International Journal of Information Management 35 (2015) 702–711

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Analyzing and predicting news popularity on Twitter

Bo Wu, Haiying Shen*

Department of Electrical and Computer Engineering, Clemson University, United States

ARTICLE INFO

Article history:
Received 29 September 2014
Received in revised form 11 May 2015
Accepted 21 July 2015

Keywords:
Twitter
Social news media
News popularity prediction

ABSTRACT

Twitter is not only a social network, but also an increasingly important news media. In Twitter, retweeting is the most important information propagation mechanism, and supernodes (news medias) that have many followers are the most important information sources. Therefore, it is important to understand the news retweet propagation from supernodes and predict news popularity quickly at the very first few seconds upon publishing. Such understanding and prediction will benefit many applications such as social media management, advertisement and interaction optimization between news medias and followers. In this paper, we identify the characteristics of news propagation from supernodes from the trace data we crawled from Twitter. Based on the characteristics, we build a news popularity prediction model that can predict the final number of retweets of a news tweet very quickly. Through trace-driven experiments, we then validate our prediction model by comparing our predicted popularity and real popularity, and show its superior performance in comparison with the regression prediction model. From the study, we found that the average interaction frequency between the retweeters and the news source is correlated with news popularity. Also, the negative sentiment of news has some correlations with retweet popularity while the positive sentiment of news does not have such obvious correlation.

Published by Elsevier Ltd.

1. Introduction

Predicting news popularity on Twitter is critically important from many aspects. There are many different news agencies on Twitter such as CNN and BBC. Each news agency publishes dozens of news tweets every day and it takes a long time to read all of them. Many of the news portal websites already have the mechanism to recommend the most popular news to users based on the pageviews. However, these recommendation mechanisms are based on the pageviews that have already occurred, so the recommendations are not timely, especially when we consider the time effectiveness of news. Also, it benefits advertisement. Since each news tweet is followed by a URL pointing to a web page, more popular news means more pageviews *in the future* and a higher value for advertisement. The news popularity prediction enables the companies to maximize revenue through differential pricing for access to contents with different popularity.

Instead of obtaining followers by friendship, many supernodes in Twitter obtained many followers contributed by their influence. The most influential Twitter accounts belong to the news agencies. For example, CNNBRK has 8,867,029 followers on September 29th, 2012. 50 percent of the retweets are from supernodes (Shaomei, Jake, Winter, & Duncan, 2011), while previous works paid little attention on it. Thus, in this work, we focus on the top news agencies accounts on Twitter. We study Twitter's news propagation characteristics, based on which, we develop a Twitter news popularity prediction model. Our work makes three main contributions: (1) We collected data from Twitter and analyzed the propagation characteristics of news from both macro- and micro-level perspectives. We propose an algorithm to find the parents of retweeters of a tweet, which are used to reconstruct retweet propagation topology. Here, the parents of retweeters of a tweet are the nodes which retweeters retweet the tweet from. For example, user *A* retweeted a tweet a generated by user *B*. User *C* is *A*'s follower. *C* saw the tweet from *A* and then, retweeted the tweet *A*. Then *A* is the parent of *C* of tweet *A*. We study the distribution of retweets of a tweet over time. We also found that the tweet popularity is correlated with the interaction frequency between the retweeters and the supernode. Further, we discovered that the negative sentiment of news has some correlation with tweet popularity while the positive sentiment of news does not have such obvious correlation; (2) Based on our observed characteristics of retweets, we propose a news popularity prediction model based on news propagation process. The model can predict a news tweet's popularity based on the number of its retweets soon after being published. It can predict (i) the total number of retweets of a tweet only from a supernode, (ii) the number of retweets of a tweet at a certain time after it is published, (iii) the total number of retweets in a certain hop distance from

* Corresponding author.
E-mail addresses: bwu2@clemson.edu (B. Wu), shenh@clemson.edu (H. Shen).

http://dx.doi.org/10.1016/j.ijinfomgt.2015.07.003
0268-4012/Published by Elsevier Ltd.

a supernode, and (iv) the final total number of retweets; (3) We evaluated our popularity prediction model based on the real trace data from Twitter. We found our prediction model is accurate and its predicted results conform with the observed popularity, and it outperforms the regression prediction model in terms of prediction accuracy.

The rest of this paper is organized as follows. Section 2 briefly describes the related work. Section 3 describes our analysis on news propagation characteristics based on real trace. Section 4 presents the design of news popularity prediction model. Section 5 evaluates our model based on the real trace. Finally, Section 6 summarizes the paper with remarks on future work.

2. Related work

Previous study has shown that Twitter is not only a popular social network, but also a popular news media (Kwak, Lee, Park, & Moon, 2010). Previous studies on information propagation on online social networks can be mainly classified to two categories: macro-level and micro-level.

The works in the macro-level category mainly involve the information propagation model and the virus spreading model. Evans and Cheng (2009) measured the characteristics of twitters based on the data from the Sysomos business intelligence company. Cha, Kwak, Rodriguez, Ahn, and Moon (2007) and Cheng, Dale, and Liu (2007) measured the characteristics of YouTube videos. Leskovec, Backstrom, and Kleinberg (2009) developed a framework for tracking popular, short and distinctive phrases online. Wallsten (2008) found strong evidence that the relationship between variables (e.g., audience size, blog discussion) is multidirectional in influencing viral video popularity. Broxton, Interian, Vaver, and Wattenhofer (2011) studied the influence of social networks on the propagation of YouTube videos in different categories. Sakaki, Okazaki, and Matsuo (2010) considered each Twitter user as a sensor and used machine learning method to predict the location of an earthquake. Cha, Haddadi, Benevenuto, and Gummadi (2010) presented an in-depth comparison of influence of three measures (indegree of users, retweets, and mentions) of Twitter users on information propagation. Jansen, Zhang, Sobel, and Chowdury (2009) reported research results investigating Twitter as a form of electronic information propagation platform for sharing consumer opinions concerning brands by analyzing more than 150,000 tweets containing branding comments, sentiments and opinions. Krishnamurthy, Gill, and Arlitt (2008) gathered and identified distinct classes of Twitter users and their behaviors, geographic growth patterns and current size of the network.

Some previous works have discussed the information propagation mechanism from a micro-level. Zou, Towsley, and Gong (2004) provided an email virus propagation model. Weng, Flammini, Vespignani, and Menczer (2012) studied the competition of information diffusion of different tweets by building an agent-based model. Hodas and Lerman (2013) found that position of exposing messages on the user-interface strongly affects social contagion and used this observation to improve the prediction of the temporal dynamics of user behavior. Szabo and Huberman (2008) used Digg and YouTube to model the increasing of votes on and views of content to predict the dynamics of individual submissions from initial data. Lerman and Hogg (2010) and Lerman and Ghosh (2010) further improved the model to predict the popularity of news on Digg, which is most similar to our work. However, their study did not pay attention to the topology reconstruction and inference of network diffusion which we will study in our paper. Sun, Rosenn, Marlow, and Lento (2009) presented an analysis of information diffusion chains in Facebook. Gargi, Lu, Mirrokni, and Yoon (2011) studied large-scale community detection over a real-world graph composed of millions of YouTube videos. Bakshy, Hofman, Mason, and Watts (2011) reconstructed a cascade in Twitter and discussed the influence of users. Unlike these works that paid little attention on the media characteristics of an online social network and equally treat every piece of information, we particularly study the popularity of news from supernodes on Twitter.

Fig. 1. Screenshot of a user's timeline.

Table 1
News agencies in our trace data.

