F-F network

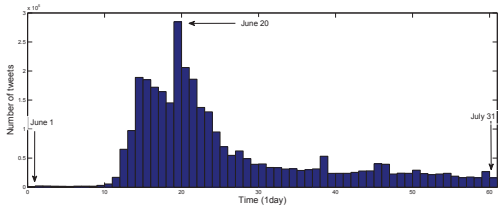


Figure 2: Number of tweets by day from June 1 2009 to Aug 1 2009. The rate gradually increased as the events unfolded in Iran and the use of Twitter provoked attention, spiking dramatically in relation to political events inside Iran as well as in relation to new events and incidents particular to the web.

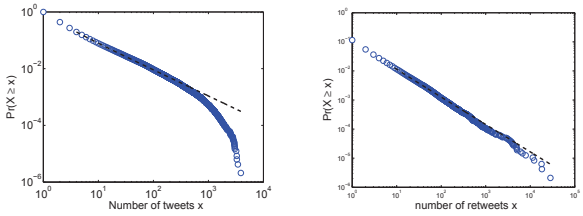


Figure 3: Cumulative distribution of number of tweets and retweets per user. Power law fit to the data with exponents -1.92 and -1.94.

The scatter plot in Figure 1(c), as well as the correlation of in-degree and out-degree, and the high reciprocity of network imply that there are a large number of mutual connections in this F-F network, which is not the case for all users on Twitter as reported in [13].

Clustering coefficient is an indication of how densely neighbors are connected. The high clustering coefficient in Table 1 suggests the presence of strong local clustering , meaning in our dataset users tend to know each other via mutual friends. Assortativity, a measure of the likelihood for nodes to connect to others with similar degrees, has been shown to be positive in social networks [20]. However, the F-F network of our dataset has a negative assortativity, which means nodes are likely to connect to nodes with different degree than their own.

4.2 Users' Activity and Authority

On Twitter, a user's activity can be measured by the number of tweets he or she posts. In Figure 2 we plot the number of tweets posted on the topic of Iran's election from June 1 to August 1, 2009. We observe that the rate at which users post relevant tweets gradually increased as the events unfolded in Iran and the use of Twitter provoked attention, spiking dramatically in relation to political events inside Iran as well as in relation to new events and incidents particular to the web. For example, on June 20 mass protests took place in Tehran and security forces responded with violence; a young Iranian woman named Neda Agha-Soltan was shot and killed by the Basij -government militia- in Tehran. Videos of the killing taken with mobile camera were posted on youtube and rapidly spread across the Internet. On that day, Twitter users' activity around the topic of Iran's election reached its peak of about 300K tweets. We analyze the activity by each user and observe that the distributions of user's activity follows power law with exponent about -1.92 (The K-S metric D is equal to 0.0078) Figure 3(a). The cutoff in power law degree distribution in Figure 3(a) is due to the limit of downloading 3200 status messages per user in Twitter API.

One might expect that users who posted a lot of tweets regarding Iranian Election would have a lot of followers who also post on this topic. Intuitively, we expect the attention (number of followers) a user gets to be correlated with the user's activity (number of tweets). However this does not seem to be the case in our Iran related users. The Pearson correlation coefficient between the number of tweets and the number of followers is only 0.040. Furthermore, a user's activity is not correlated with how many friends he or she has, as the correlation coefficient is only 0.041. We analyze users who are authoritative or prominent within the community in this case. Figure 3(b) shows that the distribution of retweets is heavy-tailed and we can fit a power-law distribution with exponent of -1.94 with a K-S metric D equal to 0.0110. Correlation between the number of retweets and number of followers is 0.1824 while the correlation between number of retweets and the number of tweets is 0.2327. Therefore if a user has more tweets and more followers, she or he will get more retweets.

5. TWEET NETWORK: INFORMATION CASCADES

A cascade is an information propagation phenomena in which an idea becomes adopted because of influence by others. A tweet network is a collection of cascades where every node represents a tweet and there is a directed edge from tweet u to v if tweet v retweeted tweet u. There are a total of 3,219,038 nodes (tweets) in our tweet network and 2,600,295

