Statistically Modeling the Effectiveness of Disaster Information in Social Media

Jiang Zhu[1] Fei Xiong[2] Dongzhen Piao[1]
Yun Liu[2] Ying Zhang[1]

[1] Carnegie Mellon University Silicon Valley, Moffett Field, CA 94035, USA
{jiang.zhu, dongzhen.piao, joy.zhang}@sv.cmu.edu
[2]Beijing Jiaotong University, Beijing 100044, China
{08111029,liuyun}@bjtu.edu.cn

***Abstract*—Twitter has increasingly become an important source of information during disasters. Authorities have responded by providing related information in Twitter. The same information channel can also be used to deliver disaster preparation information to increase the disaster readiness of the general public. Retweeting is the key mechanism to facilitate this information diffusion process. Understanding of factors that affect twitter users' retweet decision would help the authority to adopt an optimal strategy in choosing the content, style, key words, initial targeted users, time and frequency. This helps optimizing the communications of disaster messages given the unique characteristics of the Twitter medium. As a result, it will speed up the information propagation to save more lives.**

In this paper, we present the analysis of user's retweeting behavior by studying the factors that may affect this decision, including context influences, network influences and time decaying factors. We aim to build a fine-grained predictive model for retweeting. Specifically, given a tweet, we would like to predict the retweeting decision of each user within a targeted network. We use logistic regression to formulate the problem into a retweeting probability conditioned on the incoming tweet and targeted users. We use this model to examine message spread, because disaster messages do not supersede other communication in the Twitter medium (unlike the emergency alert system announcements over traditional mediums such as television and radio), resulting in a need to 'earn' visibility (e.g., through a high following or reTweeting). We also analyze how time decay would affect user's retweet decision, which in turn affect the information spread and speed. Simulation results illustrate that our model has preferable recall and precision for retweet predicting, and can forecast the trend of information diffusion in the network.

I. Introduction

Twitter, a microblogging service, has demonstrated its strength as an effective new medium. As of May 2011, Twitter has about 200M registered users and over 190M tweets sent daily [1]. With its large user base, capacity of information propagation, and realtime-ness, Twitter increasingly becomes an important source of information during disasters. Much research has since been directed to twitter's potential applications, such as education [8], scientific communication [11], politics [6], and disaster response [15]. Recently, Twitter has successfully distinguished itself from traditional media under a number of political events (i.e., the 2008 U.S. Presidential Election [12], the 2009 Iran Election and Protest [4], [2] by being more spontaneous, mobile, and disseminative.

[1]http://en.wikipedia.org/wiki/Twitter

Researchers have also found that social media such as Twitter do a better job of distributing information during emergencies than either the traditional news media or government emergency services [9] inculding the Haiti Earthquake [1]). When disaster strikes, authorities (e.g., city governments, police departments) can respond by providing related information (e.g., preparedness tips, disaster updates) in Twitter. The interactive nature of Twitter also allows us to evaluate the public's emotional response and perception of utility of disaster-related information, as well as to collect and analyze post-disaster statistics including population affected, survival rate and region after an earthquake, for example. The very same information channel can also be used to deliver disaster-aware information to increase the disaster readiness of the general public in preparation of such events.

"Retweeting" is the most powerful mechanism to diffuse information in Twitter. When a user finds a tweet worth sharing, he could "retweet" it to his followers. The information could thus reach beyond the network of the original author, while the content remains relatively intact. Most existing studies on retweeting try to analyze retweeting behaviors and related factors. For retweeting behaviors, various possible motivations are explored in [7], while the propagation graph and statistics are studied in [5] and [10]. For retweeting-related factors, [16] and [14] found that retweeted and normal tweets are different in various dimensions such as the inclusion of URL's and hashtags, publish time, wording, author publicity, and even the URL shortening service used.

In this paper, we aim to build a simple yet effective predictive model for retweeting. Specifically, given a tweet, we would like to predict the retweeting decision of each users within a targeted network. We are also interested in aggregating these decisions and studying how the information is propagated in Twitter. Such problems are challenging in several ways because (1) many factors can contribute to a user's decision, (2) dependencies exist between multiple users' decisions which depend on a number of network factors, and (3) as the size of the network increases, or as the structure of the network becomes more complicated, both prediction and runtime performances pose major challenges.

