DRAFT

# Towards Modeling Popularity of Microblogs

**Haixin MA[1], Weining QIAN[1*], Fan XIA[1], Xiaofeng HE[1], Jun XU[2], Aoying ZHOU[1]**

[1]Institute of Massive Computing, East China Normal University, Shanghai 200062, China
[2]School of Computer Science, Georgia Institute of Technology, USA

**Abstract** As one kind of social media, microblogs are widely used for sensing the real-world. The popularity of microblogs is an important measurement for evaluation of the influencial of pieces of information. The models and modeling techniques for popularity of microblogs are studied in this paper. A huge data set based on Sina Weibo, one of the most popular microblogging service, is used in the study. First, two different types of popularity, namely *number of retweets* and *number of possible views* are defined, while their relationships are discussed. Then, the temporal dynamics, including lifecycles and tipping-points, of tweets' popularity are studied. For modeling the temporal dynamics, a piece-wise sigmoid model is used. Empirical studies show the effectiveness of our modeling methods.

## 1 Introduction

Along with the development of Web 2.0 applications, social media services have attracted many users and become their hands-on toolkits for recording life, sharing ideas, and social networking. Though social media services are essentially web or mobile applications and services, they combine user-generated content and social networks together, so that information can be created, transmitted, transformed, and consumed in the cyberspace.

Microblogging is one of the most important kind of social media services. Microblogging services often support users to share their information publicly through Web or mobile applications. Pieces of information, or called *tweets*[1)], shared are usually short. A user may further *retweet*, or called *repost* a tweet from others. Thus, data from microblogging services are often used for sensing users' collective behaviors [23]. Popular microblogging services including Twitter[2)], Sina Weibo[3)], Tumblr[4)], and etc.

Popularity is an important quantative measurement for studying the influential of microblogs and authors. Modeling the evolution of popularity for a single tweet or a set of related tweets is an essential problem for better understanding the collective behaviors in social media. Furthermore, the model can be used for predicating popularity of microblogs. Thus, the model can be used in applications such as online advertisement and recommendation.

We study the problem of popularity modeling for a popular microblogging service in China, namely Sina Weibo, in this paper. Mathematical models that capture distributions and dynamics of popularity for single and group of pieces of information are constructed. Furthermore, algorithms for estimating parameters of those models are provided. Our contributions are summarized as follows.

- Popularity of microblogs are formalized in this paper. We argue that the popularity should be measured by two different measurements, namely *number of retweets*, denoated as *#retweet*, and *number of possible view*, denoted as *#pv*. The former one is used to measure the users' involvement for spreading information, while the latter is used to measure how many users may read the information. We show that though these two measurements are related, they are different aspects of the popularity in social media.

E-mail: *wnqian@sei.ecnu.edu.cn* (correspondance author)

[1)] We follow the terms used by Twitter (`http://twitter.com`) in presentation.

[2)] `http://twitter.com/`

[3)] `http://weibo.com/`

[4)] `http://tumblr.com/`

- Lifecycles and tipping points of tweets are studied based on a huge real-life social media data set collected from Sina Weibo service. Models for temporal dynamics of popularity are presented, while algorithms for estimating parameters are provided. Experimental results show the precision of the provided models and parameter estimating method.
- Extensive study and discussion on popularity of microblogs are provided. The factors that affects the popularity are analyzed. Empirical study for comparison between Sina Weibo and Twitter shows how mechanisms of services affect the popularity measurements.

The rest part of the paper is organized as follows. In Section 2, we provides a formal description of microblogging services. The method that how the data set was collected is introduced. Basic statistics of the data set are also provided. Measurements of popularity are introduced in Section 3. The relationships between measurements are analyzed based on the Sina Weibo data set. In Section 4, the temporal dynamics of tweets are analyzed. The dynamics of the popularity is modeled in Section 5, while the algorithm for parameter estimation is provided. Various aspects that related to popularity modeling are discussed in Section 6. The factors that affects the popularity are analyzed. After the related work introduced in Section 7, finally, Section 8 is devoted for concluding remarks.

## 2 Background

### 2.1 Overview of microblog services

To simplify the discussion, we follow the terms used by Twitter. A *tweet*, denoted by a triple $< t_i, u_j, c >$ is a piece of information with content $c$ shared by *user* $u_j$ at timestamp $t_i$. The content of a tweet is usually short. Twitter, for example, requires that each tweet is not longer than 140 charactors. And so is Sina Weibo. A tweet may contain URL links in its content, each of which may link to a web page, a picture, a video, or other type of multimedia resource. A tweet may also contain one or more *hashtags*. A hashtag is a tag annotated by the author of a tweet. In English microblogging services, a hashtag is usually denoted by a term without space characters following a '#'. In Chinese microblogging services, it is usually denoted by a term between a pair of '#'. Though there is no standard protocol on how to use hashtags, they are often used to annotate the topics related to the tweet. Related tweets tend to have identical or similar hashtags.

(a) Retweet in Twitter.

(b) Retweet in Sina Weibo.

**Fig. 1** Difference of retweet in Twitter and Sina Weibo.

A user may *follow* other users. The followship network is a directed graph $G : (V, E)$, in which $V$ is the set of users, and $E$ is a subset of $V \times V$. $(u, v) \in E$ means that user $u$ is following user $v$. In many real-life applications, every user in default subscribes his own messages, i.e. $\forall u, (u, u) \in E$. We denote all $v_i$'s followers as $F_i$, i.e. $F_i = \{u|(u, v_i) \in E\}$.

A user may *retweet* a tweet from another user. The *retweet* is a tweet $< t_i, u, c >$ with a pointer in its content $c$ to another tweet $< t_j, v, c' >$, which may also be a retweet. Aparently, $t_j < t_i$, which means the retweeted one must appear before the retweet. By tracing the *retweeting path*, i.e. follow the pointers in content, an *original tweet* that is the tweet not pointing to any other ones can be found. From an original tweet, a *retweet tree* can be constructed. If a tweet $T$ is a *direct retweet* of $T'$, it is denoted as $T \rightarrow T'$. *Indirect retweet*, denoted by $\mapsto$, is defined recursively as:

- If there exists a $T''$, such that $T \rightarrow T''$, and $T'' \rightarrow T'$, then $T \mapsto T'$;
- If there exists a $T''$, such that $T \mapsto T''$, and $T'' \rightarrow T'$, then $T \mapsto T'$.

The relationship of indirect retweet is transitive, which means if $T_1 \mapsto T_2$, and $T_2 \mapsto T_3$, then $T_1 \mapsto T_3$.

Microblogging services use different retweet mechanism. In Twitter, for example, users can only share the a tweet by retweeting without modifying the original tweet or adding any information to it. In Sina Weibo, a user may not modify the original tweet but attach more information. Furthermore, a user in Sina Weibo may modify information from

other users' in the retweeting path in his/her own retweet. Fig. 1 shows the difference retweet mechanisms in Twitter and Sina Weibo. In Fig. 1 (b), the original tweet is shown with gray background. the content after '//' in the retweet is actually from a user with ID '@李优liyou'. The retweet attaches a smiley face to the message. All content attached in a retweet, whether it comes from the retweeting author or others in the retweeting path, should be no longer than the required length, i.e. 140 characters. Some Twitter clients introduce non-native retweet mechanism that are similar to Sina Weibo's.

### 2.2 Data set

The data we use in this study are collected from Sina Weibo, one of the most popular microblogging services in China. A distribued crawler has developed to crawl the data set. The crawler uses APIs provided by Sina Weibo to crawl the social network, tweets, and user profiles. The crawling procedure is summarized in Fig. 2.

Our crawler works from 32 seed users. A breadth-first search is adopted from those seed users in the followship graph, which leads to 1.7 million users, which are called *core users* in the rest of this paper. The crawler further retrieves top 5000 followers[5] of the core users, which leads to 128,643,501 users in our user list and 1,397,243,218 followship relationships. All users and part of their followship network crawled form the social network we used.

Core users' tweets before the end of January 2012 are collected by the crawler. The earliest tweet we collected has a timestamp in August 2009, which is the month Sina Weibo started to provide services. Totally 649 million tweets are collected. One tweet contains information of the author, the full text, the published time and the platform where the tweet published. If the tweet is retweeted, the tweet also includes a pointer to the original tweet.

