Distinguishing Re-sharing Behaviors from Re-creating Behaviors in Information Diffusion

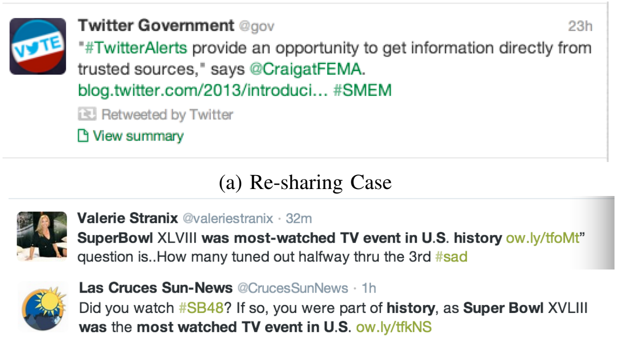
Yiran Xie*, Hongzhi Yin†, Bin Cui*, Junjie Yao‡, Quanqing Xu§
* Key Lab of High Confidence Software Technologies (MOE), School of EECS, Peking University, China
† The University of Queensland, School of Information Technology and Electrical Engineering, Qld 4072, Australia
‡ East China Normal University
§ Data Storage Institute, A*STAR, Singapore
*{xie.yiran, bin.cui}@pku.edu.cn, †h.yin1@uq.edu.au, ‡junjie.yao@sei.ecnu.edu.cn,§xu_quanqing@dsi.a-star.edu.sg

***Abstract*—Social media plays a fundamental role in the diffusion of information. There are two different ways of information diffusion in social media platforms such as Twitter and Weibo. Users can either re-share messages posted by their friends or re-create messages based on the information acquired from other non-local information sources such as the mass media. By analyzing around 60 million messages from a large micro-blog site, we find that about 69% of the diffusion volume can be attributed to users' re-sharing behaviors, and the remaining 31% are caused by users' re-creating behaviors. The information diffusions caused by the two kinds of behaviors have different characteristics and variation trends, but most existing models of information diffusion do not distinguish them. The recent availability of massive online social streams allows us to study the process of information diffusion in much finer detail. In this paper, we introduce a novel model to capture and simulate the process of information diffusion in the micro-blog platforms, which distinguishes users' re-sharing behaviors from re-creating behaviors by introducing two different components. Thus, our model not only considers the effect of the underlying network structure, but also the influence of other non-local information sources. The empirical results show the superiority of our proposed model in the fitting and prediction tasks of information diffusion.**

I. Introduction

Social media platforms represent a fundamental medium for the emergence and diffusion of information. For example, a rumor or a piece of news propagates in the form of users' re-sharing messages in micro-blog platforms. However, due to the emergence of mass media, like TV stations and online news sites, the information reaches us not only through the social neighbors but also through the influence of non-local information sources [1]. From the early stages of research on information diffusion, there has been the tension between effects carried by social structures and effects from non-local information sources.

However, it is hard to capture and study the effects of the social network and the non-local information sources directly. In this paper, we take the posts or messages that users re-share from their social neighbors (i.e., the local sources) as the proxy of the influence of social networks, and use the posts that users re-create based on the information acquired from the non-local sources to represent the effects from the external information sources. For example, when information appears at a node that has no connections to nodes that have previously mentioned the information, the emergence of information at that node can only be explained by the influence of some unobservable exogenous causes, like others posts or other online sources.This allows us to study processes of information diffusion and emergence in much finer detail than ever before.



Twitter Government @gov 23h
"#TwitterAlerts provide an opportunity to get information directly from trusted sources," says @CraigatFEMA.
blog.twitter.com/2013/introduci… #SMEM
Retweeted by Twitter
View summary

(a) Re-sharing Case

Valerie Stranix @valeriestranix · 32m
SuperBowl XLVIII **was most-watched TV event in U.S. history** ow.ly/tfoMt”
question is..How many tuned out halfway thru the 3rd #sad

Las Cruces Sun-News @CrucesSunNews · 1h
Did you watch #SB48? If so, you were part of **history**, as **Super Bowl** XVLIII **was** the **most watched TV event in U.S**. ow.ly/tfkNS

(b) Re-creating Case

Fig. 1: Two Main Ways of Information Diffusion on Micro-blog

Take the example of current popular micro-blog systems, Twitter[1]. Users generate a torrent of correlated messages, and these messages can be categorized into two different types based on the sources of information, including *re-sharing messages* and *re-creating messages*. We show them in Figure 1.