The rest of this paper is structured as follows. In Section II we describe the logistic regression model to model an individual user's retweet decision. In Section III, we discuss the factors

that may affect a user's decision to retweet a certain tweet and construct a set of features to reflect contextual influence, network influence and time influence. In Section IV, we construct a Montecarlo simulation framework to model the propagation of tweets in the Twitter network over time based on the logistic regression model for individual users. Finally, in Section V we present our experiment results and engage in discussions on the issues arise from the experiments. We conclude the paper in Section VI.

II. Problem Formulation

Let $G = (V, E)$ be a graph such that each vertex represent a twitter user $u \in V$ and each edge e_{ij} between u_i and u_j represents their relationship that user i follows user j.

The retweeting decision of user i is indexed by the vertices of G and denoted as label y_i. We also denote $\mathbf{x}$ as the observation of a tweet injected into the network G. Using $\mathbf{x}$, we can then generate a set of features $\mathbf{h}_u(\mathbf{x}; G)$ for each user $u \in V$.

According to logistic regression model, we denote retweeting probability of user i given an observation $\mathbf{x}$ in Equation (1)

$$P(y_i = 1|\mathbf{x}) = \frac{1}{1 + \exp^{-\vec{w}^T \mathbf{h}_u(\mathbf{x};G)}} \quad (1)$$

where $\mathbf{h}_u(\mathbf{x}; G)$ denotes all features and $\vec{w}$ is the parameter vector for them. Before combining individual models for the whole network G, we first describe these features $\mathbf{h}_u(\mathbf{x}; G)$ in the following section.

III. Features

In this work, all features $\mathbf{h}_u(\mathbf{x}; G)$ are generated using two sets of interrelated factors. (1) Observations of the incoming tweet $\mathbf{x} = \{a, t^w, M, D^w\}$, where a denotes the tweet author, t^w denotes the publishing time, M denotes the set of users be mentioned in the tweet, and D^w denotes the term frequency of the tweet. (2) Historical characteristics of G: these factors are fixed for a given network at some time snapshot, which encode each user's interested topics, activity level, and relationships with other users.

Because these two sets of factors can generate a large number of features, we first describe the observations based on which we enumerate our features, followed by the respective features generated.

A. Observations

Observation 1 **Content Influence.** Whenever a tweet is about a topic that the reader is interested, or it is either authored or retweeted from a user that shares the similar interests, it is more likely for the reader to retweet the respective tweet.

Observation 2 **Network Influence.** Whenever an author, a fellow retweeter, or a mentioned person in the tweet has a high degree of social connection with a reader, it is more likely for the reader to retweet the respective tweet.

Observation 3 **Time Influence.** According to [10], retweeting probability drops rapidly with time elapsed, where more than 50% of retweets take place within one hour. That is, if a tweet is published by an author or republished by a retweeter closer to its readers' active time slots, it is more likely for the readers to retweet it.

B. Feature Generation

1) Content Influence: **Topic Similarity.** On Twitter, individuals are influenced by current trend, as presented as "Trending Topics", as well as topics of their own interests. Sometimes users are influenced by friends and followers as well. In order to model this phenomena, we need to take global interests, individual interests and shared interests among different users.

Content Features are the similarities between the content of the incoming tweet and various types of contents related to a user. Here we define content as the term frequency vector for a set of tweets, whereas content similarity is defined as the cosine of two term frequency vectors.

Accordingly, the task of breaking down content features is one that enumerates the sets of tweets related to a specific user. We do so in two levels. First, we categorize all tweets about a user into three disjoint groups: (1) tweets published by the user, (2) tweets published by the user's friends, and (3) tweets published by the user's followers. Second, we further divide each group into 3 subgroups: original tweets, retweets, and replies. After the two-level breakdown, we calculate the content similarities between the incoming tweet and each of the 9 subgroups. Additionally, we obtain the global background content (all tweets on twitter), perform the second-level categorization, and calculate 3 additional similarities. Thus, a total of 3 tweet features, 9 user features and 1 relationship feature are obtained.

