On the Rise and Fall of Sina Weibo: Analysis Based on a Fixed User Group

Fan Xia [1], Qunyan Zhang [2],Chengyu Wang [3], Weining Qian [4] Aoying Zhou [5]

Institute for Data Science and Engineering, ECNU-PINGAN Innovative Research Center for Big Data
East China Normal University
{[1] fanxia, [2] 51121500043, [3] chengyuwang}@ecnu.cn
{[4] wnqian, [5] ayzhou}@sei.ecnu.edu.cn

***Abstract*—Micro-blogging service Sina Weibo in China has become the country's most free-flowing and important source of news and opinions just a few years ago. Following its launch in the summer of 2009, Sina Weibo grew quickly, attracting hundreds of millions of users and saw its biggest boom around 2011. However, several reports indicate a decrease in activity on Sina Weibo. In our study, we reveal the prosperity and decline of Sina Weibo by analyzing how a fixed user group's collective behaviors change throughout the whole development process. A huge dataset based on Sina Weibo along with search engine data is used in this study.**

In this paper we model the popularity of single tweet and multiple tweets. Then we define the statistic representing the capability of information propagation of Sina Weibo. The well-known time series prediction model, ARMA, is used to model and predict its trend. In addition, we extract both internal features, i.e. features of Sina Weibo, and external features, i.e. public's attention. Their trends are presented and analyzed. Then detailed experiments are conducted to measure the correlation and causality between them and our proposed statistic. The approaches we present in this paper clearly show the prosperity and decline of this microblogging community.

I. Introduction

Sina Weibo is a microblogging service that has been regarded as a revolution for the cyber community in China just a few years ago. Following its launch in the summer of 2009, it grew quickly, attracting hundreds of millions users. Users love its brevity and the speed at which it transmits information. However, there is no denying that things are on the downswing.

According to the survey from TechinAsia[1], Weibo's verified and influential users who have over 10,000 followers actually started to become less active in October 2012, based on the data from a third-party tool WeiboReach. Also, study by the TeleGraph[2] that sampled the activities of 1.6 million of the site's users found that activity had been dropping since late 2011, and dropped precipitously in the fall 2013. The similar trend could be found in Fig. 1. We plot the time series of active users, which will be define formally, and annotate events at important turning points. To the best of our knowledge, little research has attempted to provide a systematic overview of the whole development process of Sina Weibo and explain what is behind the drop.

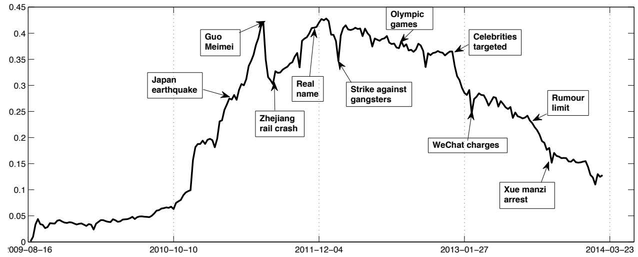


Fig. 1. AU(Active User) Time Series with Labeled Event

Sina Weibo is an information-sharing platform in essence. Its capability of information propagation is a critical measure of its popularity. We tackle the problem by first modeling the distribution of the number of retweets (donated as #retweet) for single tweet and multiple tweets. The choice of retweet number is due to the fact that retweeting is the main mean of information transition. Then we define the statistic to measure the capability and model its trend over time using time series model to reveal the evolution of Sina Weibo.

Furthermore, we try to figure out the whole process from two aspects, which are the interplay of the overall users' behaviors in Sina Weibo and public attention on similar social media products. We analyze the timestamped features extracted from Sina Weibo data, event data and search engine data at a large scale and on a fine grain. Users' collective behaviors on Sina Weibo, such as tweet, re-tweet, mention, following, publishing a hashtag or an URL, are studied to find out how they change over time. In addition, search engine data can be viewed as the public attention on certain social media including Sina Weibo, similar products like Tencent Weibo[3] and competitive products like WeChat[4]. All these features are generated and normalized as time series. At last we explore the correlation and causality among them.

Our main contributions can be summarized as follows:

[1] http://www.techinasia.com/sina-weibo-users-less-active-2013/

[2] http://www.telegraph.co.uk/news/worldnews/asia/china/10608245/China-kills-off-discussion-on-Weibo-after-internet-crackdown.html

[3] http://t.qq.com/

[4] http://wechat.com/

978-1-4799-8442-8/15/$31.00 © 2015 IEEE

- By applying tweet popularity modeling, we find Malthusian parameter of sigmoid function is quite suitable to indicate speed of information propagation of Sina Weibo. The statistic analysis shows its value changes over time.
- The ARMA model is used for modeling the distribution of Malthusian parameter through time, which clearly shows the development trend for further prediction.
- Novel, fine-grained features extracted from both Sina Weibo dataset and other data sources are defined. Their trend are analyzed in detail.
- Correlation and causality analysis reveal the development of Sina Weibo from two aspects: a) within the Sina Weibo community itself and b) relationships between real-life data from other domains.