For one thing, when reading an interesting message posted by a friend, users can re-share and comment it. Figure 1(a) is a piece of re-sharing message from Twitter. The official account of Twitter re-shared one of its sub channel's new feature announcement. As demonstrated in recent literatures [1], [2], [3], [4], [5], users' re-sharing actions result from the underlying social network. That is to say, the information propagates over the edges of social network in the form of re-sharing messages.

For another, Figure 1(b) shows two content-similar messages. Both of them are about NFL Super Bowl XLVIII with similar keywords like "superbowl", "most watched", but they have different posting time. Moreover, there are no explicit connections between these two messages, which is different from the re-sharing case in Figure 1(a). Due to the emergence of mass medias and social medias, people can acquire information and accept new opinions from various

[1] twitter.com

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

non-local information sources. Hence, there are many content-similar messages without explicit social connections in the social media platforms. We call these messages as re-creating messages.

The volume of messages that carrys the same or correlated information changes over time, which is an important quantitative measurement for studying phenomenons of information diffusion. Modeling the temporal dynamics of a group of correlated messages is essential to better understand and analyze the public attentions and opinions as well as their trends, which is of great importance for the design of many applications, such as to improve the supervision and perfect net forewarning mechanism.

However, most of existing models of information diffusions in micro-blogs assume that information only passes from a node to another node via the edges of the underlying social network in the form of re-sharing messages [2], [3], [4], and ignores the influence from non-local sources. Recently, some works have focused on the design of tracking systems to cluster correlated messages based on their hashtags or contents [6], [7]. Although many models are proposed to use these clusters of messages to analyze the information diffusion [5], [8], [9], [10], they do not distinguish the effects of the underlying social networks (i.e., re-sharing messages) from the influences of the non-local information sources (i.e., re-creating messages), and ignore an important observation that the two types of messages have different characteristics and variation trends. Here, we will pay more attention to it, from tracking topics to modeling the temporal dynamics.

In this paper, we first design a tracking framework to collect groups of correlated messages from a real micro-blog system - Sina Weibo. We then categorize messages in each group into re-sharing ones and re-creating ones. The results show that about 69% of the correlated messages are produced by users' re-sharing behaviors, and the remaining 31% are caused by users' re-creating behaviors. In other words,about 69% of the information volume in Sina Weibo can be attributed to network diffusion, and the remaining are due to the influences of non-local information sources. After that, we investigate and analyze the two types of correlated messages in the process of information diffusion, and find they have different temporal characteristics. Based on the analysis results, we propose a novel model to capture and simulate the process of information diffusion in the micro-blog platforms, which distinguishes users' re-sharing behaviors from re-creating behaviors by introducing two different components. The empirical results show the superiority of our model.

II. Framework Of Tracking Correlated Messages

In this section, we will introduce how to group messages and obtain threads of a message stream. Each thread represents the diffusion procedure of a certain topic, which is the basis of temporal characteristic analysis and diffusion modelling.

Similar to Twitter, Sina Weibo is a Chinese microblogging (weibo) website and one of the most popular sites in China. In the following analysis and experiments, we use a large-scale dataset collected from Sina Weibo, spanning from July 2013 to Dec 2013, with about 60 million messages. The raw data is about 67G in size. Each message contains some important information, including the related user, text, date and so on. Given the raw data, we propose a tracking framework to collect and group correlated messages continuously.

Definition 1: (**Message Stream**). A message stream is a line of messages $m_1, m_2, ..., m_i, ..., m_n$ ordered by the published date.