BBCBreaking	BreakingNews	nytimes
NationNow	thenation	ABC
TIME	BBCWorld	HuffingtonPost
Reuters	AP	WSJ
latimes	politico	NewYorker
USATODAY	AmericanExpress	GMA
AJEnglish	NBCNews	NewsHour
Guardiannews	usnews	Slate
759251	nytimesgloba	CBSNewsl
msnbcl	washingtonpost	thedailybeast
428333	cnni	FoxNews

3. Measurement and observation

Most of the top news agencies such as CNN, BBC and New York Times have their own Twitter accounts and millions of followers. For example, CNN has several different kinds of Twitter accounts including CNN, CNN Breaking News, CNN Live and so on. After a user logs into Twitter, the user's *timeline* shows the new tweets since (s)he logged out last time, and then reminds the user the number of new tweets every 90 s on the top of the timeline. The update time is different from user to user since different users log in at different time. The user can click the top icon to obtain the new tweets. Fig. 1 shows a screenshot of a user's timeline from Twitter. The timeline includes both news from news agencies and non-news tweets from their friends. Each tweet indicates its initial publisher with its head-icon on the left. The tweets in a timeline are ordered by their publishing (i.e., creating) time. The more recent published tweet on the top of a timeline has the highest probability to be seen by users according to the normal behaviors of users.

Users can forward any tweets they received to their followers. This behavior is called retweeting. The tweets received by their followers are called retweets and this user is called the *retweeter* of this tweet. A user's followers can see the retweet's initial publisher by the head-icon and their parent retweeter. However, they cannot see the previous parent retweeters beyond their own parent retweeter.

We collected our trace data with the Twitter API. Firstly, we manually identified 33 top news agencies with the most followers in Twitter as listed in Table 1. A news agency's page in Twitter always shows related news agencies. Thus, we started from CNN and used breadth-first search to find almost all news agencies in Twitter and the numbers of their followers. Finally, we sorted the news agencies in descending order of the number of followers and

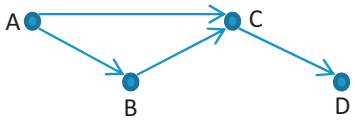


Fig. 2. Retweeting paths.

selected the top 33. Then, we used Twitter stream API to trace their retweets immediately (i.e., catching the publishing time and retweeting time) and recorded the IDs of users who retweeted the news, news content, publishing time of the news and the retweeting time of the users. In this way, we collected our data from January 20th, 2012 to April 20th, 2012 and obtained about 10^4 pieces of original news and more than 10^6 retweets in which CNNBreaking (with 1071 tweets and more than 87,500 retweets) has the most tweets and retweets. Finally, we used Twitter REST API to retrieve the friends list and most recent 20 retweets of each user ID.

To study the popularity of news from supernodes, we need to build the retweet propagation topology for a piece of news, where nodes are connected by retweeting paths. However, as mentioned previously, Twitter API does not provide the previous parent retweeters of a user's retweeter, so it is difficult to identify the entire retweeting path of a retweet.

For example, in Fig. 2, user *D* receives the news from user *C*. User *C* can receive the news from user *A*, from *B*, or from both. Then, it is difficult to determine whether the retweet that user *D* received from user *C* is originally from *A* or *B*. Thus, it is a challenge to determine where the retweet is from for each retweeter for each tweet in our trace. We then build the retweet propagation topology indirectly with two assumptions without the loss of generality listed below.

1. A user retweets a piece of news only when (s)he sees the news at the first time.
2. There is a time delay between when a user sees a piece of news and when the news was created/published.

The users that a user *A* follows are called user *A*'s *friends*. For a given tweet, suppose V is a set of retweet nodes sorted by the retweeting time for a tweet, v_0 is the source node (i.e., supernode) of the news, F_{v_i} is the set of friends of v_i and L_{v_i} is the set of retweet nodes retweeted before node v_i. t_{v_i} is the time that v_i retweeted the news since the publish time t_0. We can get the values of all the aforementioned parameters from our crawled data. We use θ to denote the users' *response delay*, defined as the time elapsed after a user sees a tweet and before (s)he retweets the tweet. We assume that the retweeter of a user is the user's friend, which is true in most cases.

In order to find the retweet topology relationship for the topology construction of a given tweet, for each retweeter v_i for the tweet, we need to find v_i's parent retweeter in the topology that retweets to v_i. Based on the above assumptions, we develop the following algorithm for this purpose:

1. For each v_i, if $L_{v_i} \bigcap F_{v_i} = \emptyset$, v_i retweeted the news from v_0 because none of v_i's friends retweeted the news.
2. If $L_{v_i} \bigcap F_{v_i} \neq \emptyset$, we sort the subset $L_{v_i} \bigcap F_{v_i} \bigcup v_0$ by retweeting time and select the node with the latest retweeting time that is smaller than $t_{v_i} - \theta$ as the parent of node v_i.

Table 2
Fraction of the # of retweets at each hop distance.

Distance in hops	1	2	3	>3
tweet1	77.1%	17.8%	3.7%	1.4%
tweet2	94.1%	4.5%	0.1%	0%
tweet3	98.3%	1.6%	0%	0.1%
tweet4	90.5%	11.8%	1.7%	6%
tweet5	96.2%	3.1%	0%	0.7%
tweet6	87.1%	10.6%	1.1%	1.1%
tweet7	94.7%	4.9%	0.3%	0.1%
tweet8	89.7%	8.9%	0.8%	0.6%

Recall our trace data includes all retweeters for a tweet, by finding each retweeter's parent using this algorithm, we can finally construct the retweet propagation topology of this tweet. In this model, the user's response delay θ is the main factor that might lead to an imprecise topology since the users' response delay may change in a relatively large range due to various reasons. Since the latest retweeting happens much earlier than their children's retweeting time in most situations, θ is very relatively small and negligible. Therefore, this algorithm can help find precise retweeting path in most situations.

Based on the above algorithm, Fig. 3 plots news retweet propagation topologies of 4 randomly selected news tweets from the 400 news tweets from CNN. We can discover that the news spreads in a cascade model and only a small fraction of the news receivers tend to retweet the news. Most of the retweets are propagated to one hop and very few retweets are propagated through 3–4 hops. Also, some retweeters have many followers who retweet the news while most of the retweeters do not have any followers who retweet the news. For instance, in the fifth topology which is for the news "Former Penn State football coach Joe Paterno has died", besides the source node, the node with the most retweeting followers is Dr. Sanjay Gupta who is the CNN Chief Medical Correspondent. Since this news tweet is related to medical treatment and he is the opinion leader in this field, his followers may be more interested in his professional opinion on this tweet and like to retweet this tweet.

In the retweeting paths from a supernode, we call retweeters *m* hop away from the supernode *m-hop retweeters*. Table 2 shows the fraction of retweets retweeted by *l*-hop retweeters ($l = 1, 2, 3, > 3$) of 8 news tweets. We can see from the statistic data that the number of retweets decreases exponentially (e.g., with a rate about 0.23 for tweet1).