URLs, hashtags, and mentions. It has been reported in [14] that URL's, hashtags, and mentions also play a role in predicting retweets. We therefore include a series of such features: whether the tweet contains a URL; how frequently the (unshortened) URL domain appears in global and in an user's retweets; whether the tweet contains a hash tag; how frequently the hashtag appears in global and an user's retweets; whether the tweet mentions other users and how often they have been mentioned elsewhere.

Overall, content influence accounts for 8 tweet features, 11 user features, and 1 relationship feature.

2) Network Influence: **Author context.** As reported in [14], a certain social credibility is essential for an author to get retweeted. Therefore, we model the credibility of an author by his (1) number of friends (2) number of followers (3) number of published tweets and (4) number of tweets being retweeted.

Social relationship Whenever an author, a fellow retweeter, or a mentioned person in the tweet has a high degree of social connection with a reader, it is more likely for the reader to retweet the respective tweet. Accordingly, we define such a social tie between two users as the number of (1) mutual friends, (2) mutual followers (3) mutual mentions, and (4) mutual retweets. For each tweet and each user, we then calculates these numbers between the user, the author, and the mentions in the tweet. As a relationship feature, we further define the number of co-retweets from the two users to an author.

In total, Network influence accounts for 4 tweet features, 12 user features, and 1 relationship feature.

3) Temporal Influence: The importance of timing in Twitter has been discussed in both [10] and [16]. According to [10], half of retweets occur within an hour, and 75% within a day. Such an

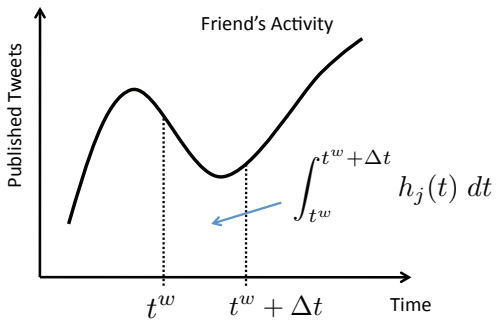

Fig. 1. Time feature is generated from user's friends' activity level

observation suggests a "window of survival", from 1 hour to 1 day, where a certain tweet gets a higher chance to get retweeted. This statistic, however, only describes an aggregate rather than a local phenomenon.

To model the timing factor at the user level, consider the scenario when using the standard Twitter Web interface. Whenever a user checks his timeline, only a certain number of unread tweets can be displayed in one page-view. On a mobile device, this number is even smaller. Therefore, tweets with higher ranks in the timeline naturally have better chances to get retweeted. To model this effect, we introduce two features: Self activity and Friend's activity. The first characterizes a user's response time, while the second characterizes the the number of accumulated tweets within this response time, which is equivalent to the timeline rank.

Self activity. To estimate a user's response time, we model a user's activity level as a poisson process similar to [13]: for any user u, we denote his activity level at some time t as $h_u(t)$, which is approximated by the average number of tweets it publishes in a periodical time slot, e.g., every Wednesday. Accordingly, the response time can be approximated using the average waiting time assuming a poisson process.

Friends' activity. With the estimated response time Δt, the number of accumulated tweets can be written as:

$$\sum_{j \in F_u} \int_{t^w}^{t^w+\Delta t} h_j(t)\ dt \tag{2}$$

where F_u denotes the set of user u's friends. This idea is illustrated in Figure 1.

We calculate both activities using periods of a day and a week. Accordingly, temporal influence accounts for 4 user features.

IV. Information Propagation by Retweeting

The logistic regression model as described in Section II can model the individual retweet decision given a tweet. This is useful to understand individual retweet decision behavior with regards to how the tweet is phased, when the tweet is presented to the user, and how relevant the tweet is towards his or her interests, etc. However, given the social nature of the Twitter system, sometimes it is more prevailing to find out the decisions of a group of users and how they are related to each other. Moreover, we want to understand how the information is propagated and how far and how fast the information can traverse through the Twitter network.

To combine individual decisions into the process of information propagation among multiple users over time, we perform time-slotted simulations using Monte-Carlo framework. This task is formulated as the following:

Given a tweet w to be injected into the system via a group of users $S_0 = u_0 | u_0 \in V$, where V is the set of users in graph $G =< V, E >$, at a given time $t = t_0$, we want to find out at time slots $t_k = t_0 + k\Delta t$, where $k = 1, 2, 3, \ldots, n$ and Δt is the simulation interval, the probability of retweet decision of all users in the network, denoted as $P(RT_u(t))$.