In addition to the social network and tweets, profiles from 14.7 million users are also collected, which include their public information on nick name, register time, web page URL, blog URL, location, a brief statement of him/her-self, number of followees, number of followers, number of tweets, number of tweets added into his/her favorites list, gender, certification, education experience and work experience.[6]

It should be noted that the items in the data set are crawled at different time. Therefore, the items are not synchronized. For example, in the data set, a user $u$ may have retweeted a tweet from another user $v$. At the time it is retweeted, user $u$ may be following user $v$. However, at some timestamp afterward, $u$ may have canceled the followship. Thus, in our data set, it is possible that $(u, v) \notin E$.

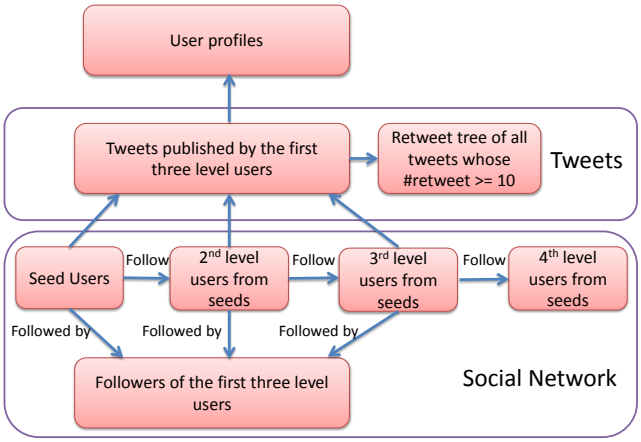



**Fig. 2** The crawling process of the data set used in this paper.

Though the data set is neither complete nor synchronized, it is sufficiently large to dipict the users' collective behavior on Sina Weibo. Fig. 3 shows the distribution of activities over hours in each day and days in each week.

## 3 Measuring Tweet Popularity

### 3.1 Hashtags, retweets, and possible views

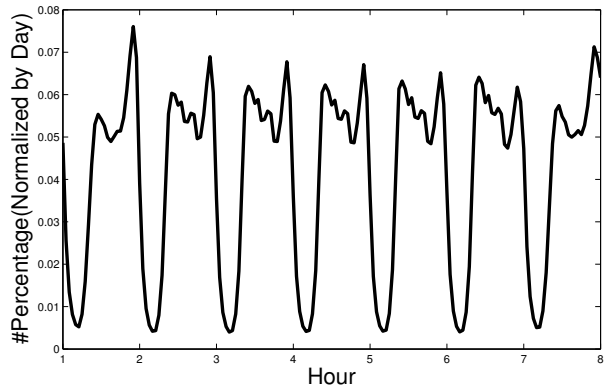
There are several different ways to measure the popularity of single tweet or a set of related tweets. Hashtags and retweets are two most important clues for measuring popularities.

Intuitively, tweets with identical hashtages are highly related. Furthermore, consecutive appearance of a hashtag in tweets of two users with following relationship often implies that the followee influenced the follower. Thus, the frequency of a hashtag, denoted by $\#hashtag$ can be treated as a measurement of popularity of those related tweets.

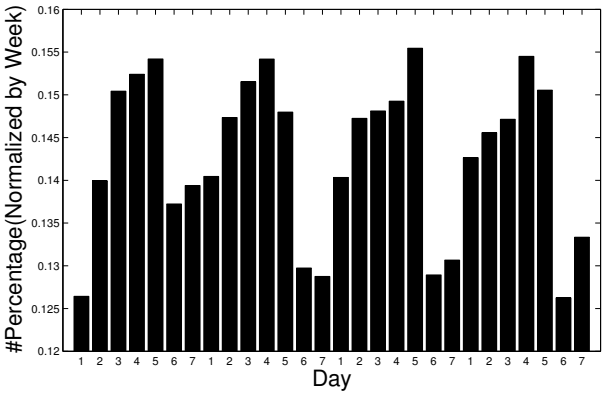
Though hashtags are usually accurate due to its folksonomy characteristics, the information propagation network implied by hashtags is often imprecise and sometimes ambiguous. For example, in Fig. 4, user $A$ follows both user $B$ and $C$, while user $B$ also follows user $C$. If a hashtag appears in tweets of all three users in the order of $C$, $B$, and $A$ within a short time period, it is difficult to make the judgement that it implies $A$ is influenced by $B$ only, or by $C$ only, or by both $B$ and $C$, or none. The problem may become even more complicated that the influential model should depend on both the

[5] The number 5000 is limited by Sina Weibo API.

[6] Profile information collected are provided and chosen to be public by users themselves.




(a) Distribution of activities over hours in 7 days (a week).




(b) Distribution of activities over days in 4 weeks.

**Fig. 3** Distribution of activities over time.

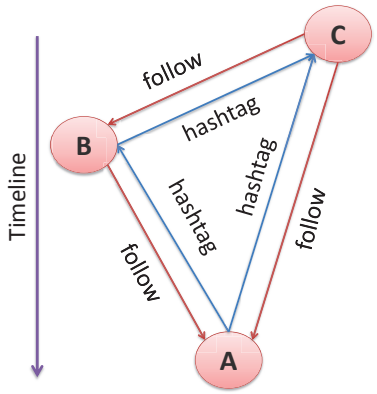



**Fig. 4** Tracing hashtag to reconstruct information propagation network.

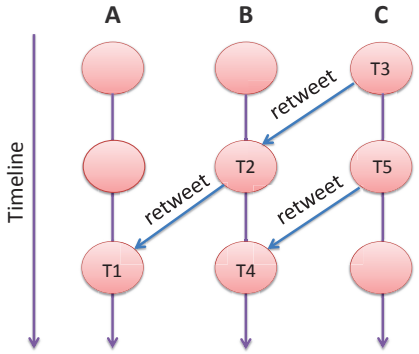



**Fig. 5** Tracing retweet cascades to reconstruct information network

content of the hashtag and the relationships among $A$, $B$, and $C$.

A second clue for measuring popularity of tweets is the retweet actions. Different to hashtags, retweets imply explicit information propagation networks. Thus, the number of retweets, denoted by #*retweet*, of one or a set of tweets, is an important popularity measurement of microblogs.

Fig. 5 shows the behaviors of three users $A$, $B$, and $C$ in a short time period. It is obvious that user $A$'s retweet at time $T1$ is an action influenced by user $B$'s retweet at time $T2$, which is further an action influenced by user $C$'s original tweet at time $T3$. Thus, following the retweeting path, we can reconstruct the information propagation network without any ambiguity.

Another advantage of using retweets to reconstruct the information propagation network is that the retweeting path may not necessarily be embedded in the followship network. A user may retweet a tweet because of occational browsing the global timeline, or recommendation by microblogging services.

When the retweet operation does not allow to add comments, retweets may also lose some kind of information in propagation, since a retweet may mean the retweeter want the followers to see the information, or the retweeter agrees with that piece of information, or he/she does not agree.

In Twitter, where retweet mechanism does not allow to add information, a user may choose not to use retweet when he/she wants to comment on a tweet. At that time, comments plus a hyperlink and/or a hashtag may be used.

A third popularity measurement on microblogs is number of *possible views*, denoted by #$pv$. Intuitively, a tweet from a user with millions of followers is much more influential than that from a user with tens of followers. Furthermore, a retweet by a popular user, who has many followers, may increase the popularity of the original tweet. Thus, we define the number of possible view of a tweet as

$$\#pv(T) = \sum_{\forall T' <t,u,c>, T' \mapsto T} \|F_u\|$$

It should be noted that a user's followers may be counted several times if he/she retweet several times in a retweeting

path.

### 3.2 Relationships of measurements

The three measurements, i.e. #*hashtag*, #*retweet*, and #*pv*, are different yet related. The distribution of #*hashtag* and #*retweet* in Sina Weibo dataset are shown in Fig. 6.