The "small-world" principle in a normal social network indicates that a person can use only a few steps to reach any other person in a social network (Newman, Albert-László, & Duncan, 2006). From Fig. 3 and Table 2, we discover that rather than fanning out widely (i.e., reaching many people in very few steps) following this principle, the width of the news propagation from a supernode in Twitter decays exponentially as the propagation hop distance increases and most of the propagation branches end in three hops. This

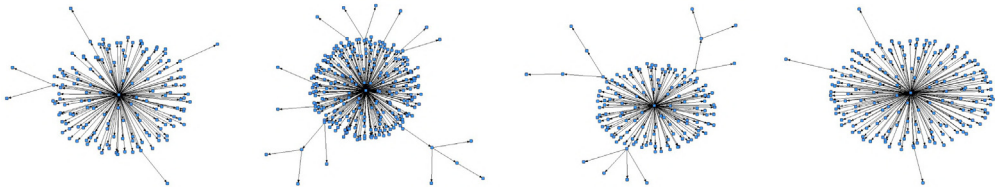


Fig. 3. Retweet propagation topologies for different tweets.

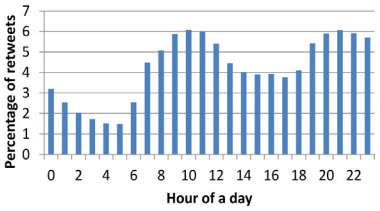
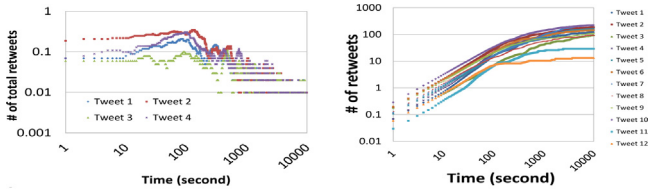


Fig. 4. Distribution of retweets in 24 h.



(a) PDF (b) CDF

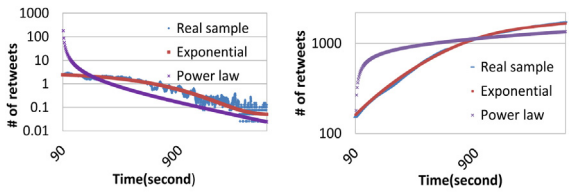
Fig. 5. PDF and CDF of the number of retweets after news publishing.

phenomenon indicates that only a very small fraction of the followers who received the news tend to retweet it.

Fig. 4 shows the distribution of the percentage of the number of retweets generated in each hour over 24 h in a day. We can clearly observe that retweets are concentrated at noon and night, i.e., 7 am–12 am and 7 pm–11 pm. It indicates the retweeting time and hence the usual tweet viewing behavior habits of Twitter users. This observation makes sense since people tend to read news as when they are taking breaks at noon and after work.

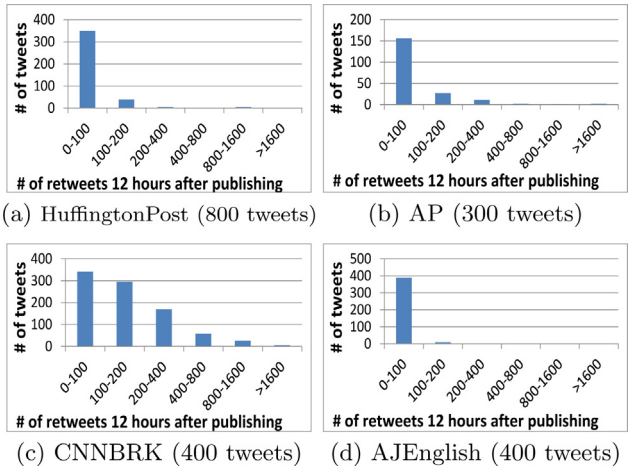
When a tweet is published by a supernode, it accrues retweets from the followers of the supernode, and then accrues retweets from the followers of these first-hop retweeters, and so on. In order to analyze the retweeting behavior (i.e., popularity) over time explicitly, we group the retweets of a tweet into two groups: (1) the direct retweets from the followers of the supernode; and (2) the indirect retweets from the other retweeters in the propagation topology. Fig. 3 and Table 2 indicate that the second group contributes little to news popularity, thus we firstly focus on the first group.

Fig. 5(a) shows the probability distributed function (PDF) of the number of retweets from direct followers of 4 randomly chosen tweets every second since the tweet is published. We can observe that there is a sharp increase at the very first beginning. We conjecture that this is because there is a delay after the news is published and before users see the news due to the 90 s update cycle of user timeline. Also we observed that the number of retweets of each tweet decreases exponentially after the sharp increase. Perhaps, this is because as the time goes on, the popularity and visibility (on the top of the timeline) of the news decrease and fewer people retweet it. We checked other randomly selected 8 tweets and found they also follow this phenomenon. In order to test whether it is an universal phenomenon, we chose 2231 news tweets; each having more than 200 retweets to measure the aggregate statistics. We find that the average time when the number of retweets reaches the top points is 87.31 s after publishing, which confirms our conjecture. Intuitively, the distribution of the number of retweets with the time may have two options: an exponential or power law distribution. In order to verify which option is correct, we draw Fig. 6 to see a real sample comparing with its most fitted exponential and power law distribution. As shown in Fig. 6, the PDF and cumulative distributed function (CDF) of number of retweets is more fitted into an exponential distribution than a power law distribution obviously. Actually, if the number of retweets after 90 s follows a power law distribution, then the number of retweets will tend to be infinite as the time goes. As a fact, the older tweets tend to have



(a) PDF (b) CDF

Fig. 6. The real sample comparing with exponential and power law.



(a) HuffingtonPost (800 tweets) (b) AP (300 tweets)

(c) CNNBRK (400 tweets) (d) AJEnglish (400 tweets)

Fig. 7. Retweeting rate distribution after news publishing of 4 news agencies.

larger number of retweets, which is not true, and the number of retweets tends to converge after a certain time. Therefore, we guess that the number of retweets does not follow a power law distribution. Furthermore, we used Kolmogorov–Smirnov test to verify whether the distribution of the number of retweets after 90 s follows an exponential distribution. The average Kolmogorov–Smirnov statistic for the 2231 news tweets is 0.0264. That is, the hypothesis that it follows exponential distribution has an approximate significance level of 0.2. Thus, it approximately satisfies exponential distribution.

Fig. 5(b) shows the cumulative distribution function (CDF) of the number of retweets from the direct followers of the randomly selected 12 tweets. As each tweet ages, the accumulation of new retweets slows down and saturates finally. We find that the number of retweets of each tweet tends to saturate after about 8000 s. We measured that over 95% of all the retweets in our dataset were created within 8000 s after their tweets are published. The life spans of different tweets of direct retweeters are similar, which confirms that it is mainly determined by the visibility (Lerman & Hogg, 2010) rather than the importance of the tweets. Users usually only view a number of tweets from the top of timelines. A tweet's visibility decreases over time as it moves towards the bottom of the timeline and new tweets occupy the top space.

We then chose the four supernodes with the most retweets from our trace. For each supernode, we grouped all the original tweets based on the total number of each tweet's retweets 12 h after publishing. Fig. 7 shows the distribution of tweets based on this number of retweets. We see that the news popularity at 12 h after publishing varies widely from news to news. A handful of news become extremely popular, accumulating thousands of retweets, while most others obtain about hundreds of retweets. Taking CNNBRK as an example, about one third of tweets have 0–100 retweets, and around 34.7% of tweets have 100–400 retweets. A small percent of tweets have retweets more than 400 and 0.6% tweets have more than 1600 retweets. Such a distribution is characterized by "inequality of popularity". One reason for this characteristic is news

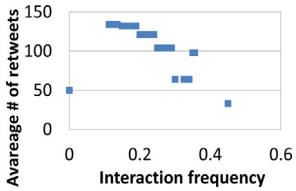


Fig. 8. Number of retweets versus interaction frequency.