At slot t_k, if this tweet w is retweeted by a user's friend at t_i, where $i < k$, we will evaluate this user u's retweet decision. User u's decision is predicted by the LR model and output a retweet probability Pr. We generate a random number $r \in (0, 1)$ and the retweet decision $R \in \{+1, -1\}$ is defined as

$$R = \begin{cases} +1 & r < Pr \\ -1 & r \geq Pr \end{cases} \tag{3}$$

where $R = +1$ indicates the user would retweet tweet w and $R = -1$ indicates the user hasn't retweeted the tweet in this round of evaluation. If the decision label is +1, we remove this user from User Set S, put the followers of it to S for the next time slot and continue the simulation. Meanwhile we remove these users from S whose iterations have reached the limit.

The steps for each time slot is formulated as in the algorithm:

```
SIMULATE_ONCE( Tweet w, User Set S, Time t )
{
    FOR EACH user u in S
        Calculate retweet probability Pr(u,w,t)
        Generate random number r in (0,1)

        IF r < Pr(u,w,t) THEN
           R = +1
        ELSE
           R = -1
        END

        IF R = +1 THEN
           Remove u from S
           PUSH Tuple (u,t) to decision set RT
           ADD children of u to S
        ENDIF

        count(u in S) = count(u in S) + 1
        IF count(u in S) > ITER_MAX THEN
           Remove u from S
           PUSH u to decision set NRT
        ENDIF

    END
    return S
}
```

The simulation will continue until no user is in S. We record the total number of users that have retweeted the tweet at each time. In order to simulate the propagation of the tweet w in the user relation graph G, we perform the following algorithm:

```
SIMULATE(Tweet w, Initial User Set S0)
{

    Initialize Random Seed
    Initialize decision set RT to empty
    Initialize decision set NRT to empty

    FOR t = 0 to T
        S = SimulateOnce(w, S0, t)
        IF S is empty THEN
            return sets (RT, NRT)
        ENDIF
    END

    return sets (RT, NRT)
}
```

We run this process for M times with different random seeds and gather the accumulated decision results for each experiment, and then calculate the average retweet count at every time slot.

$$P^{+}(u) = \frac{C(u \in \{RT\})}{M} \tag{4}$$

V. Experiments and Results

A. Data Acquisition

Adequate amount of Twitter data is fundamental in training our model and validating its accuracy. We collected twitter data over one week's period, and crawled 11,000,000 tweets and 300,000 user information.

The data we need to perform the modeling and evaluation consists of two types: tweets and user relationships. Tweets are needed in order to profile user interests and activities. User relationships, i.e., *followerships*, are collected to build networks so we can model network influences. We used a small cluster of 6 machines to parallelize data collection process. The collected data is uploaded to a storage server for later processing.

B. Data Pre-processing

After downloading data, several steps of pre-processing are performed to calculate the features. First step is choosing a valid data set for processing. It is not practical to collect the complete tweets and user relationships in a short period of time. Therefore, we adopt a filtering scheme to drop the data that does not pass a threshold test. The threshold test is described as below:

Let's denote "user with threshold T as the one that satisfies the following criterions"

- All tweets of him are downloaded
- All user relationships of him are downloaded
- $T\%$ of the friends and followers of this user have all the tweets downloaded

This threshold can range from 20% to 80%. This is to ensure we have all the tweets of a user and we know all of this user's relationships, but we only know about $T\%$ of its friends and followers' tweets.