The messages inside a stream are not isolated. They might connect with others either through explicit re-sharing relation or implicit semantic relevance. Given two messages m_i and m_j, and m_j posted later than m_i, connection types between them can be categorized into "re-share" and "re-create" according to the definitions in Table I.

TABLE I: Different Connections Between Messages

Type	Feature	Condition
Re-share	RT	m_j re-share m_i
Re-creation	URL	$url(m_j) \cap url(m_i) \neq \varnothing$
	hashtag	$hashtag(m_j) \cap hashtag(m_i) \neq \varnothing$
	text	$text(m_j) \cap text(m_i) \neq \varnothing$

As we have known, there are two kinds of messages in information diffusion, i.e., re-sharing messages and re-creating messages. RT represents the behavior that one directly re-shares the previous message or adds some comments at the same time, which forms the re-sharing connections. For the re-creating messages, the common URLs, hashtags and words between two messages can be used to measure the their semantic similarity, and these similar messages usually take on slightly different appearances. Though there might be no direct connections between those messages, the semantic similarity illustrates the implicit relationship between them. In fact, users might acquire information from exogenous sources such as mass media, and then re-create and propagate it in the form of message. Given a piece of re-creating message, there are a lot of semantic-similar messages with it, and we choose the strongest semantic link as the re-creating connection. Re-sharing and re-creating connections group all the messages of a topic into a complete process of information diffusion.

In order to measure the similarity of re-creating behaviors, we propose a scoring function as Equation (1), which combines various semantic factors (e.g., url , hashtag and text) and time factor.

$$\begin{aligned} S(m_i, m_j) = & \alpha |url(m_i) \cap url(m_j)| \\ & + \beta |hashtag(m_i) \cap hashtag(m_j)| \\ & + \gamma |text(m_i) \cap text(m_j)| \\ & + \rho |date(m_i) - date(m_j)| \end{aligned} \tag{1}$$

Where m_i represents a piece of candidate message and m_j is an incoming message. The function aggregates some indicant closeness measuring functions. Intersection is defined as Jaccard similarity between two texts. The time difference between them is also taken into consideration. The intuition behind is that messages with the same or similar posting time are more likely to describe the same topic, so they are more correlated. Besides, α, β, γ and ρ are parameters to tune the weight, which can be manually set to reflect system requirements.

It is considered that between m_i and m_j exists a kind of re-creating relationship, if the value of $S(m_i, m_j)$ is more than a threshold. Obviously, there may be more than one message which is similar to m_j, so we consider the one with maximum scored $S(m_i, m_j)$ as the former node who is most similar to it. While clustering correlated messages, they will be put together.

We try to put correlated messages together to get a set of non-overlapping groups, and only retain the message connections inside the group [11]. Connections between two messages m_i and m_j derive from re-creating or re-sharing behaviors.

Definition 2: (**Message Group**). A message group is a bundle of messages, where messages within the group tend to talk about the same topic. Message re-sharing connections and maximum scored re-creating connections within group are preserved, and intra-group connections are skipped.

Here we propose two strategies to preserve the message development trail in this way:

- Every group keeps its own representative information locally. For the efficiency of this allocation task, we will compare the incoming message with the centroids of the existing groups instead of all messages.
- One message either follows the forwarding direction, or only retains a maximum scored connections with its priori similar one, while the low scored connection will be omitted.

Definition 3: (**Thread**). A thread is associated with a group, which represents the variation volume $x_1, x_2, ..., x_T$ at each discrete period $t = 0, 1, 2, ..., T$. Because one group is composed by two kinds of messages (i.e., re-sharing and re-creating ones), one thread can be divided into two classes, *Re-share Thread* for re-sharing messages, and *Re-creation Thread* for re-creation messages.

Given a message stream, we try to group correlated messages, and derive threads from groups. Volumes of messages at discrete periods sequentially form a thread in a group, which becomes the foundation of analyzing and modeling the message threads.