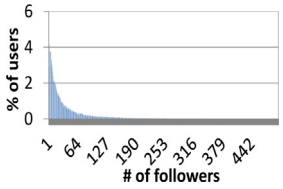
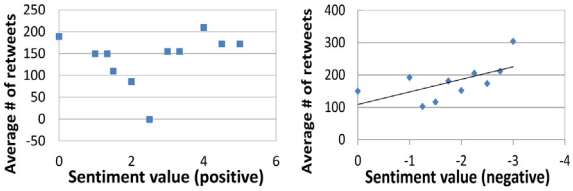


Fig. 9. Distribution of the number of users' followers.



(a) Positive sentiment (b) Negative sentiment

Fig. 10. Relationship between sentiment and final number of retweets (i.e., popularity).

contents. However, it is difficult to predict the news popularity based on new contents automatically given the wide inequality of popularity as shown in Fig. 7. Fortunately, Fig. 5 shows similar twitting (i.e., popularity gaining) pattern for each news, which facilitates popularity prediction based on the popularity gained initially in a very short time after publishing.

Granovetter (1973) indicated the strength of weak ties in information propagation, which can be used to define the strength of two nodes' relationship. Bakshy, Rosenn, Marlow, and Adamic (2012) found that weak ties are influential in the information propagation in Facebook. Next, we analyze the strength of weak ties on news propagation. The strength of ties can be measured in many different ways. We measured it by the interaction frequency between users and the news agency. Specifically, given a tweet from news source *A*, we get all users that retweeted the tweet. Then, for each of these users, we used the Twitter REST API to get the user's latest 20 retweets. The ratio of tweets created by source *A* in the 20 retweets is defined as the interaction frequency of the user with news source *A* for this tweet. The average interaction frequency of these users is defined as the interaction frequency of this tweet with news source *A*.

Fig. 8 plots the average final number of retweets of news tweets versus the interaction frequency of a tweet with the supernode. We see that the number of retweets decreases as the interaction frequency increases. The correlation between the interaction frequency and the final number of retweets is −0.21, which significantly deviates from zero. Thus, a tweet retweeted by users having low interaction frequency with the supernode tends to receive more retweets finally and hence gain higher popularity. Our conjecture of the reason of this observation is that the users with low interaction with a supernode tend to only retweet important news tweets from the supernode, while users with high interaction with the supernode tend to retweet every news tweets. Therefore, if the news tweet is retweeted by more users with low interaction with the news source, then it is a sign that the news tweet may be more important. Therefore, the news tweet may receive more retweets finally. Here, we define the two end nodes of a tie can be people or news sources. A weak tie denotes a tie that the two end nodes do not interact frequently. This result verifies the weak tie theory (Bakshy et al., 2012; Granovetter, 1973) that people tend to receive important information from weak ties rather than from strong ties.

We then show the distribution of the number of the followers of Twitter accounts. We got the data from Evans and Cheng (2009) that provides 11.5 million Twitter accounts collected during 2009. Fig. 9 shows the distribution of these Twitter accounts. We see that it follows power law distribution and that 93.6% of the users have less than 100 followers, while 98% of the users have less than 400 followers. Meanwhile, 1.35% of users have more 500 followers and only 0.68% of more than 1000 followers. This figure implies that a small percent of users are supernodes that own many followers. Non-supernodes constitute the majority users and they have similar number of followers. This observation helps us determine the average number of followers of non-supernodes needed in the prediction model.

Further, we analyze the relationship between the news sentiment and the popularity of news. We chose the AFINN (Pevzner & Tesler, 2011) sentiment database to calculate the sentiment of news on Twitter because AFINN was initially built to analyze Twitter sentiment. Each word in AFINN has a positive sentiment value which presents the positive sentiment or a negative sentiment value which presents the negative sentiment. We define the sentiment value of a tweet as the sum of all the values of the words in the tweet. We classified all tweets to two groups: group 1 contains all tweets with negative values and group 2 contains all tweets with positive values. We plot two figures to show the number of retweets versus the positive sentiment value of group 1 and versus the negative sentiment value of group 2, respectively.

Fig. 10(a) shows the average number of retweets of each group of tweets that have the same positive sentiment value. We see that there are no obvious relationship between news popularity and the positive sentiment value. Fig. 10(b) shows the average number of retweets of each group of tweets that have the same negative sentiment value. We observe that generally the average number of retweets decreases as the negative sentiment value increases. The correlation between negative sentiment value and the average retweets number is 0.38, which indicates that bad news tend to become more popular. For example, the news of "Former Penn State football coach Joe Paterno has died, his family confirms. He was 85." has finally gained 2534 retweets.

4. Popularity prediction model

Though it is difficult to predict if and when an individual user will retweet a tweet, it is possible to find the retweeting probability of each tweet from many users and the retweeting time distribution (i.e., how many retweets occurs at each second). Based on our analysis in Section 3, we propose a general stochastic process-based approach to model user retweeting behavior from an aggregate of human activity and the Twitter user interface information (such as the news update rate and the arrangement of timelines). Our popularity prediction model mimics user retweeting behaviors from both micro- and macro-level perspectives:

1. *Micro-level.* It describes the tweet propagation model from a source node to its one-hop direct followers. This model can predict the total number of retweets only from the source node and the number of retweets at a certain time after being published.
2. *Macro-level.* It describes the cascade model of the whole network for retweet propagation so that we can understand the news propagation process from one node to other nodes, which helps

Table 3
Parameters used in modeling.

	Meaning
t_0	The creation time of the original news
t_1	The time when some one-hop followers receive news
t_2	A time spot after t_1
$n_r(t_1)$	# of direct followers receiving update during $[t_0, t_1]$
$n_v(t_2)$	# of online followers viewed the news during $[t_0, t_2]$
$f_2(t_2)$	Fraction of one-hop retweeters during $[t_1, t_2]$
$M(t_2)$	Total # of direct retweets from a certain node
p_1	Retweeting probability of non-supernodes' followers
l	# of hops that a tweet is retweeted
M_s	# of retweets from a supernode
M_{total}	Final # of retweets (i.e., popularity)
$\mathcal{N}$	Total # of online followers of a supernode in a day
	Parameters in Eq. (8)
q	Retweeting probability of supernodes' followers
$T = 90$ s	Tweet update cycle in timelines
$\overline{N} = 34$	The average number of followers of a user
N	# of a supernode's online followers at a certain time
λ	A parameter in the distribution of user response time
n_u	Average # of upcoming users per second

predict the total number of retweets in a certain hop distance from a supernode and the final total number of retweets.

Table 3 lists all notations we use in modeling.

4.1. Micro-scope popularity prediction

Recall that in Fig. 5(a), the number of retweets exhibits a sharp increase at the very beginning due to the timeline update cycle followed by an exponential distribution. This observation is different from many of the previous works (Changchun, Towsley, & Weibo, 2003; Liben-Nowell & Kleinberg, 2008) which observed that user response time follows an exponential distribution but neglects the initial peak.