After applying the threshold to all the data of a user, we calculate this user's interests as word term frequency of all the tweets this user has published. This step is split into the following steps:

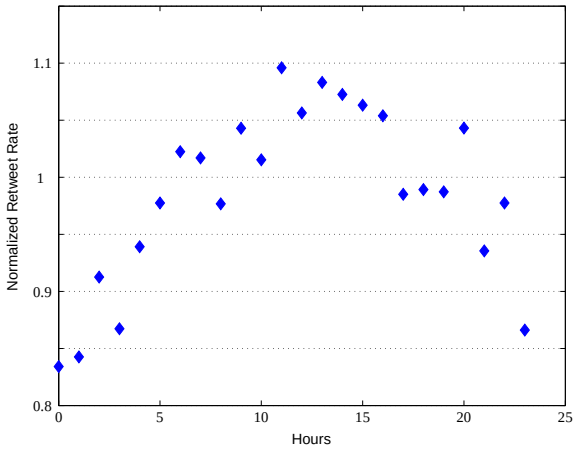

Fig. 2. Normalized retweet rate versus hours

- Remove mentions (@)
- Remove URLs
- Remove digits, punctuation, etc.
- Remove stop words (ultra frequently appearing words, such as "the", "you", etc)
- Word stemming (For example, both "education" and "educate" are stemmed to be "educat")

Then, using the processed tweets, we can get the word term frequency of each tweet.

C. Experiments and Discussions

We discovered in this data set that only 1000 out of 30000 users have both friends and followers, and we collected 25704 tweets associated with these users. Having both friends and follower is essential to carry our experiment because our model requires the features that reflect the interests of both friends and followers.

We trained the individual retweet logistic regression model using LIBLINEAR [3][2]. Due to the limited size of the data set we have gathered so far, we used cross-validation to evaluate the model. The 10-folds cross-validation archive 93.27% in accuracy, 73.47% and 40.26% in precision and recall, respectively.

In order to gain more insights into the factors that affect the retweeting decision, we used the complete data set to train the logistic regression model and generate the model parameters.

Most of the model parameters follow our hypothesis as described in Section III. However, some of the weights are counter-intuitive. e.g. model parameter for feature "interest similarity between the tweet and followers' published tweets" indicates that the interests of followers do not have a significant impact on a user's retweet decision. This is against our initial thought that might suggest otherwise. While not the focus of this report, we decide to continue to investigate issues like this in the future publications.

As far as we know, content similarity, URL, number of followers and followees are strongly associated with retweets [14].