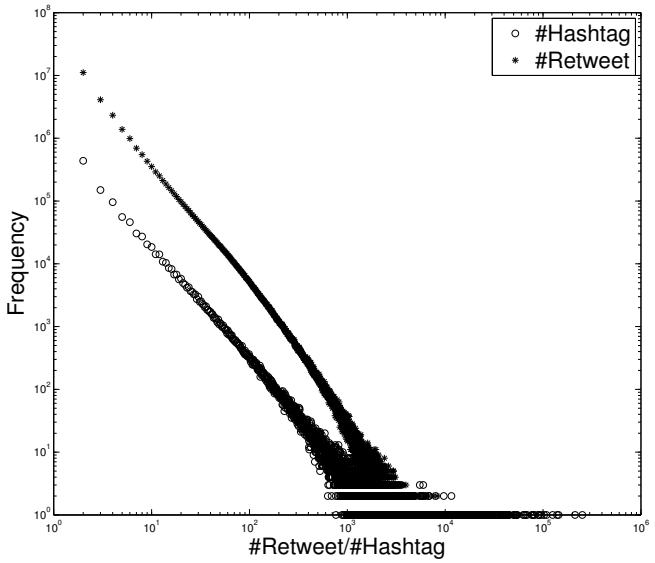



**Fig. 6** The distributions of #Hashtag and #Retweet in Sina Weibo dataset.

The log-log scale figure suggests that both #*hashtag* and #*retweet* are highly skewed. Power-law distribution, which has been proved successful in indicating the skewness in content popularity, is used to model the distributions. In power-law distribution, the attribute of an object, e.g. its size $x$, has the frequency $\alpha x^{\kappa}$, where $\alpha$ is a constant and $\kappa$ is the distribution exponent. The distributions of #*hashtag* and #*retweet* can be modeled by power-law distribution perfectly with $\kappa = -1.5$ and $\kappa = -1.9$ respectively, which is shown in Fig. 7. The coefficient of determination for the models with the data [5], denoted by $R^2$, both exceeds 0.99 in 95% confidence bounds.

Fig. 6 also suggests that retweets are more common in Sina Weibo than hashtags. About 51 percent tweets are retweeted tweets, which is far larger than tweets containing hashtag which is just 4 percent. It should be noted that the wide-use of retweet in Sina Weibo is highly related to the design of its mechanism. For comparison, in Twitter, based on the data set from SNAP[7] on 467 million tweets from 20 million users covering a period from June 1 2009 to December 31 2009 [10], only 15 percent tweets are retweeted.

The information propagation networks reconstructed from hashtags and retweets verify the conclusion. In Sina Weibo, about 4.2 million retweet trees have size larger than 10, while only 0.2 million information propagation trees via hashtag have size larger than 10.

The distribution of tweets over different #*retweet* is shown in Fig. 8. It is shown that 93.8% retweeted tweets have #*retweet* less than 10 (the blue columns). The red columns in the figure show the distribution of $\sum$#*retweet* over all retweeted tweets. It is shown that the impact of tweets with #*retweet* between 10 and 1000, is almost twice as that of 93.8% tweets with very small #*retweet*. It implies that tweets with #*retweet* larger than 10 should be pay more attention than less popular ones.

Since our study is mainly based on Sina Weibo dataset, we focus on #*retweet* instead of #*hashtag* of popular tweets, whose #*retweet* is larger than 10, in the rest of the paper.

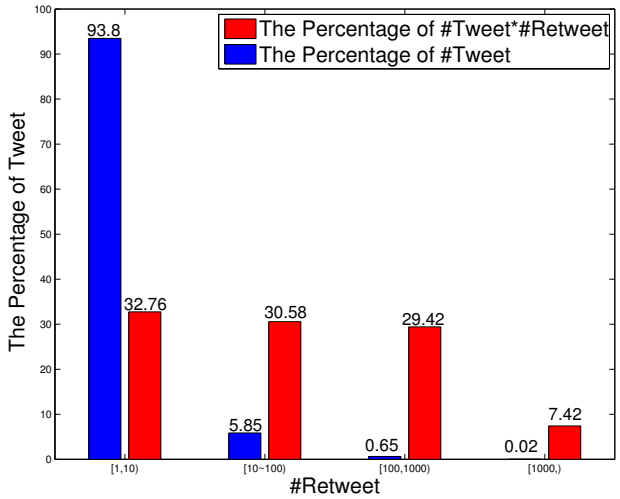
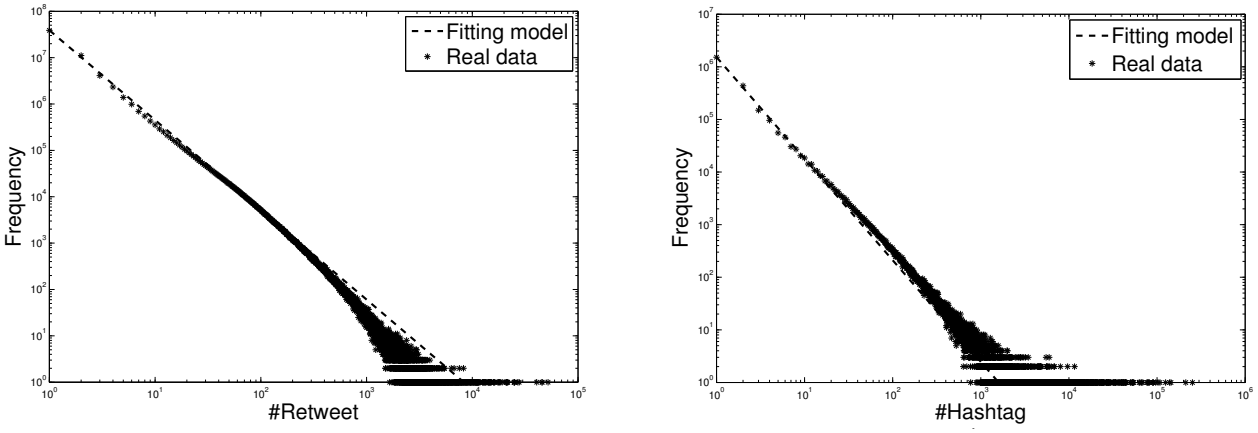



**Fig. 8** The distribution of retweeted tweets and sum of #*retweet* over different #*retweet* regions.

The plotting graph of #*pv* and #*retweet* is shown in Fig. 9. It is shown that tweets with hight #*retweet* tend to have high #*pv*. However, a close look at their correlationships, as it is shown in Table 1, shows that they are only weakly correlated for tweets with #*retweet* larger than 1000. For other tweets, they are not correlated. Recall that tweets with #*retweet* between 10 and 1000 have more impact than other ones, which is shown by Fig. 8 red columns, it can be concluded that #*pv* should be treated as a separate measurements for popularity of microblogs.

[7] http://snap.stanford.edu/data/twitter7.html.




(a) Modeling the distribution of #*retweet*.

(b) Modeling the distribution of #*hashtag*..

**Fig. 7** Modeling distributions of #*retweet* and #hashtag.

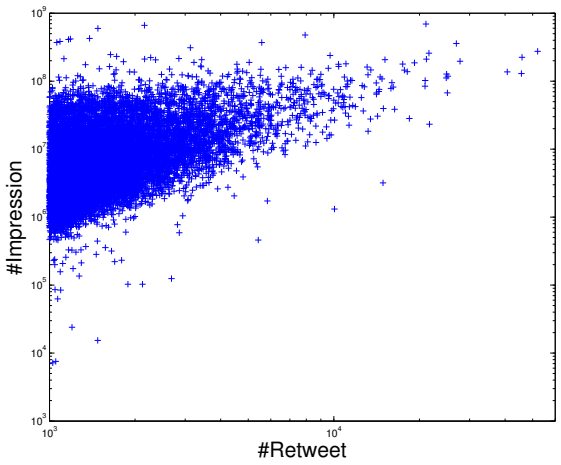



**Fig. 9** Correlationships between #*pv* and #*retweet*.

**Table 1** The correlationship between #*pv* and #*retweet*.

| #*retweet* | $R^2$ |
|---|---|
| [10, 100) | 0.11 |
| [100, 1000) | 0.21 |
| [1000, ∞) | 0.45 |

The number of followers of a spectific user $u$, i.e. $||F_u||$, detnoted as $\#followers$, is often used in Sina Weibo for measure the importance of a user. However, by plotting its relationships with #*pv* and #*retweet*, as it is shown in Fig. 10, it is shown that there is loose relationships between $\#followers$ and other two popularity measurements. Tweets from users with same $\#followers$ may have #*retweet* values with a difference of two orders of magnitude, and #*pv* values with a difference of four orders of magnitude. Since $\#followers$ is a measurement with users instead of tweets, we argue that it is not suitable for measuring popularity of microblogs.