Firstly, we consider the sharp increasing at the very beginning in Fig. 5(a). Suppose the timeline update cycle is T, the creation time of the original news is t_0. We use t_1 to denote the time spot when the news is received by a set of one-hop users and t_2 to denote a time spot after t_1. Then, the probability that one user receives the update of the news at time spot t_1 after it is published follows uniform distribution:

$$p(t_1) = \frac{1}{T}, \quad t_1 \in [0, T) \tag{1}$$

We use N to denote the number of currently online direct followers at time t_0, and $n_r(t_1)$ to denote the number of direct followers who have received the update during the time period $[t_0, t_1]$, and $N - n_r(t_1)$ users received the update during the time period $(t_1, T]$. Then, $n_r(t_1)$ equals:

$$n_r(t_1) = \begin{cases} \dfrac{t_1 N}{T}, & t_1 \in [0, T) \\ N, & t_1 \in [T, +\infty) \end{cases} \tag{2}$$

We define response time (denoted by T) as the time from receiving a tweet to reviewing the tweet. After a user receives an update, there exists a response time for the user to view the new update. From Fig. 5(a), which reflects the users' response time distribution, we see if the $[t_0, t_1]$ is large enough, the response time distribution approximately follows an exponential distribution as discovered in Changchun et al. (2003). Accordingly, we consider the user's response time as an exponential distribution with parameter λ as below:

$$g(t; \lambda) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{3}$$

We get the fraction of the set of one-hop direct users who retweet the news during the time interval $[t_1, t_2]$ (denoted by $f_2(t_2)$) by cumulating Formula (3) from time t_1 to t_2:

$$f_2(t_2) = 1 - e^{-\lambda(t_2 - t_1)}, \quad t_2 \in [t_1, +\infty) \tag{4}$$

Based on Formulas (1) and (3), we can obtain the total number of online followers who have viewed the news during the time interval $[t_0, t_2]$, denoted by $n_v(t_2)$. It is the product of the number of users that have received the tweet by t_1 ($n_r(t_1)$) and the probability that a receiver views (responds) to the tweet by t_2($f_2(t_2)$).

$$n_v(t_2) = \begin{cases} \displaystyle\int_0^{t_2} p(t_1)dt_1 \int_{t_1}^{t_2} g(t; \lambda)dt, & t_2 \in [0, T) \\ \displaystyle\int_0^{T} p(t_1)dt_1 \int_{t_1}^{t_2} g(t; \lambda)dt, & t_2 \in [T, +\infty) \end{cases} \tag{5}$$

In addition to the existing online direct users, we also need to consider the users that log in during $[t_0, t_2]$ in order to calculate the total number of direct users who have viewed a news tweet from a source node during $[t_0, t_2]$. We call these users *upcoming users*. In our model, we consider that the retweeting behavior happens independently. This is reasonable because Twitter is a sparse social network (i.e., most of a node's friends and followers do not know each other) especially for the social network topologies of the supernodes. Unlike the social network topologies of normal users that connect with each other through friendship, most of a supernode's followers do not know each other and the retweeters beyond 1-hop distance are not indicated in the timelines. Thus, we could ignore the influence of a retweeter on his/her friends' retweeting behavior.

Suppose that the average number of users who log in per second equals n_u. We use $M(t_2)$ to denote the total number of direct retweets from one certain node. $M(t_2)$ equals the sum of the retweets from existing online users and the retweets from the upcoming users. We use q to denote the average retweeting probability (the ratio of followers that retweet the tweet) of a certain news tweet. Since the upcoming users will check the news once they log in, the number of retweets of a newly received tweet can be calculated by $q \cdot n_u t_2$. Finally, based on Formula (5), we get:

$$M(t_2) = q \cdot n_v(t_2) + q \cdot n_u t_2 = \begin{cases} q \displaystyle\int_0^{t_2} p(t_1)dt_1 \int_{t_1}^{t_2} g(t; \lambda)dt + q n_u t_2, & t_2 \in [0, T) \\ q \displaystyle\int_0^{T} p(t_1)dt_1 \int_{t_1}^{t_2} g(t; \lambda)dt + q n_u t_2, & t_2 \in [T, +\infty) \end{cases} \tag{6}$$

From Formula (6), we derive Formula (7) based on Formulas (2), (3), (4), (5).

$$M(t_2) = \begin{cases} q(N - \dfrac{Nq(1 - e^{-\lambda t_2}(1 + t_2))}{T} + n_u t_2), & t_2 \in [0, T) \\ q(N - \dfrac{Ne^{-\lambda t_2}(e^{T\lambda} - 1)}{T\lambda} + n_u t_2), & t_2 \in [T, +\infty) \end{cases} = q\mathcal{M}_{t_2} \tag{7}$$

where

$$\mathcal{M}_{t_2} = \begin{cases} N - \dfrac{Nq(1 - e^{-\lambda t_2}(1 + t_2))}{T} + n_u t_2, & t_2 \in [0, T) \\ N - \dfrac{Ne^{-\lambda t_2}(e^{T\lambda} - 1)}{T\lambda} + n_u t_2, & t_2 \in [T, +\infty) \end{cases} \tag{8}$$

Predicting the number of retweets from a supernode (M_s). Based on Formula (7), then $M_s = M(t_2 \to +\infty)$ is the number of retweeters of the source node. The parameters of $t_2, q, T, \overline{N}, \lambda, N$ and n_u are needed in Formula (7). Recall that Fig. 5(b) shows that the number of retweets of each tweet tends to saturate after about 8000 s after publishing. We also verified this observation using all tweets in our trace data. Thus, we can use $t_2 = 8000$ when conducting M_s prediction. q is determined based on the number of retweets until the Nth second after news publishing, that is, it is determined at the initial stage of propagation. $T, \overline{N}, \lambda, N$ and n_u are determined offline based on trace data. In Section 4.3, we will introduce how to determine these parameters. Then, M_s can be directly calculated based on Formula (7) at Nth second after news publishing. The prediction of M_s also helps us estimate the final number of retweets of a news tweet in Section 4.2.

Predicting the number of retweets at a certain time after publishing. We can also use Formula (7) to predict the total number of retweets of a given tweet at each second after it is published. Below, we present the details for this prediction. At second 1 after the tweet is published, t_2 in Formula (7) equals 1. Using t_2, q, N values and other parameters listed in Table 3, we can calculate the number of retweets at second 1 using Formula (7), denoted by $M_1 = M(t_2 = 1)$. As the tweet is retweeted in a cascading manner, we then recursively calculate the number of retweets generated at each subsequent second regarding the retweeting node as a root node. The number of retweets at second 2 equals $M_2 = M_1 \cdot M(t_2 = 1)$, the number of retweets at second 3 equals $M_3 = M_1 \cdot M(t_2 = 2) + M_2 \cdot M(t_2 = 1)$, the number of retweets at second i equals $M_i = M_1 \cdot M(t_2 = i-1) + M_2 \cdot M(t_2 = i-2) \cdots M_{i-1} \cdot M(t_2 = 1)$.

4.2. Macro-scope popularity prediction

In the previous section, we introduced a model to predict the number of retweets for a news tweet from a certain node. That is, the number of children of the node in the retweet propagation topology of the tweet. In this section, we present a model to predict the total number of retweets at each hop distance from the source node in the retweet propagation topology, and the total number of nodes in the retweet propagation topology (i.e., popularity).

When a news tweet is retweeted by one node, all of its followers see the news. A retweeting tree consists of a supernode that initially creates a tweet and non-supernodes that receive the tweet and retweet it. The basic retweeting process from these non-supernodes is the same as the process from the supernode except that the numbers of followers of these non-supernodes are different from each other while the number of followers of the single supernode is fixed. Based on the observation from Fig. 9, the retweeting probabilities of all nodes except the supernode's followers are stochastically the same considering the user aggregate retweet behavior from a macro-level. That is, given a set of the followers of a non-supernode, when the size of the sample from the set is large enough, the retweeting probabilities of the nodes in the sample are stochastically the same.

