Identifying Information Spreaders in Twitter Follower Networks

Xufei Wang, Huan Liu, Peng Zhang, Baoxin Li
Computer Science and Engineering
Arizona State University, Tempe, AZ 85287, USA
{xufei.wang, huan.liu, pzhang41, baoxin.li}@asu.edu

ABSTRACT

A number of research efforts on Twitter have been contributed towards understanding various factors that are related to retweetability, analyzing retweeting and diffusion patterns, predicting retweets, etc. One fundamental research question remains untackled: given a user and her followers, which of the followers are likely to spread her tweets to the world (the information spreader identification problem)? Answering this new and open problem helps to bridge the gap between analyzing retweetbility and understanding information diffusion. Using a large scale Twitter data set, we first find that retweet history is not an ideal method for identifying information spreaders, especially for the long tail users. Backed by statistical analysis, we set forward to extract meaningful features and present a set of feasible approaches for identifying information spreaders in the Twitter follower networks. Our study reports interesting findings, sheds light on many practical applications, helps understand the mechanisms of relaying information from one user to her followers, and offers future lines of research.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Science**]: Sociology

General Terms

Experimentation

Keywords

Retweet, Information Spreaders, Ranking

1. INTRODUCTION

The microblogging service Twitter has grown in popularity by providing and sharing real-time information of up to 140 character-long tweets. One of the distinguishing features of Twitter is the retweet mechanism which reposts tweets that are written by other users and shares them with one's own followers. Retweeting is deemed as a key mechanism of information diffusion in Twitter [27]. A number of research efforts have been performed towards investigating the factors that impact retweeting in the Twitter network [3, 27], studying retweet patterns or understanding retweeting behaviors [15, 19, 32], predicting which tweets are likely to be retweeted [20, 22, 33], and predicting information diffusion by analyzing the properties of tweets and users [28, 31].

1.1 The Information Spreader Identification Problem

An *information spreader* on Twitter is a user who is willing to retweet (actively) from her friends and share the information to her own followers. On Twitter, most users are passive information consumers [18], how information is spread to the silent majorities arouse significant research interests in many perspectives [8, 16, 25, 29, 31]. However, one fundamental question remains untackled: given a user and her followers, which of them are likely to retweet her tweets? The formal definition will be given in Section 2.

It is worth emphasizing that the information spreader problem is different from other well known problems on Twitter. Retweet prediction aims to answer whether a tweet is likely to be retweeted [20, 22, 33]. Information diffusion on Twitter [28, 31] attempts to understand why and how an idea is spread in the Twitter follower network. Both of them are not designed to answer *who* spreads or relays new ideas in social networks. Our work enables to gain more insights in understanding the retweetability and information diffusion in the Twitter follower network.

There are apparent differences between the proposed problem and the indentication of important or influential persons [2, 6, 15, 30] on Twitter as well. First, the important or influential persons on Twitter are identified at the global level, i.e., they are important on the whole social network. However, the information spreaders are identified at personal social networks, i.e., they are identified on the user and her followers. Second, in (online) social networks, as we will later show, an information spreader could be anyone but not necessarily an important person.

1.2 Motivations and Contributions

Knowing these information spreaders is important in many fields. The motivations are summarized as follows,

- Expedite the speed of information diffusion. By knowing willing-to-retweet followers, the social networking services could harness these users to deliver real-time information to others more efficiently and effectively.

WSDM 2013 Rome, Italy

In addition, this knowledge could also be leveraged to bring the information producers and consumers closer.

- Increase the accessibility of information. The willing-to-retweet users are more likely to spread information to a broader spectrum of people in social networks. Thus, it could increase the exposure of the information to people who are potentially interested. For instance, it would be desirable for people who are seeking answers desperately for a question to have as many people as possible to be aware of the problem. It would also be desirable to political groups to propagate their propositions through the "word-of-mouth" message forwarding to as many people as possible.

- Ease the adoption of new ideas. The diffusion of innovations theory suggests that different sorts of people will adopt a new idea at different times after it is introduced. For example, opinion leaders are likely to exert influence to early adopters, while peer pressure plays its role if a user is surrounded by those who have already adopted the idea [24]. The willing-to-retweet followers could act as opinion leaders or the late majority who would influence the rest population on accepting new ideas.

- Discover the backbone of information pathways. Not all users participate equally at disseminating information in social networks. Actually only a small part of users and edges consist of a subgraph which has the potential to spread information quickly on social networks [12, 14]. The identification of the willing-to-retweet users will likely help characterize the backbone of social networks.

We believe that one of the contributions of this work is in the area of understanding and modeling information diffusion in the Twitter social network. In addition, identifying information spreaders is relevant to a number of interesting applications of social networks such as designing and adapting viral marketing strategies in ways of harnessing the power of social media. The contributions are summarized as follows,

- Propose a new problem to identify information spreaders in the Twitter follower networks, and

- Propose and empirically evaluate a set of feasible approaches to solve the problem.

Our primary objective is three fold: understand retweet patterns between pairs of following users in a Twitter social network, examine the effectiveness of features originated from both social network and user generated content, and evaluate various approaches in identifying information spreaders. Next we will formally define the novel problem of identifying information spreaders in Twitter follower networks.

2. PROBLEM STATEMENT

We first introduce notations to be used in the rest of the paper. The Twitter social network can be modeled as a directed graph $G = \{U, E\}$, where $U = \{u_1, u_2, \ldots, u_n\}$ is the set of users and E is the following relationship between users. A typical Twitter user u has a set of followers ($Follower(u)$) and friends ($Friend(u)$) which is known as followees before. We denote contacts ($Contact(u)$) as the union of the user's followers and friends, that is,

$$Contact(u) = Follower(u) \cup Friend(u) \tag{1}$$

Friends, followers and contacts are called *neighbors* of a user as they are connected in a certain manner. The cardinality of a set represents its size, e.g., $|Friend(u)|$ represents the number of friends of user u.