III. Analysis of Re-sharing and Re-creating Messages

In the following we proceed to analyze characteristics of re-creating behaviors and re-sharing behaviors. The rise and decay analysis of the thread are demonstrated. Meanwhile, the temporal dynamics of the re-creating and re-sharing behaviors are emphasized specifically. Both behaviors and the global behavior show different characteristics.

A. Re-sharing Percentage

In each group, around 69% messages are attributed to the re-sharing action, which means re-sharing action holds the dominate position. Around 86% messages are reposted, and each one is shared averagely for 9.92 times. 55% of total re-sharing actions happen on the first day, 79% of tem take place in the first week, which is a little slower than Twitter [2] due to the complexity of social network in China. The re-creating action, is of great importance to model the dynamics of information diffusion, although it accounts for about 31% of messages. Re-creating messages are comparatively fewer because of the complicated thinking and replicating process, while re-sharing behaviors are simpler.

B. Temporal Analysis

1) Rise Analysis: Figure 2 shows the volume accumulation at some stages of its life after a smoothing for temporal dynamics. Compared to the basic line connecting diagonal, the volume increases more slowly at the beginning, which is a preparing period and there are only a few messages posted as the information may not spread out, while users become more active at the rear part. On the whole, the volume increases evenly without special jumps.

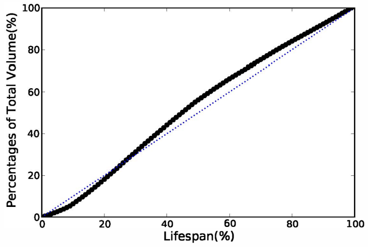


Fig. 2: Relation between volume accumulation and lifespan

2) Decay Analysis: We examine how the dynamics decline in every group. Figure 3 illustrates the distribution of volume and time in a log-log scale. We find a stable power-law exponent of around -1.7, which shows a bit slower than the burst nature of human behavior [12], because longer interactive time is needed between human and media.

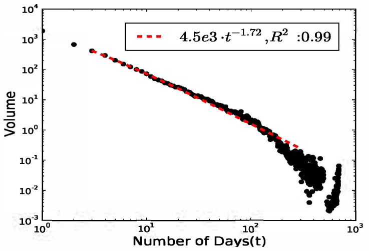


Fig. 3: Decay in log-log scale fitting power-law function with exponents -1.72

Comparing with the curve changes of Figure 2, volume here keeps declining on the whole. In fact, most groups live a short life. These short-lived groups, mainly contribute to the opening part of Figure 3. So *Rise Analysis* and *Decay Analysis* with different horizontal axis draw incomparable shapes.

C. Correlation Analysis

1) Peak Analysis: The *peak* is the max volume of one variation, at the peak time t_p. As we expected, the peak is not always at the median of temporal dynamics. By examination, it appears around 15.0% of the whole lifecycle. However every peak situates at different positions, so we shift all the peaks at the same position with $t_p = 0$, and normalize the volume of the peak to 1, in order to easily make analysis. Furthermore, we put the peak dynamics of re-sharing and re-creating behaviors respectively into the same figure for easy comparison.

Interestingly, the statistical results show that the peaks of two different kinds of messages are basically consistent with that of the whole stream. For sharing actions alone, the peak averagely appears at the 15.3% of the lifespan. It can be attributed to the leading position of re-sharing behaviors in quantity. For re-creating actions, the volume goes up a bit earlier at the 12.6% of the lifespan. Actually, Figure 4 also confirms this fact. During 30 hours before the peak time, the trend of re-sharing actions is nearly in accord with the original stream, while the trend of re-creating ones is more active. After the peak time, though we can find a small fluctuate appearing, the re-sharing line is close to the original stream all the time, which goes down smoothly. Meanwhile, the re-creating line rises up earlier and falls down a little more slowly than others. This is because this kind of behavior needs to cost some time in acceptance of messages from multiple sources. However, its contrasting performance does act to the original variation.

Usually people prefer to follow others or to share interesting information, and meanwhile significant news is more attractive to people, so the high time is shorter but the volume is bigger in the re-sharing situation. However, influenced by mass media, people experience a process of discovery, selection and consideration to post textual similar messages, which costs much more time. In fact, many texts are posted within several hours, including massive junk messages that mostly belong to the re-sharing actions. Therefore we can find that the peak of micro-blog is taller and slender.