## 4 Temporal Dynamics of Tweet Popularity

We now discuss the time dynamics of popularity for tweets. We start by looking at the life-cycle of a tweet.

### 4.1 The life-cycle of a tweet

The life-cycle of a tweet begins with its creatation and ends at the time of the last retweet. Intuitively, if a tweet is not retweeted anymore, it will be pushed out of the first page of anyone's timeline, thus become forgotten, i.e. go to the end of its life-cycle. However, in real-life microblogging services, it is more complicated that some tweets may occationally be retweeted after a long period of time due to spamming, robots, or other reasons. Therefore, we use the time taken for the first 80% retweets as a measurement for life-cycle of a tweet, which is called *80-lifecycle*.

The distribution of 80-lifecycle on tweets with #*retweet* larger than 1000 is shown in Fig. 11. About 59% popular tweets have completed their 80% retweet within 24 hours, while 76% popular tweets have completed this within 48 hours.

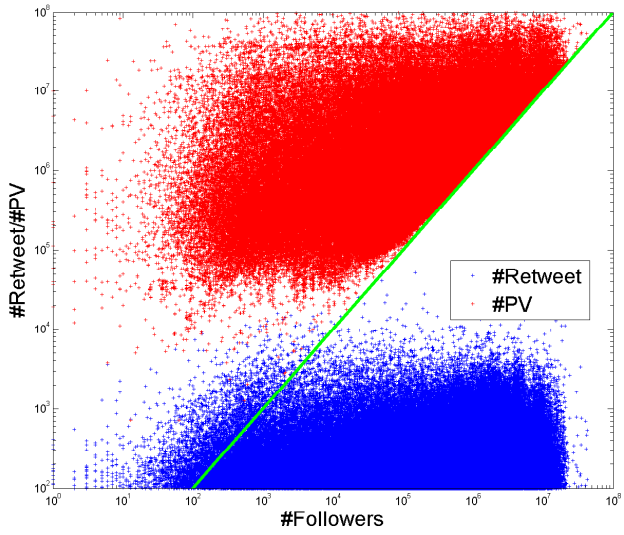



**Fig. 10** The relationships between #*pv*, #*retweet*, and #*followers*.

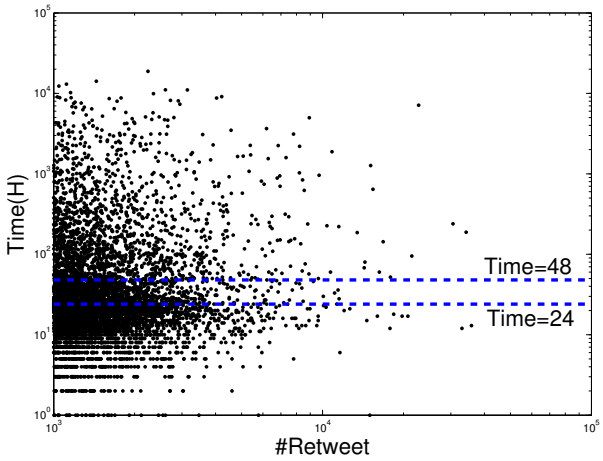



**Fig. 11** The life-cycle distribution of tweets with #*reteet* larger than 1000.

For tweets whose #*retweet* is in the range of [10, 1000), 76% tweets' $80-lifecycle$ values are less than 24 hours, and 85% tweets' $80-lifecycle$ values are less than 48 hours. The minimum, maximum, mean, median, and quantiles of their $80-lifecycle$ values is shown in Fig. 12.

For tweets with less #*retweet*, their lifecycles are shorter, since they have less possibilities to be seen. Actually, for tweets whose #*retweet* is less than 10, 87.3% tweets have $80-lifecycle$ less than 24 hours, and 91.4% tweets have $80-lifecycle$ less than 48 hours.

### 4.2 Tipping points in life-cycles of tweets

A close observation on life-cycles of tweets shows that for some tweets, #*retweet* does not increase in a steady speed. At some short time period, #*retweet* increases much more radpidly than common. This phase, often called *tipping point*, is coined in its sociological use by Morton Grodzins [6].

Formally, for a tweet, let $\Gamma(t)$ denote the cumulated number of retweets at time $t$, i.e. observed #*retweet* at time $t$, where $t$ is the time which is relatvie to tweet published time, $t$ is a tipping point if the following condtions hold:

1. $(\Gamma(t+\varepsilon)-\Gamma(t)) > \kappa$
2. $(\Gamma(t)-\Gamma(t-\varepsilon)) < \kappa$
3. $(\Gamma(t+\varepsilon)-\Gamma(t)) > \mu * (\Gamma(t)-\Gamma(t-\varepsilon))$
4. $(\Gamma(t+\varepsilon)-\Gamma(t)) > N/log(N)$

Here, $2\epsilon$ is the small time window size for observing the change of $\Gamma(t)$. $N$ is the final #*retweet* value, while $\kappa = N*0.8/T_{80}$, in which $T_{80}$ is the time used to finish the tweet's 80% retweet. Thus, intuitively, $\kappa$ is the average speed of retweet for the first 80% retweets. Finally, $\mu$ is the threshold on change rate of the slope.

Fig. 13 shows life-cycles of four tweets. Except $T4$, all other three tweets encounter tipping-points. $T3$ even met three ones in its life-cycle, though its final #*retweet* is not the largest. The figure shows that the number of tipping points and the happening time points may vary from tweet to tweet, while it is not necessary related to life-cycle or popularity.

The number of tipping points under different parameter combinations, calculated based on 0.43 million tweets with #*retweet* larger than 100, is shown in Fig. 14. When $\varepsilon = 1.5$ hours and $\mu = 20$, about 6.07% tweets meet tipping points. Fig. 15 shows the distribution of the number of tipping point in one tweet under the same parameter setting.

## 5 Popularity of a Single Tweet

### 5.1 Sigmoid function

The *sigmoid function* was first introduced by Gabriel Tarade for modeling the adoption rate of a new idea [16], and then widely used for modeling various phenomena [7, 13]. The function is defined as:

$$S(t) = \frac{1}{1+e^{-t}}. \quad (1)$$

The curve of the function, i.e. *sigmoid curve* has a shape similar to that of "S", so that it is also called *S-curve*. It is then often used to model measurments with small beginning