Common friends CFR refer to the set of users who are followed by two users u_i and u_j. Similarly, we define the common followers CFO and common contacts CCO as the users who are shared by the two corresponding sets, i.e.,

$$\begin{aligned} CFR(u_i, u_j) &= Friend(u_i) \cap Friend(u_j) \\ CFO(u_i, u_j) &= Follower(u_i) \cap Follower(u_j) \\ CCO(u_i, u_j) &= Contact(u_i) \cap Contact(u_j) \end{aligned} \tag{2}$$

We aggregate all tweets that are owned by user u, then form a term-frequency vector $t(u)$, excluding stop words. Similarly, the set of hashtags and URLs that are associated to user u are represented as term-frequency vectors $ht(u)$ and $url(u)$, respectively.

Given a user u and her followers, our primary focus is to rank the followers by their likelihood of retweeting anyone of her tweets, considering a wide range of features from the Twitter social network and user generated content. The top-k most likely to retweet followers are returned as information spreaders of this user. Let $P(f_i|u)$ be the retweet likelihood of the i-th follower from u, the objective function of identifying information spreaders is defined as follows,

$$\begin{aligned} \max_{\{f_i\}_{i=1}^{k}} \quad & \sum_{i=1}^{k} P(f_i|u) \\ s.t. \quad & f_i \in Follower(u) \end{aligned} \tag{3}$$

To demonstrate that the proposed problem is meaningful and doable, a data set should be selected carefully and it should meet several requirements. First, in order to draw sound conclusions on highly dynamic users behaviors in online social networks, a large scale data set is essential. Second, the data set should contain rich information such as user profiles, user generated content, recordable online activities, measurable interactions between users, the social network, etc. Third, the flow of information from one person to another should be clearly stated or easily obtained. These are key challenges to obtain a usable data set for the purpose of identifying information spreaders. We found that the following collected Twitter data, which gathered enormous history of events about what happened in the Middle East, is a fit for our needs.

3. DATA COLLECTION AND STATISTICS

Below, we discuss our data collection procedure which is followed by investigating the possibility of using retweeting history to identify information spreaders.

3.1 Collection Methodology

We systematically collected tweets, user profile and the social network through the Twitter API. This process involved the usage of certain parameters, namely: keywords, hashtags, and geographic regions. We collected more than 660 thousand users and 16 million tweets which were generated about or from the countries: Egypt, Syria, Libya,

Table 1: Parameters Used to Data Collection

Country	Keywords/Hashtags	Geo-Boundary
Egypt	#egypt,#muslimbrotherhood,#tahrir,#mubarak,#cairo,#jan25,#july8,#scaf,#noscaf	(22.1,24.8),(31.2,34.0)
Syria	#syria,#assad,#aleppovolcano,#alawite,#homs	(32.8,35.9),(37.3,42.3)
Libya	#libya,#gaddafi,#benghazi,#brega,#misrata,#nalut,#nafusa,#rhaibat	(23.4,10.0),(33.0,25.0)
Bahrain	#bahrain,#bah	(50.4,25.8),(50.8,26.3)
Yemen	#yemen,#sanaa,#lbb,#taiz,#aden,#saleh,#hodeidah,#abyan,#zanjibar,#arhab	(12.9,42.9),(19.0,52.2)

Table 2: Statistics of the Twitter Data Set

Measure	Value	Measure	Value
Users	666,168	Mean Friends	130.20
Mean Followers	130.20	Mean Contacts	217.09
Links	86,710,704	Bidir. Links	19.9%
Tweets	16,043,422	Retweets	3,874,449
URL	6,531,602	URL Ratio	40.33%
Hashtag	37,276,618	Hashtag Ratio	97.88%
Reply	472,160	Reply Ratio	3.98%
Mention	972,042	Mention Ratio	5.49%

Bahrain and Yemen. The tweets were crawled using the streaming API over the period of 7 months starting from February 1, 2011 to August 31, 2011. A full list of the parameters used is presented in Table 1. Column 2 in the table contains the keywords and/or hashtags used. Column 3 contains the geographic boundary box surrounding each country used to crawl all the geo-located tweets from that region. The box is specified as the SW corner (longitude, latitude) of the geographic box followed by the NE corner (longitude, latitude) of the box, separated by a comma. Essentially, if a tweet contains one of the hashtags *or* it is geo-located within above regions, it is likely to be collected. The crawled Tweets during this period account for approximately 10% of the all Tweets that are hosted by Twitter[1]. We are willing to share the data set upon request in accordance with Twitter's Terms of Use, so other researchers can benefit from our work and discover more intersting findings.

An average Twitter user has around 130 friends and 217 contacts, while the node degree distribution follows a power law: the majority of users have few connections, while a small set of authority users aggregates thousands or even millions of connections. Consistent with prior studies on the Twitter network [15], only around 20% of the links are reciprocal. We computed several other important statistics: the retweet ratio is around 24%, suggesting that information diffusion in the collected data set is prevalent; statistically, 40% of the tweets have at least one URL, whereas, only 4% of the tweets have no hashtags; interactions only account for a small part of the tweets, e.g., around 4% and 5% of the tweets are replies and mentions, respectively. Overall, Table 2 summarizes the statistics about the data set.

3.2 When Retweet History May Not Work

An intuitive approach for identifying information spreaders is to analyze the retweet history. We found that leveraging the retweet history may not be an effective approach for this purpose. Below we explain the potential reasons with statistics from the data.

The 7-month data is split into seven continuous time frames by month, i.e., we split all tweets into seven bins: data in February, March, . . ., August and analyze the retweet correlation between the earlier months and the seventh month (i.e., August). We first demonstrate the retweet likelihood in the seventh month when we know the retweet history, then we study under what conditions that the retweet history may not be useful for inferring future retweets.