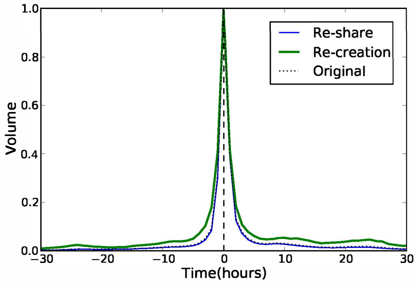


Fig. 4: Volume increases and declines over 30 hours respectively.

2) Rise Analysis: In order to investigate the rise thoroughly in a specific sight, Figure 5 illustrates how many days are needed for the volume to accumulate to the certain percentages of the peak. Meanwhile, the variations of correlated texts with corresponding peaks are shown together in Figure 5.

Interestingly, the lines are regularly easy to be divided into three periods, for instance, the original one at the percentages 6% and 40%. In the first period, it increases very fast, which means the number of messages grows up with much more time than the later periods. After the slow-growth stage, the development steps into the normal period, but also experiences a little stop at half the way to the top. Finally, it goes up to the top faster, which means the trend goes hot. The closer to the peak, the faster the growth changes. Besides, a stop before a peak is found.

The re-sharing action goes into three steps earlier than the original, but keeps step with it soon. We can infer that re-creating curve increase faster, because they use less time to achieve certain volume, which is harmonious with peak analysis. Meanwhile, compared to the other lines, re-creating actions have a longer and more gentle preparing period, which is because the re-sharing actions show quicker reaction than similar text diffusion in mass media.

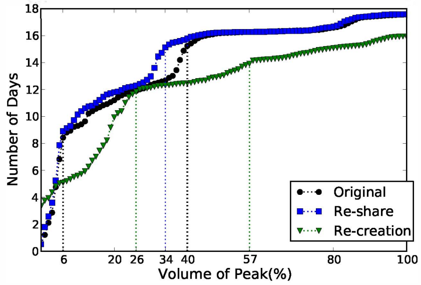
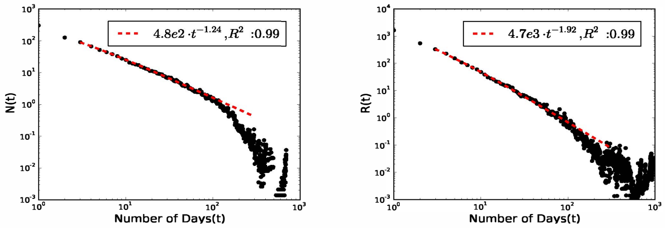


Fig. 5: Time spent to go up to peak

3) Decay Analysis: We derive two kinds of clusters according to the re-sharing behaviors and the re-creating behaviors, respectively. We have found the general decay of correlated texts. Next, we will discuss how the trend persists under the two different kinds of clusters respectively.

For the re-creating messages, users post information when they accept new opinions from the real world, or read texts from other sites. Let $N(t)$ denote the number of the re-creating messages at time t of all groups. Figure 6(a) presents the distribution of $N(t)$ over t, which perfectly follows a power-law with exponent -1.24. Meanwhile, $R(t)$ denotes the number of the re-sharing messages, and Figure 6(b) shows the distribution of the $R(t)$ over t, which follows a power-law with exponent -1.92. Re-creating behaviors keep more persistent. We can loosely think that the re-sharing action goes down faster than re-creating action, resulting from the faster diffusion through re-sharing actions.



(a) Decay of Re-creation

(b) Decay of Re-share

Fig. 6: Decays of both behaviors follow power-law distribution

As above, we have mainly analyzed temporal dynamics of two kinds of behaviors. These characteristics are the outside image of their different mechanisms. Besides investigating difference between re-sharing and re-creating behaviors, all the factors will also be considered into modeling threads in the next section, and meanwhile become the important parts.