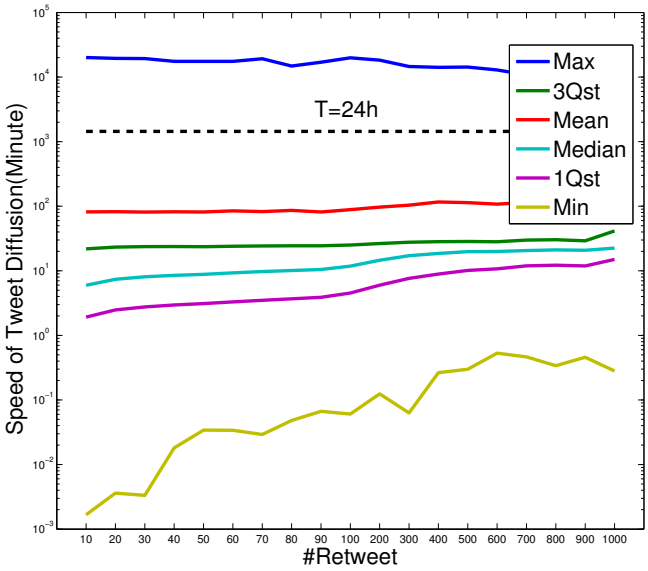



**Fig. 12** The propagation speed of tweets with #*retweet* in range of [10, 1000).

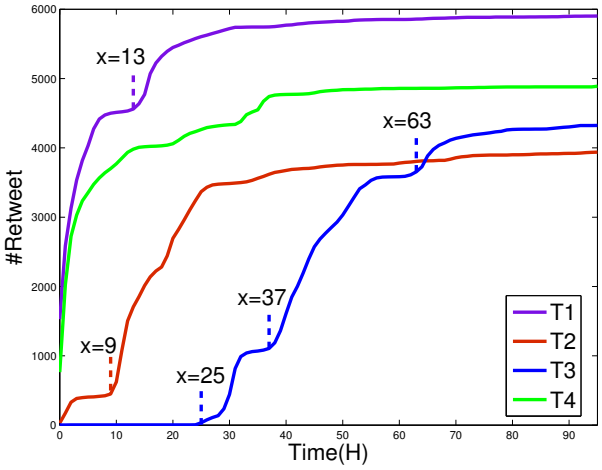



**Fig. 13** Life-cycles of four tweets with tippoing points annoted.

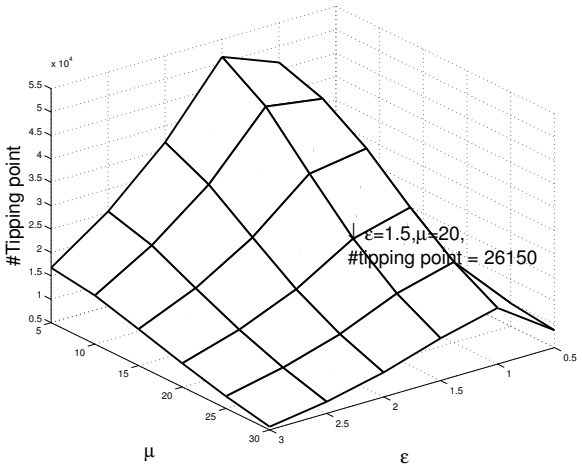



**Fig. 14** The number of tipping points under different parameters.

and accelerated growing over time [13]. Intuitively, popularity measurements, e.g. #*retweet* and #*pv*, of popular tweets with tipping points also follow such progressive increasing pattern. Thus, we use the following function:

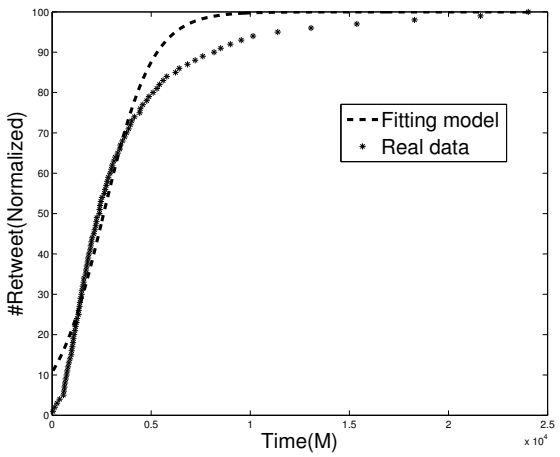
$$F(x) = \frac{N}{1 + a \cdot e^{-b(x-c)}}, \tag{2}$$

to model the popularity of tweets. Here, $N$ is the final #*retweet* of a tweet. Parameter $a$ and $c$ are used to control the horizontal position of the curve, while parameter $b$ controls the acceleration speed of the curve. To show the relationships between parameters and the shape of the curve intuitively, five curves with different parameter combinations are shown in Fig. 16.

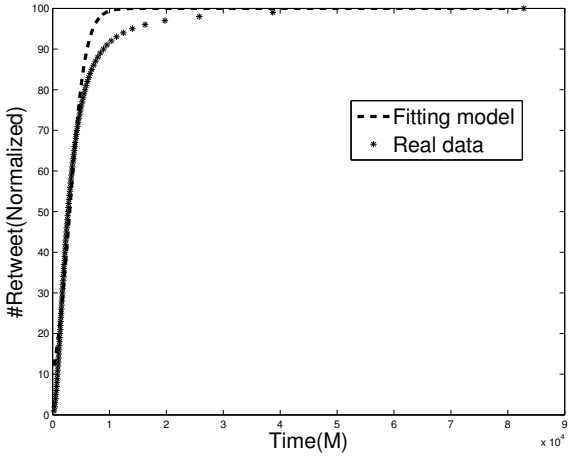
First, we use the S-curve to fit the tweet popularity change with time in average. The average #*retweet* over time of tweets with the measurement between 10 and 100 is shown in Fig. 17 (a), while that of tweets whose final #*retweet* larger than 100 is shown in (b). The average #*pv* over time of those two set of tweets are shown in Fig. 17 (c) and (d) respectively. The S-curve models, which are shown in the figure as well, fit the temporal dynamics of real data well with $R^2$ all larger than 0.95 in 95% confidence bounds.

To estimate the parameters, the trust region reflective algorithm [4] is used. Trust region reflective algorithm is an optimization algorithm that uses the point $x_k$ to define a trust region around the $x_k$ and iteratively search for the $x_k + 1$ inside the trust region to minimize the model function. Then, $x_k + 1$ is used to define the new target region, so that $x_k + 2$ can be obtained. The above operation is repeated iteratively until the process do not exhibt any new point. Therefore, $k$ is actually the iteration index.

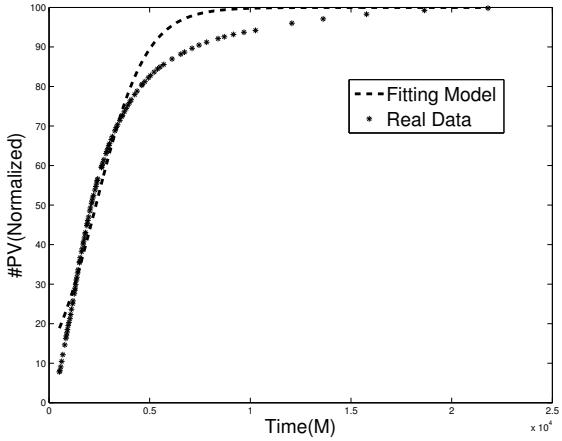
By applying the optimization algorithm on totally 208909 tweets with #*retweet* larger than 10, 87.54% models can fit the real data with $R^2$ larger than 0.8. For those models that well fit the read data, the distribution of parameters are shown in Fig. 18. Both $a$ and $b$ follow the normal distribution in log scale. It should be noted that the clear bound in Fig. 18 is caused by that tweets with #*retweet* less than 10 are not modeled, since popularity dynamics of single such tweet is of less impact, which is shown in Fig. 8.



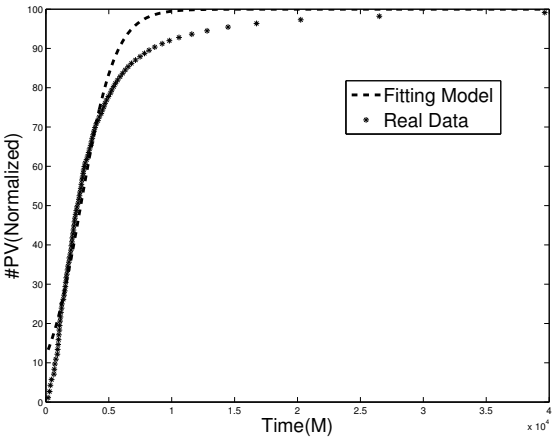

(a) Modeling the distribution of #*retweet*.




(b) Modeling the distribution of #*hashtag*.




(c) Modeling the distribution of #*retweet*.




(d) Modeling the distribution of #*hashtag*.

**Fig. 17** Modeling tweet popularity (avg) with S-curve.

## 5.2 Modeling for tweets with tipping-points

Aparently for tweets with tipping points, such as those shown in Fig. 13, the temporal dynamics of their popularity do not follow simple sigmoid function. A piecewise sigmoid function is used for modeling such tweets.