The rest of the paper is organized as follows. In Section 2 we first give an introduction to various kinds of dataset used by our study. In Section 3, the popularity of single tweet and multiple tweets are modeled using sigmoid and LGM model respectively. Then we model and predict the trend of the parameter that reflects the speed of information propagation in Section 4. In Section 5 we discuss how various kind of features are extracted. Their trend, correlation and causality analysis are also presented. Related work is described in Section 6. We finish the paper with conclusion and discussion in Section 7.

II. Dataset

We first introduce all the dataset used in our study, which consists of the internal data from Sina Weibo, including both raw data and annotated data, and the external data from Google search engine.

Sina Weibo Data This raw dataset was collected from weibo.com. The framework and crawling strategy of the crawler have been described in [1]. The dataset contains 1.6 million users' profiles, social networks and their timelines from August 2009 to the end of March 2014, including about 1.8 billion tweets. Attributes of each part have also been described in [1].

The completeness of timelines of the user group is very important for extracting the time series of users' engagement and behavior changes on Sina Weibo. We are one of the pioneers in China to crawl Weibo data, which makes our dataset cover the whole life span of Sina Weibo. Though it should be noted that the items in the dataset are neither synchronized nor complete, it is sufficiently large to depict the users' usage status of microblogging service.

Hot Event Data The hot event dataset contains tweets related to hot events. To obtain event-related tweets, we have defined an event set, which consists of 220 events such as *Wenzhou Train Collision*, *Guo Meimei* and *Japan earthquake*. *Hashtag* is kind of metadata to identify the topic of tweets. However, most of the users do not have the habit to cite hashtags in Sina Weibo. Therefore, we use a series of regular expressions to find the relevant tweets. To refine the expressions, we used search engine of Sina Weibo to evaluate the keywords we provided. Both the relevance to the event and the number of tweets should be considered when we select the expressions.

Search Engine Data To understand public attention to social media such as Sina Weibo, Tencent Weibo and WeChat, we create a search query-based dataset. We downloaded the weekly search volume data in the world and in China for a set of seed queries such as “weibo”, “weixin” and “wechat” from Google Trend, which is a Google service that provides search volume data from January 2004. To fully capture search activities related to social media, we expand these seed keywords with top relevant search terms recommended by Google Trend. In this paper, we focus on two types of keywords' search volumes: a) microblogging communities (Sina Weibo, Twitter and Tencent Weibo) and b) social media platforms that are competitive to Sina Weibo (WeChat).

III. Quantifying Tweet Popularity

Microblog services are created for users to create and share information at anytime, anywhere. The capability of information propagation is an important criterion to measure the energy of a microblog platform. Retweet is such a behavior in Weibo, which transmits information to be seen by more people. Thus, the number of retweets, donated by #retweet, of one or a set of tweets, is an important popularity measurement in microblogs.

A. Popularity of a Single Tweet

We first study the retweet behavior of a single tweet to measure the popularity. In [1], the sigmoid function is used to model the phenomenon that #retweet increases over time in an “S” curve for a single tweet. In this paper, we adopt the following three-parameter sigmoid function that #retweet of a single tweet $S(t)$ increases with time t:

$$S(t) = \frac{N}{1 + a \cdot e^{-b \cdot t}} \tag{1}$$

where N is the final #retweet of a single tweet, and parameters a and b are used to control shape of the “S” curve. More specifically, parameter a controls the horizontal position of the curve. Parameter b, also known as Malthusian parameter[2], determines rate of maximum growth. The parameters can be estimated through a trust region strategy of numerical optimization described in [3].

The parameters a and b have practical meanings in modeling the popularity of a single tweet. a indicates how much time has been passed before a tweet begins to be retweeted in a fast speed. That is, the smaller a is, the shorter time a tweet has been posted before gaining its popularity. b shows the acceleration speed of the retweet behavior. The larger b is, the more popular a tweet is when it begins to be retweeted.

By examining parameters a and b estimated from a large number of tweets, we can infer the collaborative behaviors of the entire social network. Hence we model the distributions of those two parameters in the next section.

B. Popularity Distribution in Multiple Tweets

After we carefully model the retweet behavior of a single tweet, in this section, we consider the distribution of parameters under a large number of tweets in a certain time span.

More formally, given a time span S, we collect all tweets whose #retweet is larger than 10, denoted by T. For each tweet $t_{(i)}$ in T, we model the retweet behavior as a sigmoid function and then estimate the parameters $a_{(i)}$ and $b_{(i)}$. Then $a_{(i)}$s and $b_{(i)}$s are modeled by Log Gaussian Models (LGMs)[4] respectively. We first describe the model in detail, and then conduct a goodness-of-fit verification using statistical methodology.

Log Gaussian Model A LGM forms a continuous distribution (Galton distribution) where the logarithm of the random variable is normally distributed. More specifically, if X is log-normally distributed, then $log(X)$ is normally distributed where X denotes a random variable. The distribution is denoted as $X \sim \ln N(\mu, \sigma^2)$, where μ and σ^2 are the mean and variance of the X's natural logarithm. The probability density function (p.d.f.) of the model is:

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Parameter Estimation The parameters can be estimated by Maximum Likelihood Estimation (MLE) shown as follows:

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} \ln x_{(i)}}{n} \tag{3}$$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\ln x_{(i)} - \widehat{\mu})^2}{n} \tag{4}$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are maximum likelihood estimators of μ and σ^2, receptively. $x_{(i)}$ is the ith sample from the dataset of size n. We omit the proof due to constraints of space.