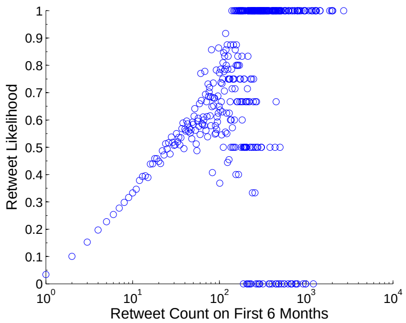


Figure 1: Retweet Likelihood Analysis. We compare the retweet history between the first 6 months and the 7th month. Users are likely to retweet from a person if they have retweeted many from that one.

We first show some statistics about future retweeting and the retweet history in Figure 1. The x-axis represents the number of retweets between a follower and his or her followee (or friend) in the first 6 months, and the y-axis represents the likelihood that this follower retweets from the same followee in the seventh month. The figure shows a roughly positive correlation (e.g., Pearson coefficient $r = 0.21$) between the retweet likelihood and the number of retweets in history with certain outliers. If a follower retweets a lot from the followee (e.g., more than 100 retweets in the last six months), it is very likely that he or she will retweet again from the same followee in the future. However, there is also an exception: around 7.8% of the users who retweet significantly in the last six months do not retweet in the seventh month.

We then start to evaluate to what extent the retweet history may be useful for inferring future retweets. Table 3 demonstrates the ratio that people become inactive (i.e., from retweet to no retweet) in August when different length of retweet history is given. Column 1 represents the time frames that are used as retweet history and the third column shows the percentages that no retweet is observed in the seventh month. We found the retweet history only tells part of the users' retweet story: within the set of users who retweet from their followees in February, only 25.8% of them retweet again in August; in case of considering all six-months' retweet history, there are still more than one third of the users do not retweet again in the future.

The reasons why retweet history may not work for future retweet prediction become clear when we examine the retweet and retweeter count distribution more carefully. The retweet count distribution shows that more than 75% of the users (who retweet at least once) only retweet once in the

[1] We verify it by the Twitter firehose data which can not be used in this paper because of legal issues.

Table 3: Retweet Inactive Ratio: the 3rd column shows the percentages of users who do not retweet from anyone in August, given vary retweet history.

Time Span	Test Month	Inactive Ratio
Feb	August	74.2%
Feb - Mar	August	66.7%
Feb - Apr	August	59.7%
- May	August	56.9%
Feb - Jun	August	50.4%
Feb - Jul	August	36.2%

course of seven months, suggesting they are extremely inactive. Therefore, the retweet history for the majority of users is unavailable, which significantly degrades its usefulness in future retweet prediction. Another reason comes from the other perspective: the number of retweeters. The statistics show that around 50% of the users have been retweeted by only one follower in the seven-month duration. This fact suggests that users tend to retweet then stop retweeting in the future. Figure 2 shows the two mentioned distributions (i.e., retweet count and retweeter count). The circled curve represents the distribution of retweet count and the asterisked curve represents the retweeter count distribution.

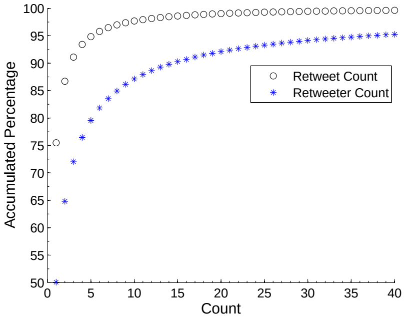


Figure 2: Retweet and Retweeter Count Distribution: in seven months, the majority of users have very few retweets; on the other hand, around half of the users have been retweeted by their followers only once. Both suggest that retweet history is limited by its usefulness for future retweet prediction.

In summary, the retweet history is a good indicator for retweet prediction only if a user is a frequent retweeter. However, the majority of users in the Twitter network are in the long tail, suggesting that they retweet infrequently and that their tweets are retweeted by very few users. For this huge set of users, the retweet history could be unavailable (i.e., cold start problem) or limited by its usefulness, therefore it may not work for future retweet prediction.

4. IDENTIFYING INFORMATION SPREADERS BY RANKING

In this section, we attempt to automatically rank a user's followers by their likelihood for future retweeting. Our hypothesis is that the degree to what extent a person may retweet from her friends can be learned from their online behaviors, interactions, etc. Boyd et al. [3] summarized several reasons why people retweet from their friends. For example, a tweet is informative thus the followers want to share it with their own followers or save it for future personal access, a stance to agree with someone, show support and presence as a listener or to start a conversation, etc.

We propose to extract some features that may contribute to the follower ranking. These features include user similarity, online interaction, structural features and their profiles. Some features are well discussed by prior works such as [20, 22, 27, 33]. Table 4 lists all features that can be roughly categorized into four groups by their functions: proximity, content, interaction, and profile.

Proximity-based features measure the similarity between an arbitrary pair of following users u_i and u_j, relative to the network topology. These features are extracted from the Twitter following network, therefore, are irrelevant to retweeting content. Features include common friends, common followers, common contacts, social status, etc.

Content-based features measure the similarity of the user-generated content between two users. The set of features used in this paper are common hashtags, common URLs and tweet similarity.

Interaction-based features indicate the frequency that two persons talk to each other. We extract the number of replies and mentions between a pair of users as the interaction features.

Profile-based features include the statistics related to each user: the number of status (or tweets), friends, followers and contact counts, list count, the language a person uses, and the account creation date.