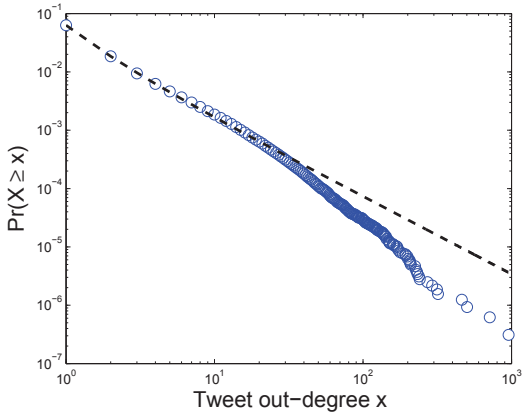


Figure 4: Cumulative distribution of out-degree in tweet network. Power law fit to the data with exponents -2.33

nodes are isolated, meaning that they did not get retweeted by others. These nodes represent the most common cascade in our dataset and we call it trivial cascades. After ignoring 2374 self edges, we got 444,843 edges in our tweet network. Applying the best power-law fit using the maximum likelihood method, we found out-degree distribution follows power law with exponent equal to -2.33 in Figure 4 and the K-S metric D is equal to 0.0045. Note the in-degree of node in our tweet network is 1 or 0, meaning that tweet is a retweet or not.

We continue with the analysis of the structure of the information propagation when certain tweets become popular and are retweeted by the other tweets. We are interested in how information propagates, how large the cascades are, how large the audiences are and how they compare with the F-F network we observed. The cascades are the subgraphs of the whole tweet network that have a single initiator and we present the information propagation from the initiator to the rest of nodes.

5.1 Cascade Shapes

We can decompose the tweet network into weakly connected components and every component represents cascades of different information propagation. For each component, we can find out the node that has zero out-degree to be the initiator for the cascade and perform breadth-first search (BFS) to obtain the rest of the cascade nodes. We want to see what are the common cascade shapes and how do the real cascades look like. To obtain the frequency and examples of the common cascade shapes, we create a signature that is composed of the number of nodes, the number of edges, the sorted in-and out-degree sequence as well as the singular value of the adjacency matrix obtained from singular value decomposition for each cascade [18]. We consider these features as a good signature, since the isomorphic graphs would have the same signature. Then we use hashing on these signature value and find the frequency of each signatures.

The top ten common nontrivial cascade shape is presented in Table 2. The cascades are ordered by frequency, and the script of the label gives frequency rank. For example, G_9 is 9th most frequency cascade with 1424 occurrences. We

Table 2: Top ten common nontrivial cascade shapes ordered by the frequency. For each graph we show the number of nodes, the number of edges and frequency.

ID	Graph	# of Nodes	# of Edges	Frequency
G_2		2	1	112895
G_3		3	2	21814
G_4		4	3	7269
G_5		3	2	5591
G_6		5	4	3482
G_7		4	3	3194
G_8		6	5	1977
G_9		5	4	1424
G_{10}		7	6	1315
G_{11}		8	7	932

find that there are 173,282 non-trivial cascades with total of 1817 different shapes. The distribution of cascade shape frequency also follows the power law distribution as exponent equal to -1.6 (The K-S metric D is equal to 0.0281). Furthermore, we notice that real cascades tend to propagate as certain shape and there are some interesting observations.

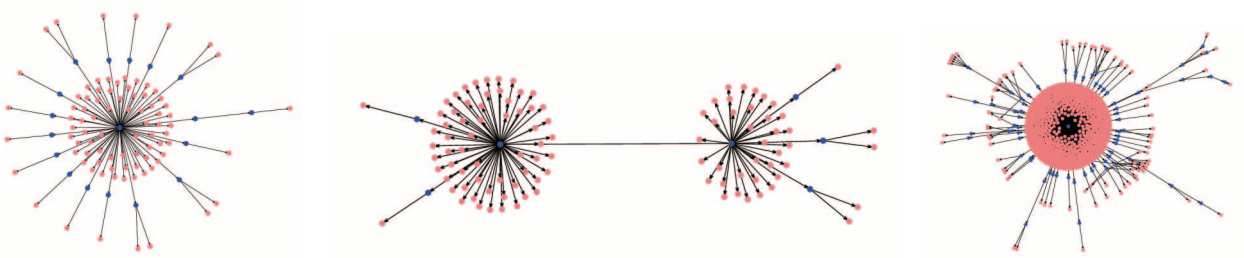
- Cascades tend to be wide, and not too deep. For example, G_3 is more common than G_5, and G_4 is more common than G_7. In general, the maximum depth of the cascade is 7 while the maximum width of cascade is about 1000.
- The cascade frequency does not simply decrease as a function of number of nodes. For example, G_4 with four nodes is more common than G_5 with three nodes. In general nodes are more likely to appear in the first depth of cascades.
- Most of the cascade have a central hub like Figure 5(a), and cascades of two or multiple hubs are less often to occur like Figure 5(b). In general, users are more likely to get information directly from the same user in the network.
- The largest cascade in the Iranian Election (Figure 5(c)) is initiated by Stephen Fry about spreading proxies that help Iranians bypass internet filters.