In the experiment, we fit the sigmoid functions of a total of 208,909 tweets with $\#retweet > 10$, and use parameters of obtained models to fit LGM model through MLE. We employ *Sum of Squared Error* (SSE), *coefficient of determination* (R^2) and *adjusted coefficient of determination* (Adjusted R^2 or $\overline{R}^2$) to evaluate the goodness of fit. R^2 ranges from 0 to 1 where 1 indicates perfect fitting and $\overline{R}^2 \leq R^2$. The result is shown in Tab. I. All the coefficients are very close to 1. We can conclude that distributions of a and b are well fitted by log-Gaussian model.

TABLE I
GOODNESS OF FIT EVALUATION

Random Variable	SSE	R^2	$\overline{R}^2$
a	0.0002	0.9443	0.9441
b	0.0008	0.9383	0.9376

IV. CHANGE IN CAPABILITY OF INFORMATION PROPAGATION OVER TIME

The practical meaning of coefficients a and b make them excellent statistics that capture the capability of information propagation. In terms of information propagation, a smaller a means the information attract attention more quickly while a greater b means the information get diffused more rapidly. In this section we start to figure out whether and how their values change over time.

A. Analysis of Variance

As shown by Fig. 1, there exist bursting events in Sina Weibo, which results in turning points in trend of active users. We first give a detailed statistical verification on whether the distribution of a and b is relevant to the time span via one-way Analysis of Variance (ANOVA)[5]. Wald Test[6] can be used to evaluate null hypothesis in ANOVA. It produces a *F statistic*, which is the ratio of between-group variance and within-group variance. Then we can calculate the *p-value* p (critical value of the F-distribution) to show the statistical significance of the experiment. Give a significant level α (typically 0.05 or 0.01 in practice), if $p < \alpha$, we can reject the null hypothesis.

In our case, we take the logarithms of parameters a and b as response variables, as we recall that, in LGM, $log\ a$ and $log\ b$ are normally distributed. We also take all six time spans $T = \{T_1, T_2, ..., T_6\}$ as six treatment groups. We denote la_i as the response variable $log\ a$ under time span i. For a, we have the following hypothesis:

$$H_0 : la_1 = la_2 = \cdots = la_6 \tag{5}$$

$$H_1 : la_1, la_2, \ldots, la_6\ are\ not\ equal \tag{6}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis. To generate each observations, we sample tweets from each treatment groups and calculate observations by fitting the sigmoid function and estimating the parameters. As we can see in the previous section, these observations are independent and identically distributed (i.i.d.).

In our experiment, we carefully sample tweets from our Weibo dataset in six time spans mentioned above and conduct the ANOVA experiments. The experiments on b give the result that $F - statistic = 27.28$ and $p - value < 2e^{-16}$. With the setup of α=0.01, the result shows the p-value of parameter b is smaller than α, which proves that we can reject the null hypothesis with great statistical significance. However, the experiment results show we could not reject that of a. The result clearly demonstrates that the speed of information propagation diverges over time. In the following section, we focus on modeling and predicting the distribution of parameter b, which we simply call Malthusian parameter.

B. Trend Modeling and Forecasting

We have shown that distributions of retweet behaviors vary through statistical analysis. In this section, we propose to use Autoregressive Moving Average Model (ARMA) to model and predict the distribution of Malthusian parameter.

Autoregressive Moving Average Model The ARMA model[7], introduced by Peter Whittle, is a powerful statistical model to analyze a stationary stochastic process. It consists of

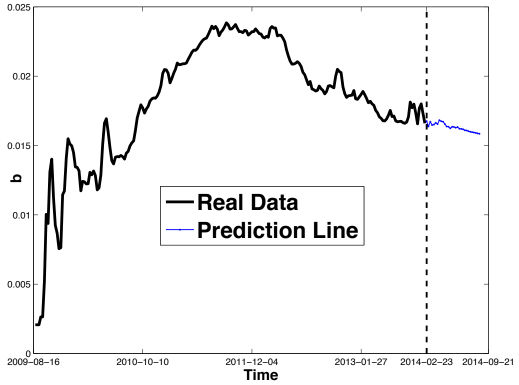


Fig. 2. Fitting and Prediction of Malthusian parameter

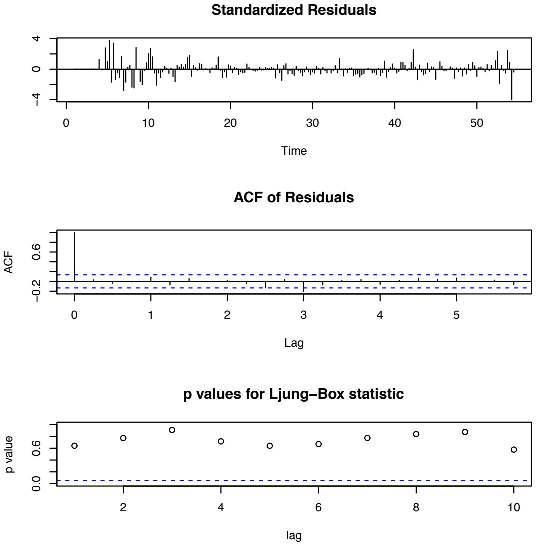


Fig. 3. Evaluation of Malthusian parameter

two polynomials, one for the auto-regression and the other for the moving average. We briefly describe the model in the following:

- **Autoregressive(AR) model** The AR model $AR(p)$ specifies that the output variable depends linearly on its own previous values. Given the order p, the AR model is written as:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-1} + \epsilon_t \tag{7}$$

  where c is a constant, all ϕ_is are parameters of the AR model and ϵ_t is the white noise.

- **Moving-average(MA) model** The MA model $MA(q)$ is an univariate time series model of the random variable with white noises. Given the order q, the MA model is written as:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-1} + \epsilon_t \tag{8}$$

  where μ is the expectation of X, all θ_is are parameters of the MA model and ϵ_t is the white noise.

- **ARMA model** The ARMA model $ARMA(p, q)$ is the combination of p autoregressive terms and q moving average terms, shown as follows:

$$X_t = \sum_{i=1}^{q} \theta_i \epsilon_{t-1} + \sum_{i=1}^{p} \phi_i X_{t-1} + \epsilon_t + c \tag{9}$$

Model Evaluation We fit the ARMA model with value of Malthusian parameter from April, 2009 to February, 2014. As shown by Fig. 2, the black line exhibits the ARMA model fitting from our real data. The blue line is the future prediction in 30 weeks using the model. The quality of the model are evaluated by three approaches shown in Fig. 3. We first analyze the residuals, which are the differences between the observed values and the estimated function values. The figure shows the residuals of the time series are normally distributed, which proves the ARMA model captures the variation patterns well and the residuals represent the white noise in the stochastic process. The other test is $Ljung - Box\ test$, introduced by G. Ljung and G. Box[8], [9], which is usually applied to the residuals of the ARMA model. The test statistic is the $Q - statistic$[8]. In Fig. 3, we present the p-values of the series, which are over 0.5. This test clearly shows the well-fitting behavior of our model.

Trend Analysis From the model, we can see the popularity trend of Sina Weibo in terms of retweet behaviors. After its launch in 2009, it continued to gain popularity with several drops in 2009 and 2010. It reached its peak in 2011, with high retweet speed. However, things started to change in 2012. Our analysis shows that the popularity of Sina Weibo has been dropping from 2012. Using the AMRA model, we also have a forecasting study. Based on our continuous tracking of Sina Weibo, we can predict that the drop in popularity of Sina Weibo will continue, as illustrated in Fig. 2.

V. Overview of Trend from Features

In this part, we will give an overview of the development process of Sina Weibo according to the features we extract. Time series curve of each feature and the relationship among them are shown.

A. *Extraction of Features*

We extract two groups of features from Sina Weibo data: activity features and graph features, which are similar to [10].

Activity Features Activity features indicate users' activities in Sina Weibo. They are counts of particular types of statistics in a time interval donated as W_{t_1,t_2}, such as number of tweets, number of users and number of hashtags. All these features have minimum timestamps at a granularity of one day. Among these, *active user* is an important concept, which is formalized as follows:

Definition [Active Users] If a user posts a tweet, we call the user is active in that day. Given a time interval W_{t_1,t_2}, $\#ad$ is the number of active days during W_{t_1,t_2}. Take a few weeks before and after the particular week as a target period to

TABLE II
DETAILED FEATURE DEFINITIONS AND DESCRIPTIONS

Activity Features	Description
TID/OID/RTID/DID	average number of tweets/original tweets/re-tweets/deleted tweets in W_{t_1,t_2}.
RT50/RT250/RT500	number of tweets whose repost is between 10 and 100, 100 and 500, larger than 500 in W_{t_1,t_2}.
THTG/TURL	number of tweets with hash-tags/URLs in W_{t_1,t_2}.
TU/VU	proportion of different users/verified users that posted a tweet in W_{t_1,t_2}.
OTU/RTU	proportion of different users posted an original tweet/re-tweet in W_{t_1,t_2}.
UHTG/UURL	proportion of different users posted an tweet with hash-tags/URLs in W_{t_1,t_2}.
UFRN/UFLW/UFBF	average number of friends/followers/bi-followers of user who post a tweet in W_{t_1,t_2}.
AU/AV	proportion of active users/verified active users in W_{t_1,t_2}.
ATID/AOID/ARID/ADID	average number of tweets/original tweets/re-tweets/deleted tweets of active users in W_{t_1,t_2}.
AHTG/AURL	proportion of tweets with hash-tags/URLs by active users in W_{t_1,t_2}.
AFRN/AFLW/AFBF	average number of friends/followers/bi-followers of active users in W_{t_1,t_2}.
EU/EV	number of users/verified users who involve in hot events in W_{t_1,t_2}.
ETID/EOID/ERID	average tweets/original tweets/re-tweets of each user involved in hot events in W_{t_1,t_2}.
EHTG/EURL	proportion of tweets with hash-tags/URLs of hot events in W_{t_1,t_2}.
EFRN/EFLW	average number of friends/followers of user involved in hot event in W_{t_1,t_2}.
Graph Features	**Description**
DEGREE	average degree of nodes of $W^c_{t_1,t_2}$.
COMPONENT	statistics on the connected component distribution for W_{t_1,t_2}. (AVG, STDV, SKEWNESS, KURTOSIS) denoted by CON_AVG, CON_VAR, CON_SKE,CON_KUR.
Search Volume Index	**Description**
IDX_W_W/IDX_S_W	sum of search volumes worldwide of weibo/wechat related keywords in W_{t_1,t_2}.
IDX_W_C/IDX_SC	sum of search volumes in China of weibo/wechat related keywords in W_{t_1,t_2}.
WEIXIN_C /WEIXIN	search volumes keyword "weixin" in China/world along with its suggested words in W_{t_1,t_2} (the same with other keywords).