Based on the above analysis, we build a model to simulate the news retweet propagation with the assumptions below:

1. The retweeting probabilities of all nodes except the supernode's followers are stochastically the same (denoted by p_1).
2. The number of the followers of a random user in the stochastic model equals the average number of followers per node in Twitter (denoted by $\overline{N}$).

In the previous work (Shaomei et al., 2011), the authors measured that the average number of the followers of randomly sampled users is 34. We then directly set $\overline{N} = 34$. Then, the propagation process for a news tweet published by a supernode is as below:

1. A news tweet is published by a supernode.
2. The followers of the supernode receive the update and the news is retweeted with a certain probability.
3. The retweet behavior happened recursively as step 2 with a certain probability on the retweeters' followers.

We can deduct from the propagation process that the number of retweets from each hop tends to be a geometric progression based on the assumptions. We present the process of the deduction below. Recall that M_s denotes the number of retweets from the supernode, and $l \to +\infty$ be the number of hops that the tweet is retweeted. The final number of retweets M_{total} equals:

$$M_{total} = M_s + M_s\overline{N}p_1 + M_s\overline{N}^2p_1^2 + \cdots + M_s\overline{N}^{l-1}p_1^{l-1} = \frac{M_s(1-\overline{N}^l p_1^l)}{1-\overline{N}p_1} \quad (9)$$

Assume that $\overline{N}p_1$ is larger than 1, then M_{total} tends to be infinite, which means it tends to approach the total number of nodes in Twitter. This is against our observation that the retweets of a tweet do not cover the entire Twitter social network. Therefore, $\overline{N}p_1$ should be no larger than 1. In this case, the total number of retweets M_{total} equals:

$$M_{total} = \frac{M_s}{1-\overline{N}p_1} \quad (10)$$

from which we derive

$$\overline{N}p_1 = 1 - \frac{M_s}{M_{total}} \quad (11)$$

Predicting the final number of retweets. Previously, we introduced M_s prediction using Formula (7). Based on the observed data $\overline{N}p_1$ from trace and $\overline{N} = 34$, we can determine p_1. For simplicity, we can use q as p_1 since both denote the retweeting probability of a node's followers. Then, based on Formula (10), we can calculate the total final number of retweets of a tweet M_{total} at the initial stage of propagation.

Predicting the number of retweets in a certain hop distance from a supernode. From Formula (9), we can calculate the number of retweets at each hop distance from the supernode given $\overline{N}$, p_1 and M_s. The number of retweets at two-hop distance equals $M_s\overline{N}p_1$, that at three-hop distance equals $M_s\overline{N}^2p_1^2$, and so on. This prediction method can be verified using the example of tweet 1 in Table 2. Based on Formula (11), we get that $\overline{N}p_1 = 1 - 77.1\%$. Then, based on Formula (9), the fraction of the number of retweets in two-hop distance equals $77.1\% \times (1–77.1\%) = 17.8\%$, the fraction of the number of retweets in three-hop distance equals $77.1\% \times (1–77.1\%)^2 = 3.7\%$, the fraction of the number of retweets in four-hop distance equals $77.1\% \times (1–77.1\%)^3 = 1.4\%$ and so on. These prediction values match the real values in the first row of Table 2, which verifies the rationality of our prediction model.

4.3. Parameter determination

In this section, we introduce how we determine the parameters needed in popularity prediction ($q, T, \overline{N}, \lambda, N, n_u$ listed in the bottom part of Table 3). Fig. 4 shows that the retweeting activity of users varies over time. We then discuss the parameter N at different times, which can directly influence the retweeting activity. Since the number of online followers of a news agency cannot be obtained directly, we need to estimate it indirectly. We consider

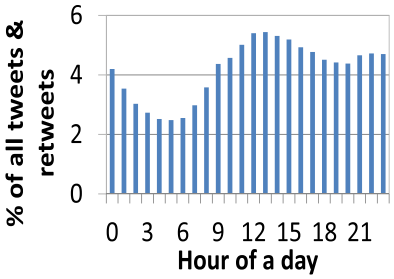


Fig. 11. Tweet activity percentage by hour.

the followers of news agency are distributed among different communities (i.e., students, employers) uniformly in the Twitter social network since the news agencies have little obvious characteristics that attract particular communities. Therefore, we could use the distribution of the number of online users per hour in a day of all users on Twitter as the distribution of the number of online followers per hour in a day of a certain supernode. The former distribution can be approximated by the distribution of the average number of all tweets and retweets published per hour in a day in Twitter as shown in Fig. 11. It is plotted based on data analysis results for a continually long time period from the Sysomos business intelligence company (Evans & Cheng, 2009). Based on this distribution, we can calculate the number of online followers of a supernode at a certain time. For example, if the total number of a supernode's unique online followers in a day is $\mathcal{N}$, then the number of its online followers at the hour 12 is $N = r_t\mathcal{N} = 5\%\mathcal{N}$, where r_t denotes the fraction of all tweets and retweets at time T.

Next, we introduce the determination of q used in Formula (7). In order to more accurately determine q to optimize the prediction results, we estimate q by minimizing the root-mean-square-error (RMSE) difference between the observed number of retweets during the first N seconds and the predicted numbers. Based on Formula (7), this can be expressed by:

$$\min \sum_{i=1}^{n} [q \cdot \mathcal{M}_i - x_i]^2 \tag{12}$$

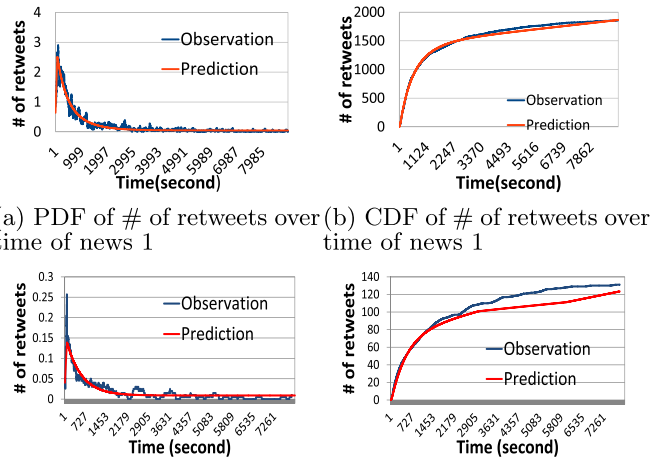
where x_i is the observed number of retweets until the ith second after news publishing. When the formula gets the minimum value, q equals:

$$q = \frac{\sum_{i=1}^{n} (\mathcal{M}_i - x_i)}{\sum_{i=1}^{n} \mathcal{M}_i^2} \tag{13}$$

Based on Formula (13), to calculate the value of q, we need to calculate $\mathcal{M}_i$ and x_i at each second. x_i can be directly retrieved from the initial stage of propagation. To calculate $\mathcal{M}_i$ based on Formula (8), we need to know the other parameters of $\overline{N}$, λ, n_u listed in Table 3. We can directly observe the parameter λ from the trace data. From Fig. 5, we observed that the response time distribution matches an exponential distribution with a parameter λ. The mean of the distribution equals $1/\lambda$. Then, we calculated the mean of the response time of all tweets and finally get $\lambda = 0.12$.