5.2 Cascade Size

We examine cascade sizes, that is how many tweets are in each cascade. We show the cascade size distributions plot in Figure 6(a). We observe that overall cascade size distribution is power-law with exponent equal -2.51 (The K-S metric D is equal to 0.0134).

5.3 Audience Size

We show the audience size distribution of observed cascade in Figure 6(b), and its complementary cumulative distribution function(CCDF) follows a stretched exponential distribution. Although the number of users participated in the cascade is not large (the maximum is just over 1000) compared to the total number of users in the network, a



(a) An example of central hub (b) An example of two hubs (c) Largest cascade

Figure 5: Real cascades observed (a)'StopAhmadi' wrote: Please @Twitter and @ev don't take down Twitter, for the iranian ppl #iranelection (b) 'RealTalibKweli' wrote: Pray for the protesters in Iran. Regardless of your politics (c) 'Stephenfry' wrote: Functioning Iran proxies 218.128.112.18:8080 218.206.94.132:808 218.253.65.99:808 219.50.16.70:8080 #iranelection - feel free to RT

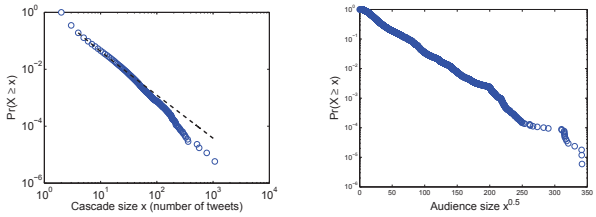


Figure 6: Cumulative distribution of cascade size and audience size. A message can reach a large audience on the network even when the number of users who retweet a message is not very large.

cascade still can reach a large portion of users on the network we studied (120K of out of 500K users are reached in a large cascade). One explanation would be a lot of users on Twitter are information seekers [11] who might post rarely, but followers other users regularly. By comparing the maximum size and the maximum audience of cascades, we can see the amazing power of Twitter in the propagation of information. Even only a small number of users are active, information can still reach many audiences by propagating on the F-F network.

6. PUBLIC TIMELINE VS. F-F NETWORK

In this section, we examine the mechanism of public timeline versus F-F network in shaping the dynamics of information propagation. We studied the interaction between network structure and information flow. What is the Twitter's role in the information propagation, a social network or a social media?

Different from social networking sites like Facebook and MySpace, Twitter has a front page which includes a search bar and a list of trending topics. This unique function not only allows users to see what the world is happening in real-time, but also provides users another source to pick up interesting tweets and retweet them. For all the retweets we observe, we check whether the retweeters are the followers of the author of the tweet. We call these retweets *followers' retweets* and this information help us to study how the information propagate through F-F network.

In Figure 7, we show the percentage of followers' retweets over the span of our dataset. Before the Election when there

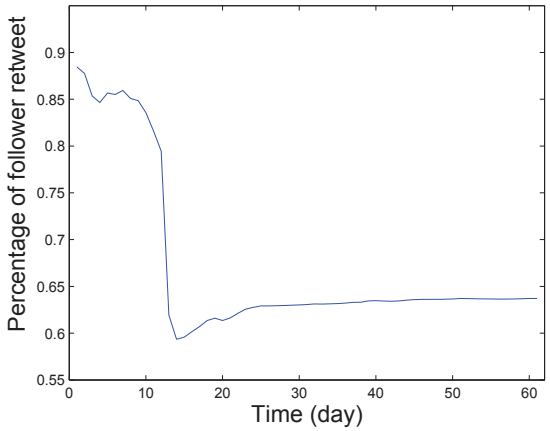


Figure 7: Percentage of followers' retweet. As the whole issue provoked attention, the percentage dropped and approached to 63.7% in the end.

were little traffic, most of the retweets were coming from the friends' posts. As the whole issue provoked attention, the percentage of retweets that are followers' retweets dropped and approached to 63.7% in the end.

Figure 8 shows the relation between number of tweets and percentage of nonfollowers' retweets per day. Once the number of tweets posted exceeded 10k per day, the percentage of nonfollowers' retweet increased by 10%. This is consistent with our hypothesis that as Iranian Election became a more popular trend on Twitter, more and more users were following this topics by retweeting them from the front page.