calculate $AVG(\#ad)$, which represents the average number of active days in W_{t_1,t_2}. $AVG(\#tweet)$ represents the average number of tweets in W_{t_1,t_2}. A user u is said to be active if $AVG(\#tweet) > \kappa_1$ and $AVG(\#ad) > \kappa_2$ (κ_1 and κ_2 are pre-defined thresholds and we have $\kappa_1 > \kappa_2$). In this paper, we set $\kappa_1 = 4$, $\kappa_2 = 3$ and the length of time interval is one week.

Graph Features Graph features measure the properties of the linked structure of the interaction graph among the users as shown in [10]. However, the definition of graph is different in our work since we emphasize the retweet behaviors of users. In this paper, we take users as nodes instead of tweets, hashtags, users, etc. Two types of links between users that we define are **Mention** and **Re-tweet**. Also, we provide a new definition of the constrained subgraph in the $W_{t1,t2}$ as follows.

Definition [Constrained Subgraph] Let G be an interaction graph $G = (V, E)$, the nodes and edges of which are defined as above. The *constrained subgraph* $G_{t_1,t_2} = (V, E)$ contains the nodes V of G that are users either retweet, mention, are retweeted or mentioned by other users in interval $[t_1, t_2]$. All the edges E in G whose end-nodes are in V are added to G_{t_1,t_2}. For re-tweet edges, we use the timestamp of the retweet. For mention edges, we use the timestamp of the tweet for the edge.

All features and their detailed descriptions are listed in Tab. II. The Map-Reduce framework is used to extract them.

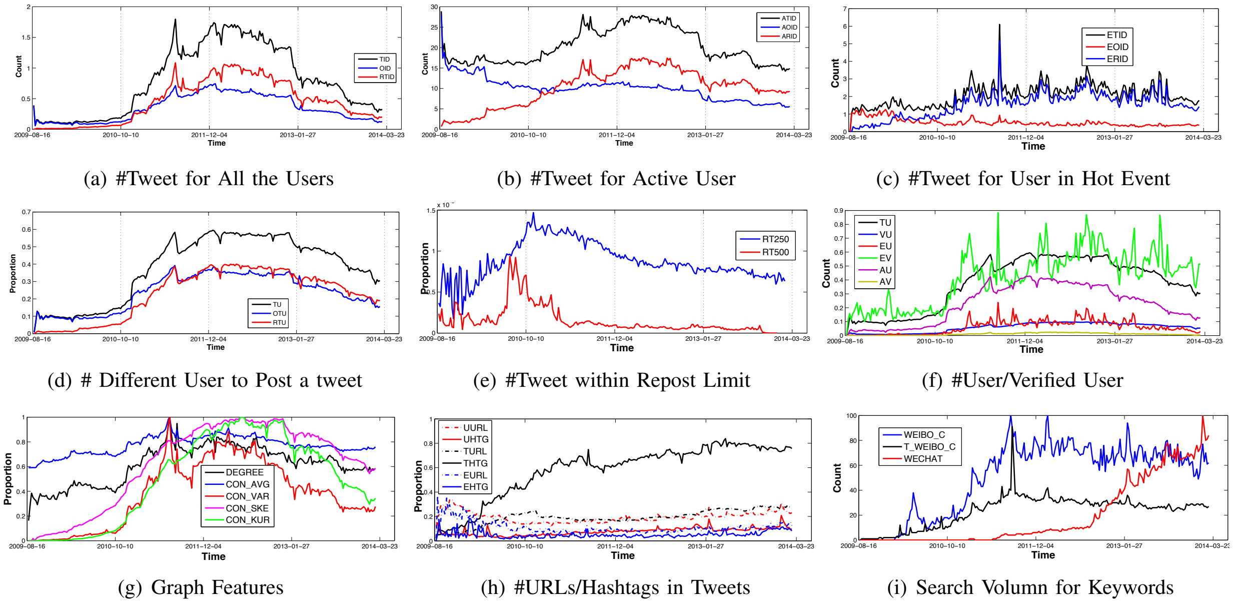
B. Generating Time Series

Although our final dataset is a snapshot of fixed set of users, it does not mean they begin to use Weibo from the same time. Hence, normalization is necessary for generating time series. For example, if we calculate the proportion of tweets in one day, simply divide the number of tweets in that day by the total number of tweets is not enough since our user group size is not the same at different time. Hence, we use average number of tweets per user per day to indicate the popularity of tweets. It should be noted that user group size in a certain day is the number of users who have registered before that day. The same normalization strategy can be used for re-tweets and active users. Other features like number of URLs, hashtags, are normalized using the number of tweets for the full day.