A piecewise sigmoid function is defined as:

$$F(x) = \begin{cases} \dfrac{N_1}{1 + a_0 \cdot e^{-b_0(x-c_0)}} & x <= x_1 \\ N_{i-1} + \dfrac{N_i - N_{i-1}}{1 + a_i \cdot e^{-b_i(x-c_i)}} & x_{i-1} < x <= x_i, 2 \le i \le \lambda \end{cases} \tag{3}$$

where

$$\sum_{i=1}^{\lambda} N_i = N. \tag{4}$$

Here, $N$ is the final #retweet, while $\lambda$ is the number of tipping points, and $x_1, x_2, \ldots, x_\lambda$ are timestamps of those tipping points.

**Algorithm 1** Modeling temporal dynamics for popularity of single tweet.

```
Detect tipping points based on four conditions;
for all piece between tipping points do
    Estimate N_{i-1} from last piece (0 for the first one);
    Estimate parameters a_i, b_i, and c_i by using trust region
    reflective algorithm;
end for
```

Algorithm 1 is used to model temporal dynamics of popularity with multiple tipping points. It first identifies tipping points based on the four conditions. Then, for each piece be-

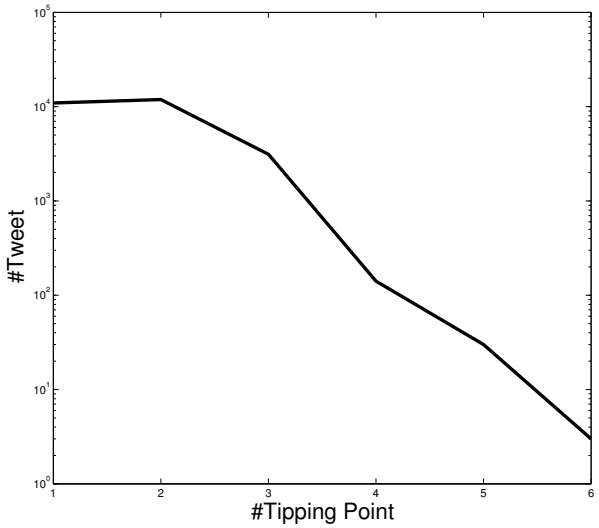



**Fig. 15** The distribution of number of tipping points over tweets.

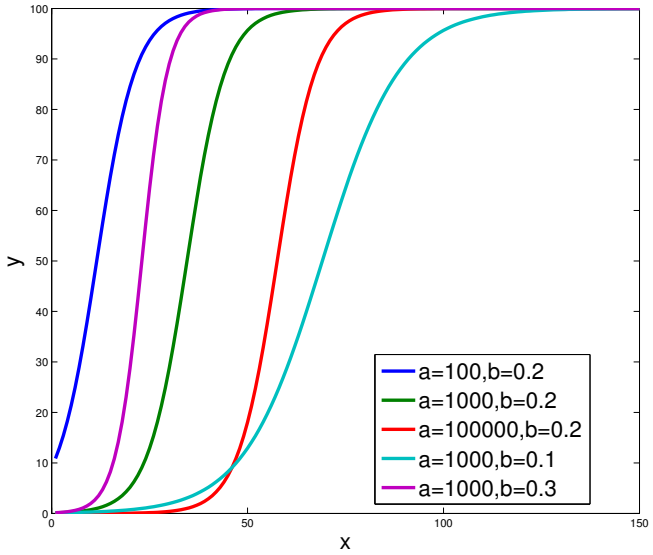



**Fig. 16** How the parameters a and b effect the S-Curve shape

tweetn two tipping points, parameters are determined.

Aparently the algorithm would not affect modeling quality for tweets without tippint points. Experimental results, which is shown in Fig. 19 over 208909 tweets with tipping points show improvement of 98.17% tweets.

## 6 Related Work

Microblogging services have attracted much attention in research communities. Both Twitter and Sina Weibo are widely studied [9,10,21]. Previous research mainly focus on characterising information diffusion network in Twitter [10, 19, 20]. The information diffusion networks used in these research are constructed based on hashtags. Therefore, they are implicit. As it is stated before, in microblogging services, such as Sina Weibo, where retweets may contain additional information, the explicit retweeting graph is another important source for studying how information is propagated.

Modeling popularity of pieces of information of different kinds, e.g. video clips on YouTube [3], channels in IPTV system [12], and tweets in microblogging services [1, 8, 11, 15, 18, 22], are the basis for understanding collective behavior of users. Our work extends the previous research in that multiple measurements of popularity are defined, while temporal dynamics of tweets are modeled.

There are research works that focus on factors that influence tweets' popularity [11, 15, 22], or predicting tweet or URL popularity [1, 8, 18]. This paper has different focus to those ones. The discussion on factors that influence tweets' popularity that are related to our work is provided in the next section.

The study on popularity of social media is often motivated by sensing real-world emerging events, e.g. earthquake [14], or predicting trends in real-world, e.g. index values of stock market [2], and results of polictical elections [17]. We leave the applications of our modeling technique as future work.

## 7 Discussion on Future Work

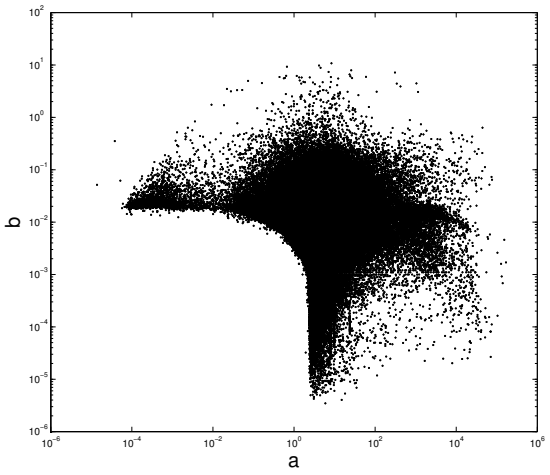
This paper focuses on measurements and modeling of popularity of single tweet. There are several important related issues that are discussed as follows.

### 7.1 Factors that influence tweet popularity

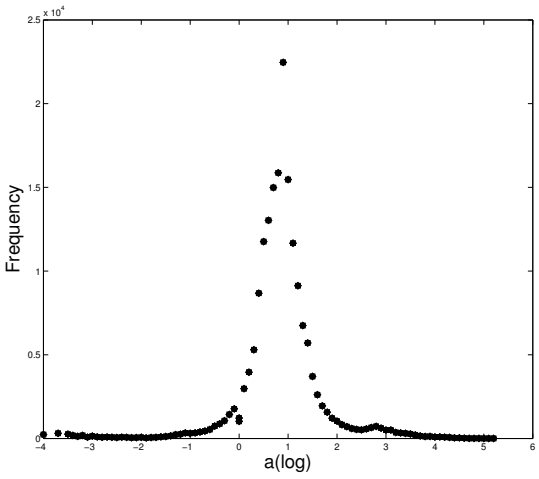
The popularity of a single tweet or of a set of related tweets is affected by many factors, one of which is the content of the tweet.

For example, Fig. 20 shows the retweet speed of deleted tweets[8)] and other tweets. A deleted tweet may be retweeted before it is deleted, while the retweet path is preserved in the timeline, though its content cannot be retrieved anymore. It is interesting that deleted tweets tend to have higher retweet speed compared to other tweets. Since the deletion of a tweet

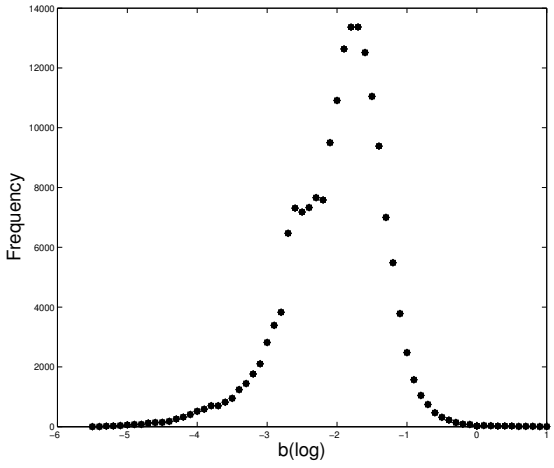
[8)] Tweets in Sina Weibo may be deleted by authors themselves, or by Sina Weibo management team due to censorship.