4.1 Feature Extraction

Firstly we introduce a few important concepts about Twitter. A tweet is a short message that is limited by 140 characters. A retweet is used to repost a message from another Twitter user. Therefore, it is an important mechanism in how information is spread on Twitter. Retweet is characterized by the abbreviation "RT" at the beginning of a tweet. The "@" sign followed by a user name indicates that it is a mention or reply to other users. In this paper, retweet is considered as information diffusion, while mention and reply are considered as interactions between users. A hashtag is used to group posts by their topics, e.g., a tweet containing hashtag "#egypt" implies that it may be related to Egypt. A hashtag could be any word or phrases that is prefixed with a "#" sign. Also many tweets are embedded with URLs.

For each tweet, we extract the following information where possible: the owner of the tweet, the hashtag(s), URL(s), mentioned user(s) and the reply-to user. Then, we aggregate the information that is related to each user and form the term-frequency vectors $t(u)$, $ht(u)$, and $url(u)$. We found an average Twitter user uses 147 words in total of which around one third of them are unique, suggesting Twitter users are likely to use same words repeatedly. The usage pattern of hashtags is similar to that of tweet terms. However, Twitter users have very close URL usage statistics: on average a person uses 16.4 URLs in their tweets, but 14 of them are unique. Although the average number of terms, hashtags, and URLs are relatively large, the majority of the users use very little of them. For instance, the median counterparts are only 13, 4 and 1 for tweet terms, hashtags, and URLs, respectively. Highlighted by the last column "NZ" (abbreviation for not zero) in Table 5, more than 90% of the users use at least one term or hashtag, whereas, almost half of the

Table 4: Feature Description

Group	Feature	Description
Proximity	Common Followers	The number of users who follow both users
	Common Friends	The number of users who are followed by both users
	Common Contacts	The number of users who have connection with both users
	Mutual Link	Indicator of whether two users follow each other
	Social Status	Larger PageRank values represent higher social status, and vice versa
Content	Common Hashtags	The number of common hashtags that are used by both users
	Common URL	The number of common URLs that are shown in both users' tweets
	Tweet Similarity	The cosine similarity between the two users' tweet vector $t(u)$
Interaction	Reply	The number of replies that one user to another
	Mention	The number of mentions that one user mentions the another in his or her tweets
Profile	Status	The number of Tweets of a user
	Lists	The number of lists that belongs to a user
	Language	The preferred language of a user
	Account	The date that the user's account is created
	Friends	The number of friends
	Followers	The number of followers
	Contacts	The number of contacts

Table 5: Feature Statistics

Measure	Unique		Duplicate		NZ
	Mean	Medium	Mean	Medium	
Terms	52.8	13	147.0	13	91.1%
Hashtag	9.1	4	52.4	4	92.7%
URL	14.1	1	16.4	1	52.9%

users have no URL in any one of their tweets.

4.2 Methods for Ranking Followers

In this section, we summarize the set of approaches that are potentially suitable for ranking a user's followers by their likelihood of retweeting. All these methods assign a retweeting score to an arbitrary pair of following relationship, i.e., $P(f_i|u) \in [0,1], f_i \in Follower(u)$. Some methods are very well developed but are applicable in other tasks. To simplify notations and for ease of understanding, we always use the hashtag as an example to derive the proposed approaches. The definitions can be generalized to other features easily. Assume u_i and u_j are two Twitter users that have a following relationship, e.g., u_i is a follower of u_j.

- **Shared Feature Counting.** Countable features in this data set include shared neighbors (i.e., friends, followers and contacts), shared hashtags and URLs. This approach is reasonable because shared features and the retweet likelihood are correlated. More details are about to be presented in Section 5.1.

$$|ht(u_i) \cap ht(u_j)| \tag{4}$$

- **Jaccard Index** measures the extent to what two sets overlap. It is a normalized similarity measure and its value is between 0 and 1.

$$\frac{|ht(u_i) \cap ht(u_j)|}{|ht(u_i) \cup ht(u_j)|} \tag{5}$$

- **Adamic/Adar Index** assigns more weights to shared features that are rarely used by other people [1]. We consider the hashtags and URLs that are used by Twitter users in the paper to compute this index. Let u_i and u_j be two users, z be a shared hashtag, $F(z)$ represent the number of users who used the feature z, the Adamic/Adar index between two users is given by

$$\sum_{z \in ht(u_i) \cap ht(u_j)} \frac{1}{\log F(z)} \tag{6}$$

we also consider a variation (i.e., Weighted Adamic/Adar Index) which takes into account the number of times that a hashtag has been shared by two users. Let z_{u_i} be the number times that a hashtag z is used by user u_i, the definition is shown as follows,

$$\sum_{z \in ht(u_i) \cap ht(u_j)} \frac{\min(z_{u_i}, z_{u_j})}{\log F(z)} \tag{7}$$

- **Tweet Similarity** is computed by assuming each user as a term-frequency vector after removing stop words. The tweet similarity between two users u_i and u_j is given by the vector similarity,

$$\frac{t(u_i) \cdot t(u_j)}{\|t(u_i)\| \cdot \|t(u_j)\|} \tag{8}$$

- **Regression Models** are used to investigate the relationship between a dependent variable and one or more independent variables. In this paper, the dependent variable is the happening of retweeting (more details in Section 5), and the independent variables are the features mentioned in Table 4 with z-score normalization. More specifically, a regression function $\mathcal{F}$ is to be learned from the dependent and independent variables, i.e., $y = \mathcal{F}(x)$, where x and y represent the vector of independent variables and the dependent variable, respectively. Two regression models are considered: the logistic regression and the random forest regression.

  Logistic regression [11] is widely used in many fields such as social science, economics and marketing. Given a pair of two users f_i and u, $f_i \in Follower(u)$, the likelihood that a user f_i will retweet from user u can be estimated by

$$p(f_i|u) = \frac{1}{1 + e^{-(w^\top x_i + b)}}, f_i \in Follower(u) \tag{9}$$

  where w and b represent the weight of the features and intersect, respectively, vector x_i is a feature vector that is associated with f_i and u.