C. Trend in Features

It is difficult to choose one single feature that can represent the activity index of Sina Weibo. Each feature extracted from Weibo dataset represents one dimension of it. In this section, multiple time series are visualized to provide an overview of how each of these features evolve over time.

In Fig.4(a), the overall trends of three tweet count series go upward first, then go downward. The first significant rise of tweet count happened around the end of October, 2010, then followed by a rapid increase in the next six months. Both in the later half of 2011 and the full year of 2012, users' average tweet count per day maintained at a relatively high level. However, it suffered continuous decline afterwards. The time of turning points are the end of 2011, mid-July of 2013 and the beginning of 2014. These three turning points are in accordance with the time of strike against celebrities, rumor control and Xue Menzi(a well-known user in Sina Weibo)'s arrest as we marked in Fig.1. Tweet count of active users is shown in Fig.4(b). The curve in the first six months is quite different from what in Fig. 4(a) for users' frequent use when Sina Weibo was just launched. Average tweet count of active users was extremely high in the beginning. Another interesting point is that the proportion of original tweets and retweets of active users. Users choose to transmit rather than



(a) #Tweet for All the Users

(b) #Tweet for Active User

(c) #Tweet for User in Hot Event

(d) # Different User to Post a tweet

(e) #Tweet within Repost Limit

(f) #User/Verified User

(g) Graph Features

(h) #URLs/Hashtags in Tweets

(i) Search Volumn for Keywords

Fig. 4. Time Series of Selected Features

create information. Thus, retweeting has become an important behavior of active users and the collective retweeting behavior is also an important criteria to quantify the activity index of Sina Weibo. Proportion of different users that post an original tweet or a retweet in Fig.4(d) exhibits similar trend as the curve in Fig.4(a), but not as sharp as that in Fig.4(a). It means users still have the habit to use Sina Weibo but tend to post fewer tweets than before. In Fig.4(c), these three indicators of hot event are quite stable, which means although the sensitivity to hot events and users' involvement are not going down, it also means that Sina Weibo tends to become a traditional media. The similar phenomenon can be found in Fig.4(f) and Fig.4(h). The proportion of tweets with rather high repost also reaches its peak around the first half of 2011.

Graph features are shown in Fig.4(g). We find that the change of standard variance time series is quite agreed with curves in Fig.4(a), and changes of skewness and kurtosis are agreed with curves in Fig.4(d). In Fig.4(i), we show three keywords search volume time series and learn the fact that "WeChat" is getting more and more attention while "Weibo" and "Tencent Weibo" search volumes tend to go in the opposite direction.

D. Time Series Analysis Techniques

Before diving into the relationship among Malthusian Parameter and various features, we first introduce two analysis techniques that will be used later.

Correlation Analysis The pair-wise Pearson's correlation coefficient r among various time series are computed. The coefficient measures the linear dependence between two series. It has the range of $[-1, 1]$, with $+1$ indicating exact positive linear dependence, -1 indicating exact negative linear dependence and 0 indicating no linear dependence between two series. By applying correlation analysis, we hope to find features that exhibit similar trends.

Causality Analysis However, high correlation between two variables doesn't imply there also exists causation relationship between them. Thus we adopt the Granger causality analysis[11] to infer the causality of the statistic concept. It is a statistic hypothesis test based on prediction. Suppose X and Y are two time series. For a time lag $s > 0$, the MSE of predicting y_{t+s} based on $(y_t, y_{t-1}, y_{t-2}, \dots)$ and its combination with $(x_t, x_{t-1}, x_{t-2}, \dots)$ could be computed respectively. X fails to Granger-cause Y if two MSE are nearly the same for all $s > 0$. An F-test is conducted to examine if the null hypothesis that $X(t)$ does not Granger-cause $Y(t)$ can be rejected.

E. Relationship between Internal Features and Malthusian Parameter

We conduct both Pearson's correlation and Granger causality analysis between internal features and Malthusian parameter. The results are presented in both Fig. 5 and Tab. III. The null hypothesis of some intern feature Granger-causes Malthusian Parameter with lag varying from 1 to 6 weeks are tested. Only the p-values under 1 lag are presented due to limitation of space.