[2]Software available at http://www.csie.ntu.edu.tw/ cjlin/liblinear

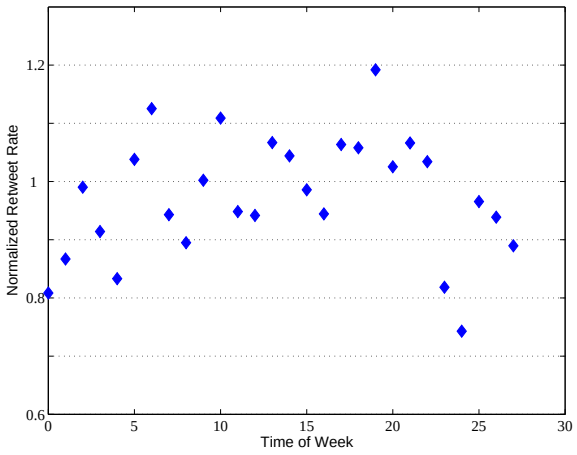

Fig. 3. Normalized retweet rate versus time of week

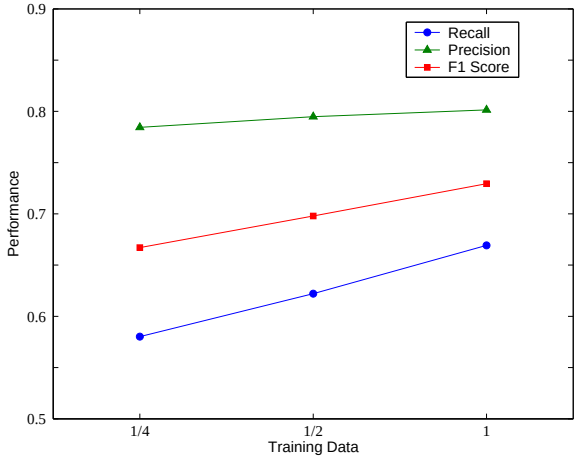

Fig. 4. Recall, precision and F1 score with different size of training data set

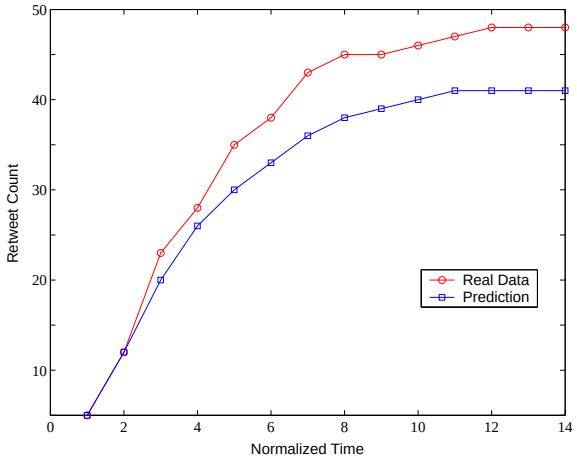

Fig. 5. Retweet count as a function of normalized time

Then we turned our attention to the significance of time features in retweet model, and studied the relationship between the created time of tweets and retweet probability. We calculated the normalized retweet rate to understand how frequently a tweet created at a certain time can be retweeted. The retweet rate at time t is defined as the retweet number divided the tweet number at time t. Then the rate can be normalized by the normalization factor which equals the total number of tweets divided the total number of retweets. The normalization assures a value of 1.0 represents the average retweet rate on tweets. Figure 2 and 3 illustrate the retweet rate at different time of day and time of week. Each day of a week is split into four periods, that is, 0:00 a.m.–6:00 a.m., 6:00 a.m.–12:00 p.m.,12.00 p.m.–6:00 p.m., 6.00 p.m.–12.00 a.m.. From the week chart, tweets around 12:00 p.m. at noon have quite high retweet probability. It is obvious that in each day of a week, tweets in the first time period are very unlikely to get retweeted, while the fourth time period of Friday, the start of the weekend, promotes retweetability significantly. The time features coincide with users' daily behavior. The results suggest our time features provide great benefits to the understanding of the retweeting mechanism in twitter. In our model, combined with receiver actions, time features are embodied as self activity and friends' activity.

We divided the data set uniformly into two groups at random, each of which includes 12852 tweets for 1000 users. One group is used as training set to train user retweeting models, while the other is used as a testing set to estimate whether a user would retweet a given tweet. In the training set of all 12852 tweets, the most active user retweets only 278 tweets, and a large number of users never retweet at all. In order to investigate how large the size of training set is sufficient to build user models, we limit the training set to 1/4 or 1/2 of original set, and compute the average recall, precision and F1 score of users among the whole predicting set. As shown in Fig. 4. The average precision of user retweeting is about 80%, which has nothing to do with the size of training set, implying 1/4 set is sufficient to build a predictive model. Evidently, as shown in the figure, increasing the number of tweets in the training set can improve average recall slightly to 67% with complete training set.

We studied the information propagation process using Monte-Carlo framework as mentioned in section IV. We used logistic regression model to get prior retweet probability for each of 1000 users, and then carried out numerical simulations on real user relationship network to compute users' posterior retweet probability for a given tweet. Thus, the spread and path of information diffusion can be obtained. In the testing set, a tweet has been retweeted for at most 47 times. We chose this tweet as testing data to explore the process of information diffusion, and the simulations begin when the first several users receive the tweet. Time slots of simulations and real retweeting time of this tweet are normalized respectively in Fig. 5, where the results are averaged over 100 different realizations. It is concluded that our model can commendably describe the evolutionary trend of tweet information. The number of final retweets in the simulations is less than real data. This is because the training set contains many tweets that have never been retweeted, so a few users who participated in retweeting are predicted to be inactive in the

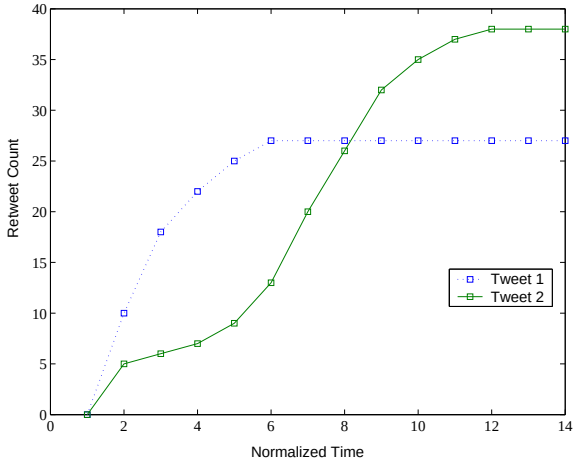

Fig. 6. Retweet trend with different entry points

propagation process.