(a) The relationships between $a$ and $b$




(b) Distributions of $a$ and $b$




(c) Distributions of $a$ and $b$

**Fig. 18** The distribution of parameters $a$ and $b$ for models with $R^2$ larger than 0.8.

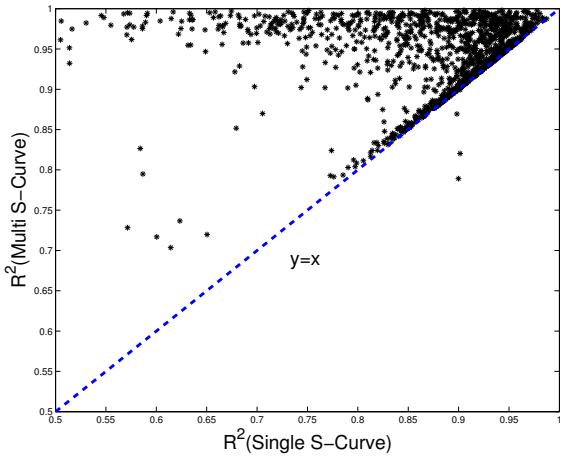



**Fig. 19** Comparison of single sigmoid function modeling with piecewise sigmoid function modeling on tweets with tipping points.

is not the reason for its high retweet speed, it is natural to conclude that users tend to be more sensitive to content of tweets that are possible to be deleted. It also suggests that content analysis is an important while challenging task for better understanding the popularity of tweets.

A second example, shown in Fig. 21, is the temporal dynamics for number of tweets, denoted as #*tweet*, over four sets of tweets on four events. It is shown that different events may result in totally different shape of #*tweet* curve. This is because that content of tweets are actually projection of events in real-world. An external event may greatly affect the change of popularity of tweets on the event. This example suggests that, though modeling tweet popularity is possible, to predict the popularity of tweets on a specific event is difficult. It should be noted that to predict the popularity of a single tweet is still possible, since the life-cycle of a tweet is usually short, which is shown in Fig. 11.

A natural question then is that whether the popularity of tweets on specific events depict any patterns. The distribution of #*retweet* over tweets on five popular short-term events are shown in Fig 22. It is shown that all of five curves satisfy power-law distribution. However, the involvement activities are different. For event of Trial of LI Zhuang's Case, more percentage of tweets have high #*retweet* compared to other events. This is because that the discussion on the case needs more professional knowledge on laws and politics. On the contrary, for the burst of event on GUO Meimei, there are a larger percentage of tweets with low #*retweet*. This example also show that the long-tail does not appear in tweets of all events. It again suggests that content is an important factor

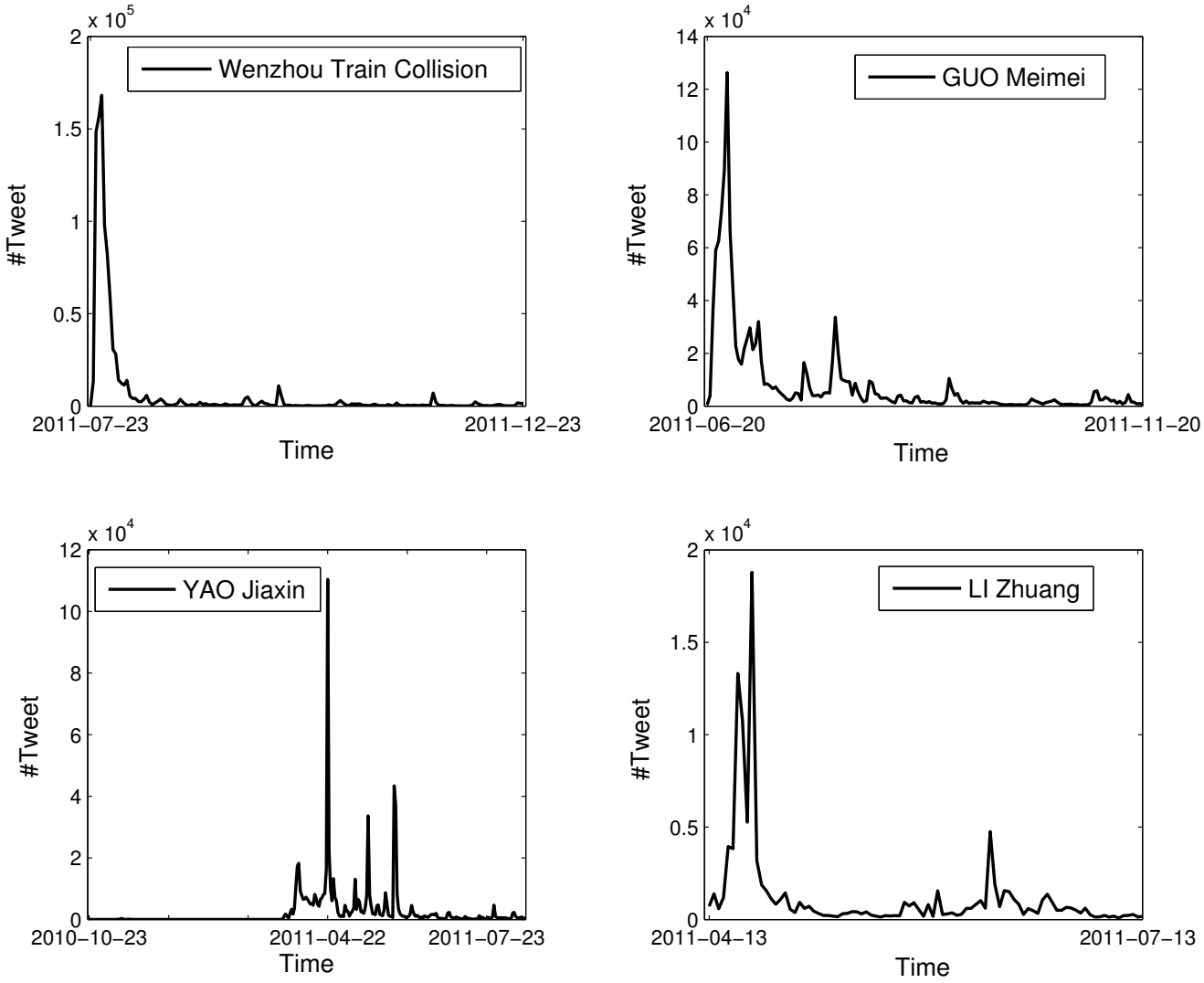



**Fig. 21** Bursts of set of tweets.

that affect the popularity of tweets.

Since a microblogging service is actual an open world that is affected by external events, we argue that for better predicting tweets popularity, we should also use external information sources, such as knowledge base constructed by experts or based on authority information sources, search or click logs, and real-time news or financial feeds. We leave the study on factors that influence tweets' popularity as our future work.

### 7.2 Sina Weibo vs. Twitter

The study in this paper is mainly based on Sina Weibo data. Though the authors believe that the discussion is also appliable on other microblogging services, it should be noted that different microblogging services do have different characteristics due to different mechanism design. Besides the difference of retweet in Sina Weibo and Twitter, which has already been discussed, there are two other differences that should be noted.