Tab. III lists top 10 correlated features in descending order of correlation value. We notice that they portrait the collective behavior of users from various aspects. First, activity graph features CON_SKE and CON_KUR describe characteristics of the structure of users' interaction. With higher value of CON_SKE and CON_KUR, people opt to discuss in large communities rather than dispersing in many warrens. Second, top related features such as RTU, RTID and TU reflect users' willingness to post tweets and share with others. At last,

TABLE III
FEATURES OF TOP 10 PEARSON CORRELATION AND P-VALUE OF THEIR GRANGER CAUSALITY ANALYSIS

Feature	Correlation	Causality(Lag 1 Week)
AV	0.6572	0.0043
RTU	0.6567	0.0083
CON_KUR	0.6553	0.0055
TU	0.6544	0.0095
CON_SKE	0.6543	0.0074
RTID	0.6541	0.0046
UFRN	0.6521	0.0096
AU	0.6508	0.0063
VU	0.6472	0.0107
OUT	0.6445	0.0093

features involving activity of users, i.e. AV, AU, VU and OUT, determine the scale of users taking part in the propagation. UFRN directly reflects activity of neighbors in the one hop ego-centric network, which has influence on users' activity.

Furthermore we conduct Granger causality analysis to figure out whether there exists causality relation between them. Fig. 5 illustrates that in our case the p-values trend to be small for features with high correlation. As shown in Tab. III, most of the null hypothesis can be rejected under the significance level of 0.01. Thus we can accept the alternative hypothesis that those features Granger-cause Malthusian Parameter and conclude that those collective behavior has impact on the change in capability of information diffusion.

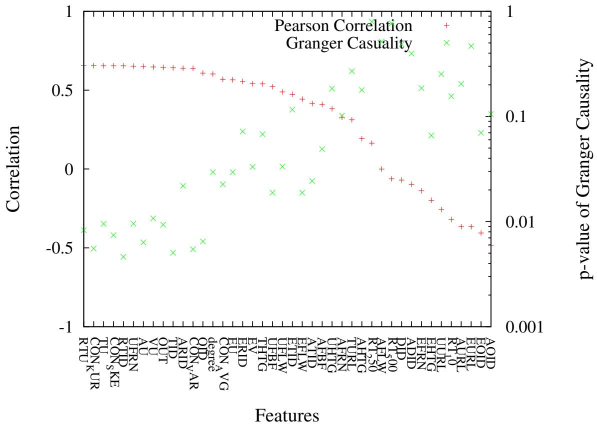


Fig. 5. Results of Pearson's Correlation And Granger Causality

F. Relationship Between Features

In this section, relationship between features is analyzed in order to examine the correlations in behaviors within Sina Weibo and across different domains. Especially we are concerned with the question weather variations of public attention related to social media correlate with changes of the activity in Sina Weibo. Due to the fact different social media services start up at different time, all the following analysis are applied to overlap part of feature time series.

The result of Pearson' Correlation is given by Fig. 6. It is easy to understand that we find strong correlations in most cases within Weibo dataset, especially for the number of tweet and users. An interesting exception is that most of features involving URLs, i.e. AURL, EURL and TURL, trend to have negative correlation with features involving the number of tweets, retweets and users, which requires further exploration.

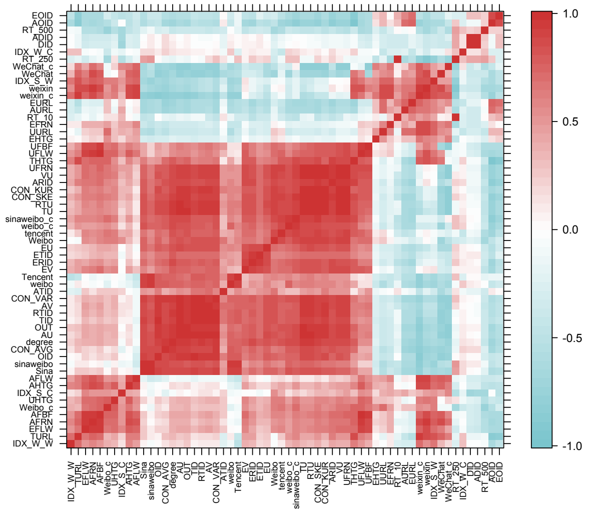


Fig. 6. Pearson's Correlation Results

TABLE IV
PART OF RESULTS OF CORRELATION AND CAUSALITY ANALYSIS BETWEEN WEIXIN AND INTERNAL FEATURES

Feature	weixin		sinaweibo	
	Correlation	Causality	Correlation	Causality
AV	-0.4469	0.0073	0.7819	0.015
RTU	-0.3216	0.0004	0.7317	0.0138
CON_KUR	-0.1962	2.83e-09	0.6831	0.0003
TU	-0.4098	6.46e-06	0.7320	0.0021
CON_SKE	-0.1909	4.04e-09	0.6966	0.0001
RTID	-0.5124	0.0009	0.8231	0.0179
UFRN	0.0803	0.0003	0.6285	0.0007
AU	-0.6508	3.43e-05	0.8165	0.0045
VU	0.02492	0.0059	0.6321	0.0059
OUT	-0.6862	7.05e-06	0.7989	0.0171

Due to the limit of space, we only present the analysis results between "weixin" , "sinaweibo" and ten features listed previously in Tab. IV. The existence of extremely smaller values from Granger Causality analysis may be caused by the fact that the assumption of normality used by the test statistic is incorrect in those cases and p-value depends heavily on the tail behavior. However, it is still an indicator of heavy significance. The result clearly demonstrates public attention has a great impact on Sina Weibo.