Therefore F-F network are the primary mechanism for spreading information before topics were promoted to the front page while after promotion a lot of users may retweet from the public timeline. This suggests Twitter serves a role of social networking as forwarding users' tweet to their followers. At the same time, Twitter is a social media where users can get fresh and related tweets from the public timeline.

6.1 Information Propagation via F-F Network

To study the information propagation via F-F network, we analze the retweet characteristic of tweets by estimation

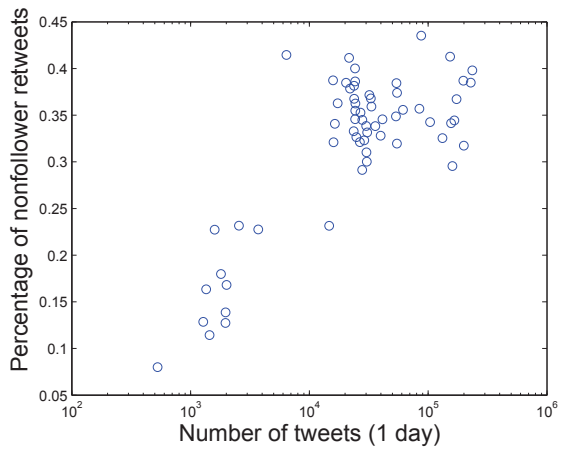


Figure 8: Number of tweets versus percentage of nonfollowers' retweets per day. Once the number of tweets posted exceeded 10k per day, the percentage of nonfollowers' retweet increased by 10%.

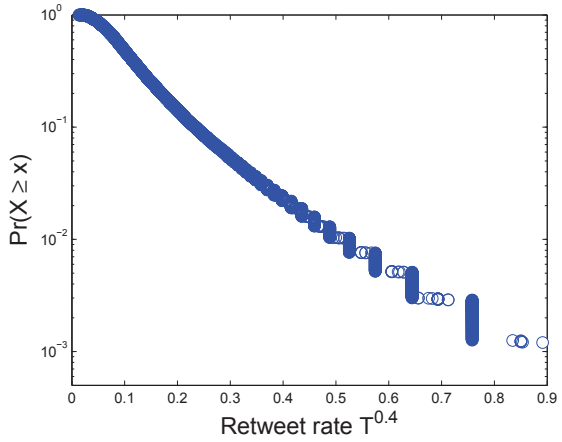


Figure 9: Cumulative distribution of retweet rate decays with a stretched-exponential law.

retweet rate $T(x)$ of tweet x as follows:

$$T(x) = \frac{\text{number of retweets that x received from followers}}{\text{number of followers that author of x has}}$$

We find the most of retweet rates are zero since they did not get retweeted. The high variance in retweet rate suggests different content may have different popularity among their followers. We represent the popularity of each content by the retweet rate of the cascade and show the retweet rate distribution for different content in Figure 9. CCDF of retweet rates decays with a stretched-exponential law. The mean retweet rate of these non-trivial cascades is 0.0136 but standard deviation is as high as 0.0501. These observations confirms our hypothesis that the content plays a key role in the structure of information propagation.

The observation that cascades tend to be wide not too deep indicates the retweet rate may decay as the cascades spreads away from the source. We define the *retweet rate decay factor* at hop N as the ratio between retweet rate at hop N and retweet rate at hop $N-1$. For example, a factor of 0.5 means the retweet rate at hop N is half of retweet rate at hop $N-1$. All these retweets are a maximum of six hops away from the source. We find that the mean of decay factors are all about 0.2 while the standard deviations are very high, which suggests the variance of the tweet content affects the decay factor. *Therefore the retweet rate decays exponentially as the cascades spreads away from the source* and one possible explanation would be that the freshness of the tweet would drop as the time goes on.

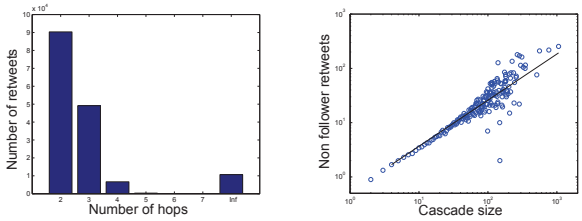


Figure 10: Nonfollowers' retweets. (a) shows 57.47 % of nonfollowers' retweets can be reached by two hops so most of the time, users are just retweeting their friends' friends.(b) shows larger cascades included more retweets of this nature.