Random Forest [4] is an ensemble learning method which consists of many decision trees and can be used in both prediction and regression tasks. It takes advantages of high accuracy, being efficient and robust to noise, having no overfitting, etc [23].

5. EMPIRICAL EVALUATIONS

We first demonstrate some statistics about the features, showing that they are not randomly selected. Then we present our findings and discussions.

5.1 Searching For Meaningful Features

In this section, we provide detailed information about the selected features by studying the correlation between features and the likelihood of retweeting.

Common Neighbors. The followers, friends and contacts of a user are all deemed as neighbors. We examine all following pairs in the Twitter social network and find that the retweet likelihood increases as the number of common neighbors increases. As shown in Figure 3 in which the x-axis represents the number of common followers, friends or contacts, and the y-axis represents the percentage that a follower retweets from the corresponding followee. However, the common neighbors are not strong indicators of retweeting likelihood as we notice that the percentages are less than 3%, even two users share as many as 100 neighbors. We also examine users who share more than 100 neighbors, but the likelihood did not increase significantly.

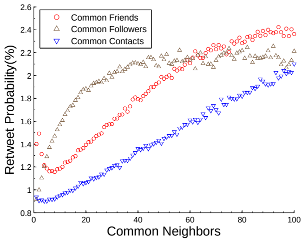


Figure 3: The Retweet Probability w.r.t. the Number of Common Neighbors: the more common contacts between a pair of two users, the higher retweet rate from followers to followees.

Common Hashtags. We first compute the number of hashtags that are shared by a pair of users, then study to what extent this number is correlated to retweeting or not retweeting. Two variations of common hashtag computation strategies are used in the experiments: weighted and unweighted. The unweighted variation is exactly computed by Equation (4), while the weighted version is slightly different by taking the shared frequency into account. Its definition is given below,

$$\sum_{z \in ht(u_i) \cap ht(u_j)} \min(z_{u_i}, z_{u_j}) \tag{10}$$

where z_{u_i} and z_{u_j} represent the number of times that a hashtag z is used by users u_i and u_j, respectively. Our hypothesis is that two users who use the same set of tags more frequently are more likely to retweet from each other. The statistic in Figure 4 verifies this hypothesis: more common hashtags between two users, higher likelihood of retweeting. The observed high retweet probability in this figure suggests that hashtag is a strong indicator for retweet prediction.

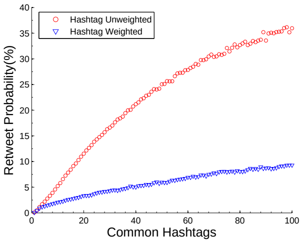


Figure 4: Retweet Probability vs. Number of Common Hashtags. Both weighted and unweighted statistics show that the more hashtags that are shared by two users, the higher likelihood that retweet happens.

Common URLs. URLs in tweets are mostly references to external sources where the tweet is inspired. Similar to hashtags, we consider the shared URLs by two different strategies: weighted and unweighted. The measures for common URLs are transplantable from the definitions of common hashtags. As shown in Figure 5, the retweet likelihood is positively correlated to the number of shared URLs between two users. The retweet probability increases quickly when only few URLs are shared, but then the trend becomes flat as more URLs are shared.

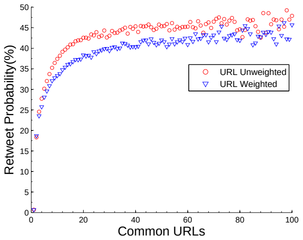


Figure 5: Retweet Probability vs. Number of Common URLs. Both weighted and unweighted measures show that the retweet likelihood and the shared URLs are positively correlated.

Tweet Similarity. Given a pair of users u_i and u_j, the tweet similarity is defined as the cosine similarity of their tweet vectors $t(u_i)$ and $t(u_j)$. Intuitively, two users with a higher tweet similarity are more likely to share certain interests, thus increasing the likelihood of retweeting. We find agreeable evidence in Figure 6 in which the x-axis represents the similarity between two users, and the y-axis represents the likelihood of retweeting. The trend shows that the retweet likelihood is low for users who have a small tweet similarity, whereas, if two users have a high tweet similarity, the likelihood of retweeting is significantly increased.

Social Status is a relative rank or a position that a person holds within a social network. We deem the PageRank [21] value for each user as their social status in the Twitter network. Given two users u_i and u_j, we say u_i has a higher status than u_j if u_i has a larger PageRank value than that of u_j. After examining all retweet pairs in the Twitter network, we find around 85% of retweets are contributed by

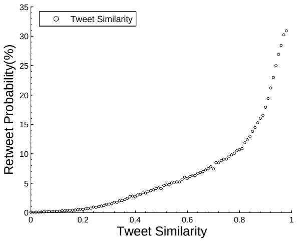


Figure 6: Retweet Probability vs. Tweet Similarity. Large tweet similarity suggests high retweet likelihood.

users in a lower status. The statistics are summarized in Figure 7. Here weighted means that we first make a vote on the number of retweets between two users, then determine whether the retweet between two users are from lower status to higher status or in the way. For example, if user u_i is at a higher status than u_j, and u_i retweets 2 times from u_j, u_j retweets 5 times from u_i, then in the weighted case, the retweet relationship between the two users is "u_j retweets from u_i", or lower status.

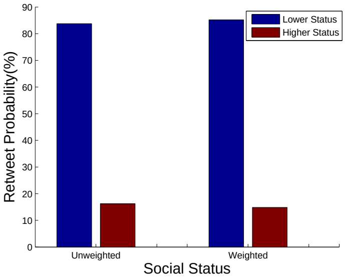


Figure 7: Retweet Probability vs. Social Status. Most Twitter users retweet from others with high Pagerank.