It is difficult to measure the number of total online followers at a certain hour (N) and the number of upcoming users (n_u), we adopted a machine learning method to determine the parameters based on prediction performance. Let us use t_i to denote the publishing time of news i, then $N = r_{t_i}\mathcal{N}$ where r_{t_i} is determined by referring to Fig. 11. Specifically, we use the parameters $\mathcal{N}$ and n_u to form the following objective function:

$$\min \sum_{i=1}^{m} [M(r_{t_i}\mathcal{N}, \lambda, n_u) - x_i]^2, \tag{14}$$



(a) PDF of # of retweets over time of news 1 (b) CDF of # of retweets over time of news 1

(c) PDF of # of retweets over time of news 2 (d) CDF of # of retweets over time of news 2

Fig. 12. The consistency between the predicted number of retweets and the actual number of retweets.

where m is the number of tweets published by a supernode in our trace data, $M(r_{t_i}\mathcal{N}, \lambda, n_u)$ and x_i are the predicted and observed number of the retweets of news i, respectively. We use a gradient descent algorithm to solve the optimization problem in Formula (14). The algorithm loops over all the observations and updates the parameters and finally finds the optimal values of $\mathcal{N}$ and n_u that achieve the objective.

In a nutshell, T, $\overline{N}$, λ, N, n_u are determined offline. After a very short time since news publishing, q is determined using Formula (13). Then, using all these parameters, we can conduct news popularity prediction.

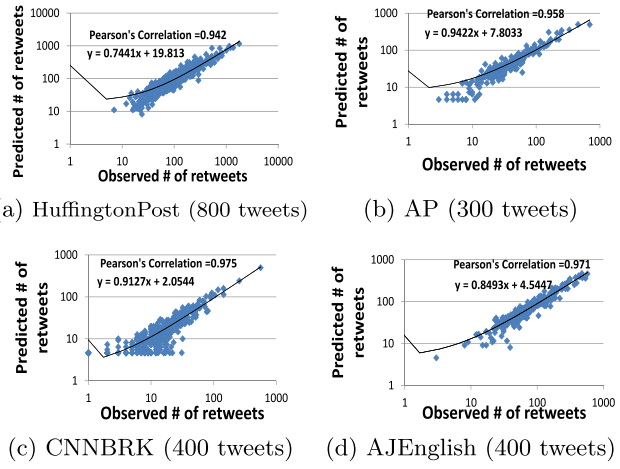
5. Trace-driven prediction performance evaluation

5.1. Model-based prediction

Based on our prediction model introduced in Section 4, parameter q is needed to simulate the retweeting process of a tweet. We use the retweet trace of the first 100 second of each tweet to estimate q of each tweet based on Formula (13). We then use the method introduced in Section 4.1 to predict the number of generated retweets in each second after a tweet's publishing. In the test, we assume p_1 has the same value as q for simplicity. In fact, the direct retweet probability from the followers of the source (q) should be different from the indirected retweet probability (p_1) since the followers of the initial author should be more interested in the retweets from the author. Calculating p_1 offline before prediction would improve the prediction accuracy in our presented prediction results below.

Fig. 12 shows the PDF and CDF of the observed and predicted number of retweets of randomly chosen 2 tweets from the 12 tweets in Fig. 5(b) at each second over time. Both figures show that the predicted results are consistent with the observed actual results. This result verifies that our prediction model can accurately predict the popularity of the tweet over time and our previous assumptions are reasonable.

Next, we verify the correctness of the prediction on the final number of retweets of a tweet. We use the method introduced in Section 4.2 for this prediction. Fig. 13 shows the predicted results and observed results for all news published by 4 news agencies in Fig. 7. We measured that the predicted number of retweets have around 89% correlation with the observed numbers of retweets. Table 4 shows the correlation between prediction and observation and the slopes of the best fit lines shown in Fig. 13 for the 33



(a) HuffingtonPost (800 tweets) (b) AP (300 tweets)

(c) CNNBRK (400 tweets) (d) AJEnglish (400 tweets)

Fig. 13. The prediction results of 4 news agencies with most retweets.

Table 4
Comparison of the two prediction models.

News medias	S_{m1}	C_{m1}	S_{m2}	C_{m2}
BBCBreaking	0.9143	0.9711	0.6415	0.7439
BreakingNews	0.8977	0.9641	0.5937	0.7223
nytimes	0.8721	0.9377	0.6371	0.6975
NationNow	0.9605	0.9713	0.5820	0.7064
thenation	0.8276	0.9440	0.6161	0.7501
ABC	0.8572	0.9227	0.5950	0.6982
TIME	0.9040	0.9512	0.5811	0.9805
BBCWorld	0.8939	0.9737	0.6412	0.7744
HuffingtonPost	0.7441	0.9427	0.6113	0.6802
Reuters	0.8280	0.9766	0.5607	0.7162
AP	0.9422	0.9586	0.5975	0.7139
WSJ	0.9318	0.9309	0.6779	0.7338
latimes	0.8516	0.9145	0.5607	0.7184
politico	0.9196	0.9273	0.6517	0.6990
New Yorker	0.8684	0.9793	0.5674	0.6837
USATODAY	0.9054	0.9481	0.6312	0.7328
AmericanExpress	0.8277	0.9051	0.5929	0.6564
GMA	0.8313	0.9619	0.6024	0.6993
AJEnglish	0.8493	0.9718	0.7155	0.8986
NBCNews	0.7522	0.8728	0.5552	0.7188
NewsHour	0.7325	0.9167	0.5744	0.7434
Guardiannews	0.8521	0.9795	0.8359	0.8987
usnews	0.8981	0.9756	0.6046	0.7020
Slate	0.7747	0.9194	0.6275	0.7505
CNN	0.9127	0.9753	0.5886	0.6991
nytimesglobal	0.9150	0.9645	0.6272	0.7252
CBSNewsl	0.8864	0.9562	0.6027	0.7124
msnbc	0.6552	0.8070	0.5646	0.6788
washingtonpost	0.8643	0.9530	0.6250	0.7397
thedailybeast	0.9416	0.9209	0.8111	0.9334
cnnbrk	0.9106	0.9735	0.5858	0.6633
cnni	0.8730	0.8649	0.6324	0.7477
FoxNews	0.8285	0.9533	0.5902	0.7030
Average	0.8613	0.9419	0.6206	0.7400

news medias in our dataset. S_{m1} denotes the slope and C_{m1} denotes the correlation of our prediction model. Both correlation and slope reflect the prediction accuracy and range from −1 to 1. A value closer to 1 means a higher prediction accuracy. We see that both C_{m1} and S_{m1} values are very close to 1, with the average values equal to 0.9420 and 0.8613, respectively. The results indicate that our prediction model can use the very early popularity observations to provide an approximately accurate prediction of the final number of retweets for a news tweet.

Section 4.3 introduced a method to determine the value of q. Using popularity observation from a longer time since news publishing will lead to more accurate estimation of q but longer delay, and vice versa. A more accurate estimation of q in turn leads to more accurate prediction of news popularity. We then test the effects of the length of time period used to determine q on the final prediction correlation with the observed data. Once a tweet is published, we used the observed data from the first 100 s, 200 s and so on until 2000 s to conduction prediction. We conducted prediction on all news of CNNBRK, and then calculate the prediction correlation with the observed data. The results are reported in Fig. 14. The figure shows that the observations from a longer time period for q prediction lead to prediction correlation closer to 1, as more information generates more accurate q prediction. Thus, we should balance the prediction accuracy and the delay in prediction in order to guarantee the timely and precise prediction.

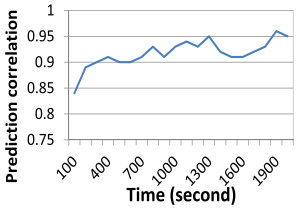


Fig. 14. Time used to estimate q versus prediction correlation.