Our study shows that retweet rate is content-dependent as well as depth-dependent. Hence, it might not be possible to ignore the modeling of content-level properties when it comes to studying information propagation through F-F network.

6.2 Information Propagation via Public Timeline

If the retweet does not come from their friends, where did users see these tweets? We study the retweets outside the F-F network and find the shortest paths it took the retweeters to reach the author through F-F network. In Figure 10(a), we can see 57.47 % of retweets can be reached by two hops so most of the time, users are just retweeting their friends' friends. Another possible explanation is: some users would like to give credits to the original users when they see their friends retweeting. While 88.75% of the nonfollowers' retweets can be reached within three hops confirms the small world phenomenon. However there are still 6.79% of the retweets cannot be reached through F-F networks, meaning these retweets must be coming from the search bar on the public timeline. This observations suggests Twitter also serves the role of social media in propagating information.

Furthermore,we analyze how nonfollowers' retweet contribute to the overall cascade. Figure 10(b) shows the number of nonfollowers' retweets for each cascade and the slope in the log-log plot is about 1. There are linear relationships between the number of outside retweets and the size of cascade and it suggests rich-get-richer phenomena and tweets are equally likely to get retweeted outside F-F network.

7. CONTENT OF CASCADES

We have shown in the previous section, the content of tweet plays a key role in determining its retweet rate, consequently its dynamics of information flow. Then what kind of content is popular for this topic? Who are the most retweeted users on this issue? These are some questions we

would like to answer in this section. Understanding these questions would help us keep most updated information in real-time.

7.1 Content of Tweets

Study of contents of collected data in its context can be a compelling aspect of data analysis. We looked at the contents of medium and large cascades (with over 30 nodes) in our data set and observed several noteworthy characteristics. The contents of tweets in medium and large cascades can be categorized as follows:

- *Breaking news* An important characteristic of the Twitter network is the real-time nature of much of the information in tweets. For the dataset studied in this paper, real-time reports of events in Iran were important to individuals following the post-election unrest and so a large number of tweets include breaking news. These tweets were sometimes sent by official news media in the form of links to the news piece on their website. In some other cases tweets were either updates by Iranian people in Iran, or individuals who had direct contact with eyewitnesses in Iran. Some of these tweets kept spreading long after the incident had passed.

- *Non-time-sensitive material* Sharing photos and videos, political analysis, personal accounts of protests in blogs, and instructions for the Twitter community on how to get involved, were among other types of content in tweets. These tweets commonly included links to websites that contain the information. The two largest cascades in the dataset are about spreading proxies that help Iranians bypass censorship that blocks many websites. Other popular tweets include instructions on engagement of Twitter community in support of protests, directions on how to conduct Denial of Service attacks on Iranian government websites, first aid information for people in Iran, and instructions on how to avoid spreading rumors and detect reliable information. Other tweets shared plans for future actions on the ground in Iran, such as time and locations of future protests or plans for a national strike.

  In our dataset, 487,005 distinct URLs were used 1,582,537 times. Frequency distribution of URLs was power-law with an exponent equal to -2.14, which suggests the rich-get-richer phenomenon [4] (with K-S metric D of 0.0047). The most popular URL found in our dataset is http://helpiranelection.com/ (appearing about 200K times). The website adds a green overlay or a green ribbon to a user's Twitter avatar in support of the protesters in Iran who also used the color green.

- *Rumors and misinformation* Unverified information from unknown sources can lead to spread of rumors and misinformation on Twitter. It appears that the Twitter community was relatively successful in recognizing reliable users as sources of information. Nevertheless there were rumors that spread during the period of our study. Specifically one rumor that tanks had appeared on the streets in Tehran spread easily on Twitter. On a few occasions rumors about the arrest of opposition leader Mir Hussein Mousavi were spread either intentionally or due to some level of fear and hyper-sensitivity to the possibility of such an event.

- *Spam* We find some irrelevant hashtags came with our tweets, for example #jobs and #loan which appear more than 5000 times in our dataset. Spammers tried to use the hashtag #IranElection in order to use its popular public timeline to advertise their own websites. It has been confirmed that furniture chain Habitat took advantage of the protests in Iran to market its spring collection on Twitter [2].