Interactions. Two types of interactions are considered: the reply and the mention. Figure 5.1 suggests that interactions between two users are likely to increase the likelihood of retweeting between the two. As shown in the two figures, the trend of both interactions are similar: both of them increase significantly if two users have few interactions, then become flat. In addition, the trend of replies (Figure 8(a)) shows larger variance than that in mentions (Figure 8(b)). Although both reply and mention are strong indicators for retweeting, the ratios of replies and mentions are small which limits the effectiveness in retweet prediction.

5.2 Empirical Findings

We first introduce the ground truth construction and the measure that will be used to evaluate the performance of above methods. Then we present the experimental results.

Ground Truth Construction. The emergence of retweet between a user and her friends is deemed as ground truth. More specifically, if a user retweets at least once from her friends, then the directed link her to the friend is labeled as positive (i.e., '+1'), whereas, if no retweet occurs during the seven-month time frame, this link is labeled as negative (i.e., '–1'). Thus, for each user, followers are in two categories: the positive set in which all followers retweet at least once and the negative set in which all followers never retweet.

Evaluation Strategy. We evaluate the performance of different methods by the measure precision which is widely used in information retrieval tasks. More specifically, for each user, we rank the followers by their likelihood in retweeting from the user in descending order, then compare the top-k ranked users with the ground truth. In the following experiments, the number k is chosen as 1, 5, 10, 20, 30, 40, 50, 100 and 500. The precision that is averaged over all users in the Twitter social network is reported.

Result Interpretation. Table 6 lists the precision performance of the different methods. Each column represents the top k users that are retrieved, e.g., column 1 indicates that we only consider the first user who is recommended by the corresponding methods.

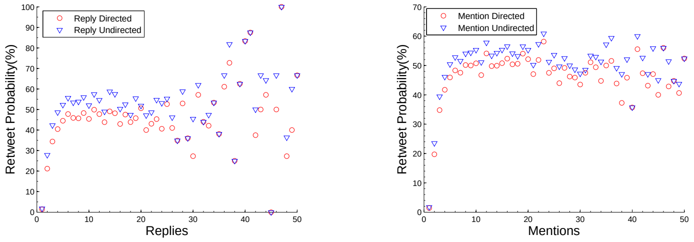
Methods based on URLs work best. In most retrieval or recommendation applications, k is typically chosen to be a small number, e.g., 10. The URL-based methods outperform the other methods with a margin, especially when the selected number k is small, e.g., the best performance of URL-based approach is 11.9% better than the second best approach when $k = 1$. We also notice that different features have different strengths in retweet prediction: URL is the best, followed by tweet similarity, hashtag, interaction, and common neighbors. Statistically, comparing the best performances of URL-based methods to those of feature based methods, the relative improvements are 30.5%, 72.4%, 99.3% and 159.3%, respectively. This result is consistent with prior studies that tweets with URLs are more likely to be retweeted by others [15, 20, 27].

There are several observations of the different treatments of the features: (1) the Adamic/Adar Index consistently outperforms the other approaches; (2) applying weights to the Adamic/Adar index does not improve the performance at all, suggesting information spreaders may be infrequent retweeters; (3) the performance of common feature counting is comparable to that of the Jaccard Index.

We found interaction features are not suitable for predicting which followers are likely to retweet because there are too few interactions in the data, e.g., only around 4% and 5% of the tweets are related to reply and mention, respectively. On the other hand, since more than 90% of users have at least one tweet, the tweet similarity is a relatively strong feature for retweet prediction.

Surprisingly, regression models that take all relevant features into account do not improve the retweet prediction any further. Logistic regression is less effective than the random forest approach. For both regression models, we randomly sample a certain amount of data as training data. Different sizes of instances (i.e., from $1,000$ to $20,000$) that are used to train the regression models are tried, and we find sizes are insensitive to the prediction performance. The results are not presented due to space limitation.

Determining the Best Strategy. For the studied Twitter users who have been retweeted at least once by their followers, the majority of them are retweeted by a very small number of followers. Figure 2 shows that around 50% of Twitter users are retweeted by only one follower. We assign users into different groups by the number of retweeters, then study which methods might be appropriate for diffrent user groups. For example, "group 1" represents the group in which users are retweeted by only one follower, and "group

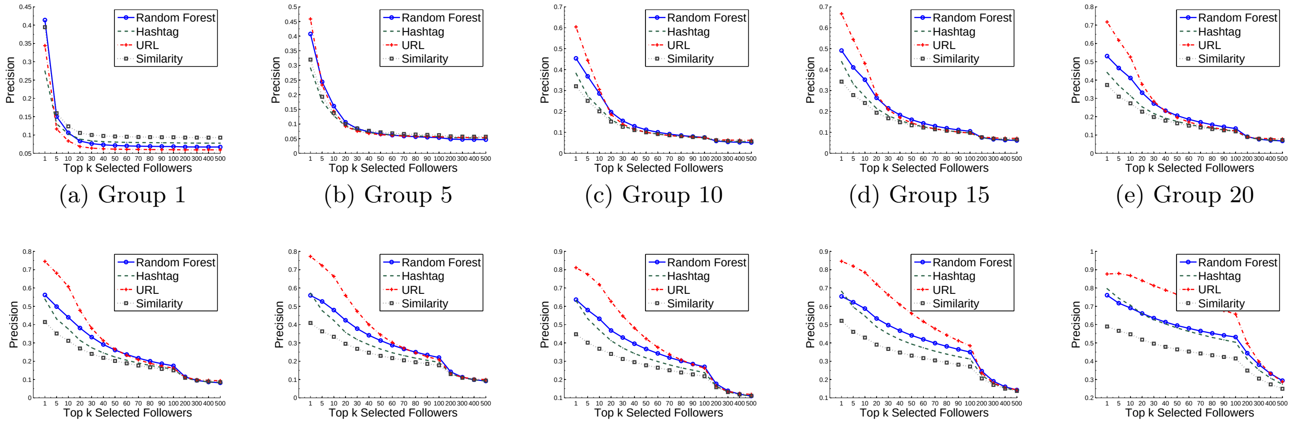


(a) Reply (b) Mention

Figure 8: Retweet Probability vs. Number of Interactions. Directed means that the communication direction is considered.