5.2. Comparison to regression prediction model

Based on the previous analysis in Fig. 5, it can be clearly observed that the number of retweets tends to decrease to zero over time, and the decreasing tendency approximately fits an exponential distribution (Changchun et al., 2003; Liben-Nowell & Kleinberg, 2008). Therefore, we choose exponential regression prediction model and use its standard exponential regression function for popularity prediction:

$$y = ae^{b\hat{t}} \tag{15}$$

where $\hat{t}$ denotes the time period since the news is published, y denotes the number of retweets generated during time $(\hat{t}-1, \hat{t})$, and A and B are the regression parameters. We determine the parameters of A and B using the retweet data during 200 s after publishing, and then use the regression prediction model to predict the final number of retweets of each news tweet. We use time period twice longer for parameter estimation than in our model because the retweeting peak in the first 90 s does not satisfy an exponential distribution and we need a longer time period to more accurately determine the parameters.

Table 4 shows the comparison results of our prediction model and the regression model, where S_{m2} and C_{m2} denote the slope and correlation of the regression model, respectively. We see that S_{m1} and C_{m1} are much closer to 1 than S_{m2} and C_{m2}, respectively. The average correlation and average slope of the regression model are 0.7400 and 0.6206, which are much lower than those of our model (0.9420 and 0.8613). These results show that our model predicts news Twitter popularity more accurately than the regression model, even though it only uses half time to determine the parameters.

5.3. Outliers analysis

We selected ten news tweets with the biggest prediction errors in our trace data and calculated their retweeting rate distributions. We noticed that the retweeting rate decreases over time. Fig. 15 shows the retweet rate distribution of one outlier. We see that the retweeting rate decreases sharply during a short time after news publishing, which is abnormal. This may be caused by the reason that our model does not consider the decreasing visibility caused

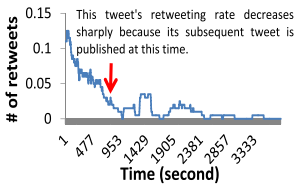


Fig. 15. The retweeting rate of an outlier.

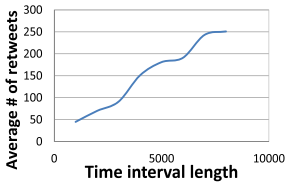


Fig. 16. Popularity versus publishing interval.

by upcoming news in the timelines and assume that the tweet is always visible.

If tweet B is the subsequent tweet after tweet A, and B appears x time period after tweet A's appearance, we call x as tweet A's time interval before its subsequent tweet. We found that the average time interval between the publishing time of these ten news tweets and their subsequent news is 43 min, which is much shorter than our observed average time interval (i.e., 114 min) between the publishing time of news. Also, we noticed that their retweeting rate decreases sharply after their subsequent news was published. The early appearance of the subsequent tweets should be the reason for the retweeting rate decrease. We then analyze the real trace to verify our conjecture. We classified the 400 tweets based on their time interval before subsequent tweets and then calculated the average number of final retweets of each group. Fig. 16 shows the average number of final retweets of each group versus each group's time interval before the subsequent news publishing. We see that the average final number of retweets increases as the time interval increases. The result illustrates that followers tend to retweet the latest published news. This result verifies our conjecture that the time interval before the subsequent news publishing affects the final number of retweets and retweeting rate of a tweet. In our future work, we will consider the decreasing visibility of tweets over time into our prediction model.

6. Conclusion

Twitter is not only an online social network, but also an increasingly important news media. Retweeting is a most important information propagation mechanism on Twitter. Supernodes are critical information source on Twitter. Therefore, understanding the news retweeting propagation from supernodes is very important for many purposes such as information management, advertisement, and social media management. In this paper, we first measured the news propagation characteristics on supernodes. Based on the characteristics, we built a news tweet popularity prediction model from both stochastic micro-level and macro-level perspectives. It can predict the total number of retweets of a tweet (i) only from a supernode, (ii) at a certain time after it is published, (iii) in a certain hop distance from a supernode, and (iv) when it saturates finally. Our trace-driven experimental results verify the high prediction accuracy of our prediction model and its superior performance in comparison with a regression prediction model. In addition, we also show the reasons for outliers in prediction. In the future, we will incorporate the news visibility consideration into the prediction model, and explore a method to more quickly estimate p_1. Also, we will evaluate and consider the influence of different factors (e.g., the time in a day, interval between tweets, sentiment, tweet length) on the tweet popularity, and analyze the retweet robot behaviors.

Acknowledgements

This research was supported in part by U.S. NSF grants NSF-1404981, IIS-1354123, CNS-1254006, CNS-1249603, Microsoft Research Faculty Fellowship 8300751.

References

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proc. of WSDM*.
Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proc. of WWW*. ACM.
Broxton, T., Interian, Y., Vaver, J., & Wattenhofer, M. (2011). *Catching a viral video*. JIIS.
Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proc. of AAAI*.
Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., & Moon, S. B. (2007). I tube, you tube, everybody tubes: Analyzing the world's larges user generated content video system. In *Proc. of IMC*.
Changchun, Z., Towsley, D., & Weibo, G. (2003). *Email virus propagation modeling and analysis*. Technical report TR-CSE-03-04.
Cheng, X., Dale, C., & Liu, J. (2007). Understanding the characteristics of internet short video sharing: YouTube as a case study. In *Proc. of SIGCOMM Conference on Internet Measurement* (p. 28).
Evans, M., & Cheng, A. (2009). *An in-depth look inside the Twitter world*. Sysmos.
Gargi, U., Lu, W., Mirrokni, V., & Yoon, S. (2011). Large-scale community detection on YouTube for Topic Discovery and Exploration. In *Proc. of AAAI Conference on Weblogs and Social Media*.
Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology*, *78*(6), 1360–1380.
Hodas, N. O., & Lerman, K. (2013). *The simple rules of social contagion*. CoRR., abs/1308.5015.
Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.
Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about Twitter. In *WOSP'08: Proceedings of the first workshop on online social networks* (pp. 19–24). ACM.
Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proc. of WWW* (pp. 591–600). ACM.
Lerman, K., & Ghosh, R. (2010). *Information contagion: An empirical study of the spread of news on Digg and Twitter social networks*. CoRR., abs/1003.2664.
Lerman, K., & Hogg, T. (2010). *Using a model of social dynamics to predict popularity of news*. CoRR., abs/1004.5354.
Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD*.
Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using internet chain-letter data. *PNAS*, *105*(12), 4633–4638.
Newman, M., Albert-László, B., & Duncan, J. W. (2006). *The structure and dynamics of networks*. NJ: Princeton University Press.
Pevzner, P., & Tesler, G. (2011). AFFIN. *Informatics and Mathematical Modelling*.
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proc. of WWW* (pp. 851–860). ACM.
Shaomei, W., Jake, M. H., Winter, A. M., & Duncan, J. W. (2011). Who says what to whom on Twitter. In *Proc. of WWW* (pp. 705–714). ACM.
Sun, E., Rosenn, I., Marlow, C., & Lento, T. (2009). Gesundheit! Modeling contagion through Facebook news feed. In *Proc. of ISWSM*.
Szabo, G., & Huberman, B. A. (2008). *Predicting the popularity of online content*. CoRR., abs/0811.0405.
Wallsten, K. (2008). Yes we can: How online viewership, blog discussion and mainstream media coverage produced a viral video phenomenon. In *Presented to the annual meeting of the American Political Science Association* Boston, MA.
Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, *2*(335).
Zou, C. C., Towsley, D., & Gong, W. (2004). *Email virus propagation modeling and analysis*. Technical report.