Overall, when a tweet is posted by a user, our model is robust to model its diffusion trend. Therefore our model can be used to assist in determining the spread and speed that emergency warnings can reach to the public when disaster strikes. By carefully selecting the entry point and the content of the tweets, the simulation results can show different diffusion trends. Fig. 6 shows simulation results with two different entry points which yield different spreads and speeds. Examination of the entry points shows that even though those entry points have similar degrees of connectivity in the tweeter relationship graph, i.e. similar number of friends and followers, their effectiveness in assisting information propagation is different. We found that a user's influence in information propagation in microblogging system can be modeled by certain factors. Without losing focus of this work, we will address this in details in future work.

VI. Conclusion

In this paper, we proposed a logistic regression model to predict individual user's retweet decision and constructed a Monte-Carlo simulation framework in order to model how the information is propagated in the twitter network. We collected real trace data from twitter.com using provided API and built a large data set to train and test our logistic regression model. We archived cross-validation accuracy, precision and recall of 93.27%, 73.47% and 40.26%, respectively. We analyzed the model parameters from the trained model and find some of the model parameters align with our hypothesis, but others are counter-intuitive. We also ran simulation study with a small set of users and examined how the information was propagating through this small network. Due to the limited size of the data set, our simulation results might not be practical enough to have real work application at this point.

As an extension of this work, we are planning to collect more data to train our models and run simulations with a large set of users and continue to investigate the issues arises from the individual logistic regression model. Also, as mentioned in previous section, we are proposing a mechanism based on "betweenness" to model user's value in terms of effectiveness in information propagation.

Acknowledgment

The research was support in part by the CyLab Mobility Research Center at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF0910273 from the Army Research Office and Fundamental Research Funds for the Central Universities (China) under Grant 2011YJS005. The authors would also like to thank Dr. Rong Yan from Facebook Inc. and Professor Bo Shen from Beijing Jiaotong University for their valuable comments and suggestions.

References

[1] Haiti earthquake: Twitter updates from the disaster zone, 2010.
[2] A. Click. Iran's Protests : Why Twitter Is the Medium of the Movement. *Time*, pages 4–7, 2009.
[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
[4] D. Gaffney. # iranElection : Quantifying Online Activism. In *Analysis*, 2010.
[5] W. Galuba and K. Aberer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *Conference on Online social networks (WOSN)*, 2010.
[6] J. Golbeck, J. M. Grimes, and A. Rogers. Twitter Use by the U . S . Congress. *Journal of the American Society for Information Science*, 61(8):1612–1621, 2010.
[7] S. Golder. Tweet , Tweet , Retweet : Conversational Aspects of Retweeting on Twitter. In *Sciences-New York*, pages 1–10, 2010.
[8] G. Grosseck and C. Holotescu. CAN WE USE TWITTER FOR EDUCATIONAL ACTIVITIES ? In *4th Scientific Conference eLSE "elearning and Software for Education"*, 2008.
[9] P. J. *New Scientist Magazine*.
[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors. In *International conference on World wide web (WWW)*, 2010.
[11] J. Letierce, A. Passant, S. Decker, and J. G. Breslin. Understanding how Twitter is used to spread scientific messages. In *Web Science Conference*, 2010.
[12] D. A. Shamma and E. F. Churchill. Tweet the Debates Understanding Community Annotation of Uncollected Sources. In *SIGMM workshop on Social media*, pages 3–10, 2009.
[13] K. C. Sia, C. J, K. Hino, Y. Chi, S. Zhu, and B. L. Tseng. *Monitoring RSS Feeds Based on User Browsing Pattern*, pages 161–168. 2007.
[14] 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 *IEEE International Conference on Social Computing (SocialCom)*, 2010.
[15] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging During Two Natural Hazards Events : What Twitter May Contribute to Situational Awareness. In *International conference on Human factors in computing systems (CHI)*, pages 1079–1088, 2010.
[16] D. Zarrella. The Science of ReTweets, 2009.