VI. RELATED WORK

In recent years, much research work has focused on the study of microblogging communities. Depend on the detailed fields they cover, the related work can be further classified into the following three categories:

Microblogging Community Analysis Microblogging communities such as Twitter and Sina Weibo have been heavily studied in the literature, which is closely related to our work. Kwak *et al.*[12] studied the entire Twittersphere. The topological characteristics of Twitter and the way people share information on the Web are studied. Java *et al.*[13] analyzed users' intentions in making microblogging posts. They found that people primary talk about daily activities and share or seek information on Twitter. Our work studies how activities of users are related to change in information diffusion.

Popularity of Tweets Retweet is an important behavior in social networks. A wide array of techniques have been proposed to model and predict its trend. For instance, Ma *et al.*[1] proposed to use the piecewise sigmoid function to model the popularity of a single tweet. Our work extends the previous research in that we propose LGM to model the distribution of parameters in the sigmoid function of a batch of tweets. Yang *et al.*[15] designed novel models to capture the speed, scale, and range of users' ongoing social interactions. The popularity of news items on social networks can be predicted using regression and classification algorithms[14]. Hong *et al.*[16] investigated factors that influence information propagation in Twitter to predict the number of future retweets.

Correlation Analysis Several studies have analyzed online social networks together with data from other sources and fields to conduct deep correlation analysis. Mao *et al.*[17] combined Twitter, news and other data to make financial prediction of market indices. Zhang *et al.*[18] studied the link between Twitter and the e-commerce platform (eBay), and discovered the correlation between social media and e-commerce events. Bollen *et al.*[19] predicted the stock market activity by tracking the moods of daily Twitter feeds through a self-organizing fuzzy neural network. Besides prediction problems, Ruiz *et al.*[20] further analyzed features in microblogging activities that have strong correlations with changes in stocks of companies.

VII. Conclusion And Discussion

In this paper, we present our in-depth analysis after collecting a huge amount of Sina Weibo dataset. The modeling of the popularity of tweets leads to the discovery that Malthusian Parameter is quite suitable to describe the collaborative behaviors. We perform statistical analysis and use the ARMA model to predict the its trend, which shows the prosperity and decline of Sina Weibo. Furthermore, we extract fine-gained novel features from the dataset of both Sina Weibo dataset and search engine data. Through the systematic study, we learn that several aspects of collective behaviors have great impact on the platform's capability of information propagation. We also show the public attention on different social media influence activities of users on Sina Weibo.

Our study can be useful to both social science and computer science. While the analysis on prosperity and decline process in Sina Weibo gives social scientists clues on information propagation, we should also notice those features could improve prediction accuracy of various measurements. For example, those extracted features could be combined to obtain a more accurate prediction of Malthusian Parameter. Furthermore, it also points out users' collective behavior can be useful in predicting popularity of tweets or events.

Acknowledgment

This work is partially supported by National High-tech R&D Program (863 Program) under grant number 2012AA011003, and National Science Foundation of China under grant numbers 61432006 and 61170086. The authors would also like to thank Ping An Technology (Shenzhen) Co., Ltd. for the support of this research.

References

[1] H. Ma, W. Qian, F. Xia, X. He, J. Xu, and A. Zhou, "Towards modeling popularity of microblogs," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 171–184, 2013.

[2] T. R. Malthus, *An essay on the principle of population*. Cosimo, Inc., 2013, vol. 1.

[3] M. Celis, J. Dennis, and R. Tapia, "A trust region strategy for nonlinear equality constrained optimization," *Numerical optimization*, vol. 1984, pp. 71–82, 1985.

[4] C. C. Heyde, "On a property of the lognormal distribution," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 392–393, 1963.

[5] R. G. Miller Jr, *Beyond ANOVA: basics of applied statistics*. CRC Press, 1997.

[6] C. Gourieroux, A. Holly, and A. Monfort, "Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters," *Econometrica: journal of the Econometric Society*, pp. 63–80, 1982.

[7] F. J. Fabozzi, S. M. Focardi, S. T. Rachev, and B. G. Arshanapalli, "Autoregressive moving average models," *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*, pp. 171–190, 2014.

[8] G. M. Ljung and G. E. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.

[9] G. E. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970.

[10] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," pp. 513–522, 2012.

[11] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," 2010.

[12] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010, pp. 591–600.

[13] A. Java, X. Song, T. Finin, and B. L. Tseng, "Why we twitter: An analysis of a microblogging community," in *WebKDD/SNA-KDD*, 2007, pp. 118–138.

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

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

[16] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *WWW (Companion Volume)*, 2011, pp. 57–58.

[17] H. Mao, S. Counts, and J. Bollen, "Predicting financial markets: Comparing survey,news, twitter and search engine data," *CoRR*, vol. abs/1112.1051, 2011.

[18] H. Zhang, N. Parikh, G. Singh, and N. Sundaresan, "Chelsea won, and you bought a t-shirt: characterizing the interplay between twitter and e-commerce," in *ASONAM*, 2013, pp. 829–836.

[19] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *J. Comput. Science*, vol. 2, no. 1, pp. 1–8, 2011.

[20] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *WSDM*, 2012, pp. 513–522.