- *Others* Some of the largest cascades are about Twitter itself. The Twitter community was very aware of its own activism and role in the Iranian struggle, although sometimes their perception of this role was exaggerated. A number of largest cascades are about the US government, such as Barack Obama's statements about the unrest. In fact the most retweeted Persian-language tweet was by the White House with a link to Obama's press conference on Iran (247 retweets). Another interesting observation is that some of the cascades -including the fourth largest cascade- are jokes, e.g. by The Onion. There were a lot of jokes, encouraging words, and funny slogans on the ground in Iran during the protests, which helped release tension and diffuse fear among protesters. Funny tweets might serve a similar function for Twitter users who were following the stressful developments on Iran around the clock.

7.2 Most Retweeted Users

Sources of cascades in medium and large cascades can be categorized as follows:

- *Official news media* Much of breaking news was tweeted by official news media. @breakingnews (breaking news from MSNBC), @cnnbrk (breaking news from CNN), @anncurry (NBC journalist), and @laraabsnews (ABC News) consistently appear in medium and large cascades.

- *Alternative media* Alternative media such as weblogs also have a presence in our dataset. Mashable, a popular social media news blog, has a significant presence in large cascades. Tehranbureau, a news blog with accurate information on Iran, also has a presence as the source of several information cascades, although it has a much less prominent presence than Mashable.

- *Iranian tweeters* A significant number of cascades were originated by Iranian Tweeters, some of these users were tweeting inside Iran (@persiankiwi) and some others were tweeting from other countries (@oxfordgirl). These users were the source of many medium-size cascades (between 30 and 150 retweets).

- *Celebrities* The two largest cascades (1074 and 771 retweets) were originated by a British actor named Stephen Fry. A British author, Neil Gaiman, was also the source of some of the large cascades. These celebrities have a substantial number of followers which helped generate huge cascades.

[2] http://news.bbc.co.uk/2/hi/uk/8116869.stm

8. CONCLUSION

In this study, we use Iran protest as a window to study the medium, structure and mechanism of information propagation. Our contributions are summarized as follows.

- **Medium of information propagation** Twitter's F-F network structure plays an important role in information propagation through retweets, the search bar and trending topics on Twitter's front page offer other avenues for the spread of information outside the explicit F-F network. We examine F-F network on Twitter and many statistics follow a power law structure, such as in-degree, out-degree distribution as well as the distribution of tweets and retweets.

- **Structure of information propagation** We find cascades tend to be wide, and not too deep, with a central hub being more common. The overall cascade size distribution follows a power-law distribution with exponent equal -2.51. Due to broadcasting of tweets, cascades reach a lot of audience on the network we studied even although user participation rate is not high.

- **Mechanism of information propagation** Users employ F-F network to discover and spread information but the popularity of tweet is determined by the content of information, which plays a key role in dynamics of information flow. The popularity of tweet decays exponentially as the cascades spreads away from the source. We also present a brief taxonomy of cascade content and source, and discuss the main categories of the tweets.

Our findings about information propagation have following applications. Understanding the principles of information propagation on F-F network as well as public timeline will be help design better application systems that address different aspects of the social media. Cascade disseminates information of video and images via URLs, hence leading internet traffic outside Twitter. Therefore our findings about structure of information propagation has a significant impact on determining and managing Internet traffic, and hence the Internet infrastructure backbone. Our analysis shows public timeline is an important medium for users to get real-time popular news, and we can take advantage of these trending topics on Twitter to do viral marketing.

9. REFERENCES

[1] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 207–214, 2005.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 54. ACM, 2006.

[3] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the tenth ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.

[4] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[5] J. Beilin and M. e. Blake. The iranian election on twitter: The first eighteen days. *Web Ecology Project*, 2010.

[6] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *hicss*, pages 1–10, 2010.

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[8] E. Damiani, S. Di Vimercati, S. Paraboschi, and P. Samarati. An open digest-based technique for spam detection. In *Proceedings of the 4th IEEE international conference on peer-to-peer computing*. Citeseer, 2004.

[9] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[10] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):8, 2009.

[11] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[12] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 617. ACM, 2006.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *WWW'10: Proceedings of the 19th International World Wide Web Conference*, April 2010.

[14] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[16] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 429. ACM, 2007.

[17] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. *Society of Applied and Industrial Mathematics: Data Mining*, 2007.

[18] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining*, pages 380–389, 2006.

[19] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, page 42. ACM, 2007.

[20] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122, 2003.

[21] T. O'Reilly and S. Milstein. *The Twitter Book*. O'Reilly Media, Inc., 2009.

[22] E. Rogers. *Diffusion of innovations*. Free Pr, 1995.

[23] D. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5766, 2002.

[24] F. Wu, B. Huberman, L. Adamic, and J. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, 2004.