Table 6: Precision Performance of Various Methods

	Method	Top k Retrieved Followers								
		1	5	10	20	30	40	50	100	500
Hashtag	Common Tags	.29	.18	.15	.13	.11	.11	.10	.09	.07
	Jaccard Index	.26	.16	.13	.11	.10	.10	.09	.08	.07
	Adamic/Adar	.33	.20	.16	.13	.12	.11	.10	.09	.07
	Weighted Adamic/Adar	.29	.18	.15	.12	.11	.11	.10	.09	.07
URL	Common URLs	.42	.25	.19	.15	.13	.12	.11	.09	.07
	Jaccard Index	.41	.24	.18	.14	.12	.11	.10	.09	.07
	Adamic/Adar	**.47**	**.28**	**.21**	**.16**	**.14**	.12	.11	.09	.07
	Weighted Adamic/Adar	.47	.28	.21	.16	.13	.12	.11	.09	.07
Neighbor	Common Friends	.09	.07	.07	.07	.07	.07	.07	.06	.06
	Jaccard Index (CFR)	.15	.10	.09	.08	.07	.07	.07	.07	.06
	Common Followers	.11	.09	.08	.08	.08	.07	.07	.07	.06
	Jaccard Index (CFO)	.15	.11	.10	.09	.08	.08	.08	.07	.06
	Common Contacts	.10	.08	.08	.07	.07	.07	.07	.07	.06
	Jaccard Index (CCO)	.16	.11	.09	.08	.08	.07	.07	.07	.06
Interaction	Reply	.15	.13	.13	.12	.12	.12	.12	.12	.12
	Mention	.18	.15	.14	.14	.14	**.14**	**.14**	**.13**	**.13**
Similarity	Tweet	.37	.21	.16	.13	.12	.11	.11	.10	.08
Regression	Logistic	.23	.15	.13	.11	.10	.10	.09	.08	.07
	Random Forest	.42	.24	.18	.14	.12	.11	.10	.09	.07



(a) Group 1 (b) Group 5 (c) Group 10 (d) Group 15 (e) Group 20

(f) Group 30 (g) Group 40 (h) Group 50 (i) Group 100 (j) Group 500

Figure 9: Precision of Different Approaches on Varying Sized User Groups. Users are grouped by *the number of followers who retweet from them*. Random Forest and Tweet Similarity are best for identifying information spreaders for users having extremely small numbers of retweeters, or the URL-based approach is preferred.

10" represents that these users are retweeted by more than 5 but at most 10 followers. These groups have different characteristics and would deserve different treatments.

We consider four methods in this experiment: Adamic/Adar Index on hashtag, Adamic/Adar Index on URL, Tweet Similarity and Random Forest. Results are presented in Figure 9 in which each figure represents the precision performance on the corresponding user group. In order to return the top 10 most likely to retweet followers, we find in "group 1", it is preferable to use Random Forest or Tweet Similarity methods, for "group 5" and "group 10", both Random Forest and URL-based approaches are good candidates. Otherwise,

Table 7: Comparison between the Important Persons and Information Spreaders. The small nDCG values and Jaccard Index suggest that the information spreaders are not the important persons in egocentric networks.

Measures and Methods		Top k Information Spreaders											
		1	5	10	15	20	25	30	35	40	45	50	100
nDCG	URL	.01	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
	HashTag	.02	.05	.06	.07	.09	.10	.11	.12	.13	.14	.14	.20
	Similarity	.02	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
	Random Forest	.01	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
Jaccard Index	URL	.01	.03	.04	.06	.07	.09	.10	.11	.11	.12	.13	.17
	HashTag	.02	.03	.04	.06	.07	.08	.09	.10	.10	.11	.12	.16
	Similarity	.02	.03	.04	.05	.07	.08	.09	.10	.11	.11	.12	.16
	Random Forest	.01	.02	.04	.05	.06	.07	.08	.09	.10	.10	.11	.15

URL-based approach is preferred. We conjecture that for user groups with an extremely small number of retweeters (long tail users), users might not share any of the single features (e.g., hashtag, URL), so it is imperative to take other information (e.g., tweets or other features) into account.

Are Information Spreaders Important Persons? Important Persons (IP) or influential persons in online social networks are usually characterized by their Pagerank values [15]. For each Twitter user, two ranked lists are present: the list of important persons (IP), and the list of information spreaders (IS). Both ranked lists are in descending order either by their Pagerank values or the likelihood of retweeting. Comparing the IS list to the IP list is able to answer the question. Two measures the discounted cumulative gain (nDCG) and the Jaccard Index are used to quantify the difference between the two lists. In nDCG, the relevance score is binary and is determined in the following way: if the i-th user $IS(i)$ appears in the first i users in the IP list, the relevance value is 1, otherwise, it is 0. That is,

$$rel_i = \begin{cases} 1 & IS(i) \in \{IP(1), IP(2), \ldots, IP(i)\} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Both measures fall between 0 and 1. Value 0 represents that two lists are completely different, and value 1 represents that the two lists are exactly the same. So if the information spreaders are equal to the important persons in each user' follower networks, we would expect that the mean nDCG value and Jaccard Index that are averaged over all Twitter users are close to 1. Results in Table 7 disproves this statement: in fact, the small values suggest that information spreaders are very unlikely to be the important persons in the egocentric networks, and even unlikely to be important persons globally. The results are obtained on the four best strategies: Adamic/Adar Index on URL and hashtag, tweet similarity and Random Forest.