The first one is the number of followers and followees are different in Twitter and Sina Weibo. In Fig. 23, it is shown that users of Sina Weibo tend to follow more other users and

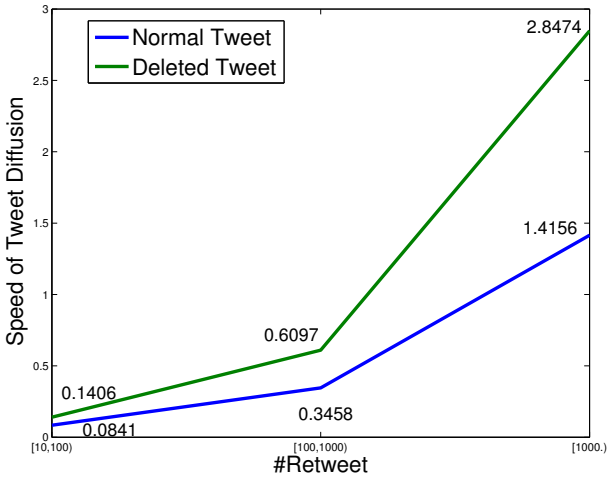



**Fig. 20** Comparison of spread speed of deleted tweets and other ones.

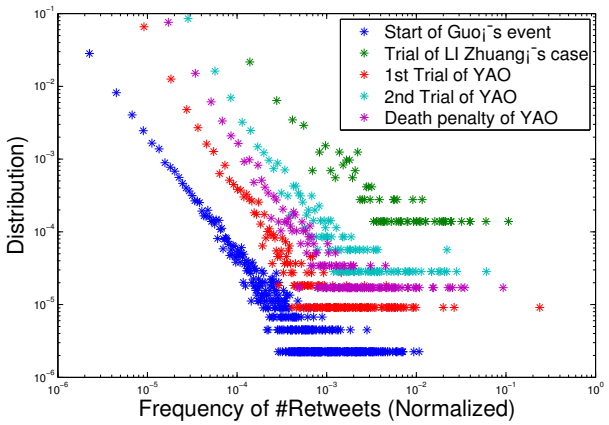



**Fig. 22** Bursts on events caused reason

tend to have more followers. Since the number of followers is related to the number of retweets, the different structures of the social networks behind Twitter and Sina Weibo should be considered, if the results of this paper are to be applied to Twitter.

A second important difference between Twitter and Sina Weibo is that the languages used by users are different. Twitter users may use many different languages, in which English is the dominant one, while most users of Sina Weibo use Chinese. Since in Chinese, there is no natural delimits between words, a hashtag must be put in a pair of #. Thus, hashtags are much less frequently used in Sina Weibo. Furthermore, in Fig. 24, it is shown that the length of tweets follows different distributions in Twitter and Sina Weibo. Therefore, the language itself should be treated as a factor that affect the popularity of tweets.

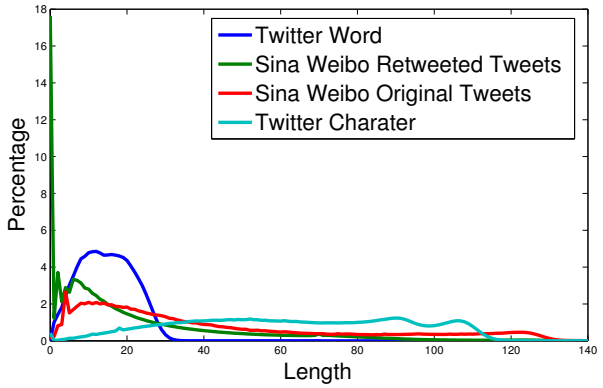



**Fig. 24** The distribution of the percentage of length.

## 8 Conclusion

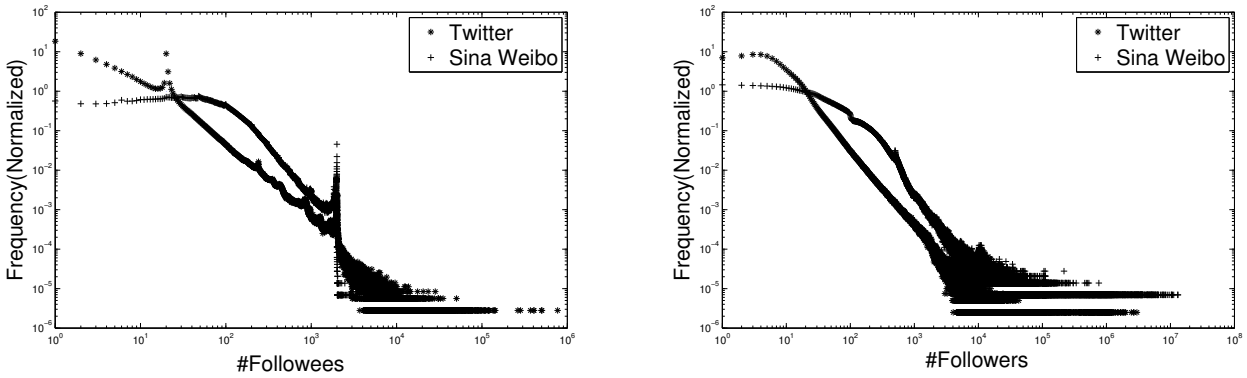
The tweets popularity of Sina Weibo is studied in-depth based on a real-life data set in this paper. Our empirical study show that: 1) different measurements should be used for better understanding of the popularity. A single measurement usually is insufficient. 2) Though the lifecycle of tweets' popularity usually follows sigmoid function, tipping points may emerge at certain time poin. Our piecewise sigmoid function may help to increase the precision of the modeling. 3) Tweets' popularity may be affected by external events greatly. Since single tweet's lifecycle is usually short, it is possible to model the temporal dyanmics of single tweets' popularity. However, for a specific group of tweets on an event, especially those with multiple stages, to model the change of their popularity is challenging.

## Acknowledgments

This work is partially supported by National Science Foundation of China under grant numbers 60925008, 61170086, and 61021004, National Basic Research (973 Program) under grant number 2010CB731402, and National High-tech R&D Program (863 Program) under grant number 2012AA011003.

## References

1. R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. *CoRR*, abs/1202.0332, 2012.




(a) Distribution of #*followee* (b) Distribution of #*follower*

**Fig. 23** The distributions of #*followee* and #*follower*.


2. J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.
3. Y. Borghol, S. Ardon, N. Carlsson, D. L. Eager, and A. Mahanti. The untold story of the clones: content-agnostic factors that impact youtube video popularity. In *KDD*, 2012.
4. M. Celis, J. Dennis, and R. Tapia. A trust region strategy for nonlinear equality constrained optimization. *Numerical Optimization*, 1984:71–82, 1985.
5. F. David, R. Pisani, and R. Purves. *Statistics.* Norton, 1995.
6. M. Grodzins. *The metropolitan area as a racial problem*. University of Pittsburgh Press, 1969.
7. J. Han and C. Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. *From Natural to Artificial Neural Computation*, pages 195–201, 1995.
8. L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW (Companion Volume)*, pages 57–58, 2011.
9. A. Java, X. Song, T. Finin, and B. L. Tseng. Why we twitter: An analysis of a microblogging community. In *WebKDD/SNA-KDD*, pages 118–138, 2007.
10. H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.
11. Z. Liu, L. Liu, and H. Li. Determinants of information retweeting in microblogging. *Internet Research*, 22(4), 2012.
12. T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao, and J. J. Xu. Modeling channel popularity dynamics in a large iptv system. In *SIGMETRICS/Performance*, 2009.
13. E. Rogers. *Diffusion of innovations*. Simon and Schuster, 1995.
14. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 2010.
15. B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom/PASSAT*, pages 177–184, 2010.
16. G. Tarde. The laws of imitation, trans. *Elsie Clews Parsons (New York, 1903), pp. 278f*, page 221, 1903.
17. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, 2010.
18. J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*, 2010.
19. J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.
20. J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.
21. L. Yu, S. Asur, and B. Huberman. What trends in chinese social media. *arXiv preprint arXiv:1107.3522*, 2011.
22. L. Zhang, T. Peng, Y. Zhang, and X. Wang. Content or context: Which carries more weight in predicting popularity of tweets in china. In *65th Annual Conference of World Association for Public Opinion Research (WAPOR)*, 2012.
23. A. Zhou, W. Qian, and H. Ma. Social media data analysis for revealing collective behaviors. In Q. Yang, D. Agarwal, and J. Pei, editors, *KDD*, page 1402. ACM, 2012.


## Appendix: Reference Pages of Events

- Wenzhou Train Collision: `http://en.wikipedia.org/wiki/Wenzhou_train_collision`
- LI Zhuang's Case: `http://en.wikipedia.org/wiki/2010_in_organized_crime`
- YAOJiaxin Murder Case: `http://en.wikipedia.org/wiki/Yao_Jiaxin_murder_case`
- GUO Mmeimei: `http://wiki.china.org.cn/wiki/index.php/Guo_Meimei`