6. RELATED WORK

Twitter characterizes itself as "a real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting". With the availability of large scale Twitter data, many research possibilities arise. We summarize relevant works in this section.

Retweet Pattern Analysis. Retweet is deemed as an effective means to relay information to users who are not necessary direct followers. Kawk et al. studied several interesting topics related to retweet patterns, e.g., the audience size of retweet, retweet tree, temporal aspects of retweet [15]. They found the distributions of the height of retweet trees and the number of participating uses in retweet trees follow a power law: with a small set of retweet trees aggregate a large number of people and spread to longer distances, but most tweet trees only involve a few persons and short distances. They also found that retweeting is time sensitive, i.e., half of retweeting occur within an hour, and 75% under a day. However, they also point out that around 10% of retweets take place a month later.

Retweet Factor Analysis. Many researchers reason the factors that might affect the retweetability of a tweet. Boyd et al. interpret the retweeting practice as a way of conversation in which Twitter participants "retweet others and look to be retweeted" [3]. Based on user feedback of reasons why they retweet and on what they retweet most, they find diverse motivations such as "to amplify or spread tweets to new audiences" and "to entertain of inform a specific audience". Suh et al. find that URLs and hashtags have strong relationships with retweetability [27].

Retweet Prediction. Furthermore, a number of research have been conducted to predict the occurrence of a retweeting [7, 20, 22, 26, 33, 32]. Naveed et al. viewed the likelihood of retweetability as a function of interestingness to generate a model to describe the content-based characteristics of retweets [20]. Petrović et al. also attempt to predict whether a tweet is likely to be retweeted by considering a set of social features and tweet features. They claimed that the automatic retweet prediction performance is as good as the human prediction. They also found that the social features dominate the performance, while the tweet features also add a substantial boost [22]. Zaman et al. propose to predict whether a person will retweet a given tweet from another user by using a collaborative filtering approach [33]. The proposed problem is different from above mentioned retweet predictions. We are interested in the aggregated behavior about which followers are likely to retweet from their friends, but not a single tweet. The proposed problem has close connection to information diffusion.

Information Diffusion on Twitter. Both retweeting and the spread usage of hashtags are treated as information diffusion in Twitter [8, 16, 25, 29, 31]. It is long believed that weak ties are more likely to be sources of novel information, rather than strong ties [9]. Compared to the spread of hashtags, retweeting depends more the Twitter social network. Romero et al. examine the hashtags that are spread on Twitter and observe significant variations on the spread of hashtags on different topics. They conclude that the repeated exposures to hashtags have significant marginal effects on their adoption by other users [25]. Tsur and Rappoport show that the combination of content features with temporal and topological features all contribute to predicting the spread of an idea in a given time frame [29].

Other relevant work on Twitter include friend recommendation or link prediction [5, 17], quantifying influence and identifying influential users in Twitter [2, 6, 10, 30], understanding the factors that affect response such as reply or retweet [7], the usage of Twitter [13, 34], etc.

7. CONCLUSION AND FUTURE WORK

In this work, we propose a novel problem of identifying information spreaders. Identifying information spreaders in social networks is relevant and useful to a broad spectrum of applications such as increasing the depth and breadth of information diffusion, affecting the adoption of new ideas and discovering the backbone of information pathways. We propose a number of feasible approaches based on proximity, content, interaction and profile features. We find simple methods outperform complex methods. The information spreaders are not important persons. Many extensions are worth further exploration. For example, what is the role of tie strength in shaping information spreaders? Another interesting task is to identify topical information spreaders. In addition, more sophisticated methods are to be developed and analyzed.

8. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.

[3] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, 2010.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] M. J. Brzozowski and D. M. Romero. Who should i follow? recommending people in directed social networks. In *ICWSM*, 2011.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[7] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *ACM HT*, 2012.

[8] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *LSM*, page 58-65, 2011.

[9] M. Granovetter. The strength of weak ties. *American Journal of Socialogy*, 78(6):1360–1380, May 1973.

[10] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *ACM RecSys*, 2010.

[11] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience Publication, 2000.

[12] B. A. Huberman, D. M. Romero, and F. Fu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):1–9, 2009.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *WebKDD and SNA-KDD*, 2007.

[14] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *KDD*, 2008.

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[16] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *ICWSM*, 2010.

[17] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[18] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. In *SocialCom*, 2011.

[19] M. Nagarajan, H. Purohit, and A. Sheth. A qualitative examination of topical tweet and retweet practices. In *ICWSM*, 2010.

[20] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *ACM WebSci*, 2011.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[22] S. Petrović, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, pages 586–589, 2011.

[23] M. Robnik-Sikonja. Improving random forests. In *ECML*, pages 359–370, 2004.

[24] E. M. Rogers. *Diffusion of Innovations*. Free Press, 5 edition, 2003.

[25] D. M. Romero and B. M. J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *WWW*, pages 695–704, 2011.

[26] K. Starbird and L. Palen. Will the revolution be retweeted? information diffusion and the 2011 egyptian uprising. In *CSCW*, 2012.

[27] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*, 2010.

[28] E. Tonkin, H. D. Pfeiffer, and G. Tourte. Twitter, information sharing and the london riots? *ASIS&T*, 38(2):49–57, 2012.

[29] O. Tsur and A. Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *ICWSM*, pages 643–652, 2012.

[30] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[31] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*, 2010.

[32] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *CIKM*, 2010.

[33] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern. Predicting information spreading in twitter. In *CSSWC Workshop*, 2010.

[34] D. Zhao and M. B. Rosson. How and why people twitter: The role that micro-blogging plays in informal communication at work. In *GROUP*, 2009.