IV. Proposed Model

So far, we have gotten plenty of groups talking about the same topics, and corresponding threads from a message stream. The threads of these groups are temporally changing, resulting from users' different opinions and behaviors. Modeling threads is important to understand and predict users' behaviors. We have investigated the different characteristics of re-creating behaviors and re-sharing behaviors. Next, we emphasize how the correlated texts contribute to modeling threads. As it follows the features of groups, good performances in fittings and predictions are achieved.

A. *Basic Model*

As the former definition, we shift all the threads, make them begin at time 0, and assume that time runs in discrete slices $t = 0, 1, 2, ..., T$. During each time slice, a set of short messages might be published, influenced by different driving factors. Specifically, we use x_t to denote the new increasing volume at the time period $[t-1, t](t \in [1, 2, ..., T])$ of one thread. So far, we can present a thread of one group as $x_1, x_2, ..., x_T$.

Usually, it is hard to confirm when a dynamic exactly goes into the state. We define the *beginning period* of thread as the time period with a set of small and unimportant fluctuates before the topic really raises the common concern. It is the first period discussed in the *Rise Analysis* of last section. The begining period is insignificant, but brings lot of noises to model the truly valuable dynamics.

Here, we use variable t_0 to represent the beginning time of a real variation, which is used to divide the original thread into two parts, including the beginning period and the remaining period. After the time t_0, a wave of users post their opinions, which indicates the real variation begins.

In a group, we can consider that diffusions begin with a source node, and users tend to follow a more popular one, which gradually develops an approximate scale-free network. According to the theory of *preferential attachment* [13], we denote $\Gamma(k)$ as the proportion of existed connection k of a set of nodes to that of all the nodes during a period, representing the existence of preferential attachment. We approximately consider it follows a power law, i.e., $\Gamma(k) \sim k^{\alpha}$. Supposing that there are x_{t-1} messages (i.e., nodes) at the time $t-1$, the number of increasing edges of the updated graph at the time t could be nearly considered as x_{t-1}^{α}. Here α describes the growth factor of the group.

Moreover, besides influence from social network, unknown influence outside should be taken into consideration. In every discrete time slice, the outer influence, such as interest or bursts, also leads to several connections and gives the opportunity to the evolvement of social network. Therefore, preferential attachment here has a more appropriate form, $\Gamma(k) \sim (k+b)^{\beta}$. Besides the effect of the stable social structure, the parameter b controls the likelihood of a new node appearing because of outside unknown reason is discovered. β denotes the speed of growth in this scenario. On the whole, $f(x_{t-1}) = (x_{t-1}+b)^{\beta}$ is proportional to the incasement of new messages for the time t in the groups.

However, a piece of information cannot be spread out indefinitely for a very long time, due to the delay of human response and the capacity of the network. Time directly leads to the limitation [14], and the delay usually takes a heavy-tail form. The second ingredient, depending on the age of diffusion, should be taken into account while modeling streams. $t^{-\gamma}$ represents the decay of the attraction of the information. As described above, let $g(t) = (t-t_0)^{-\gamma}$, adjusting the beginning point of the thread, helps better understanding of diffusion.

So far, two basic ingredients, preferential attachment $f(x_{t-1})$ and decay $g(t)$, have proved to influence the extension in the next time period and should be taken into account while modeling thread.

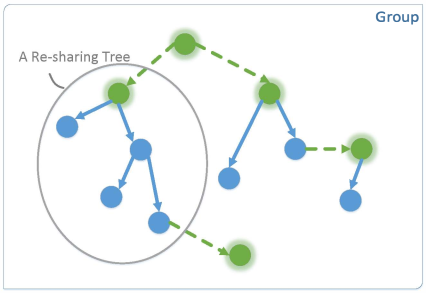


Fig. 7: Group Case

Based on the above analysis, we can get a basic model as follows:

Basic Model

$$x_t = a \cdot f(x_{t-1}) \cdot g(t)$$

We use parameter a to denote a normalization constant for the full distribution, which is directly related to the interest degree of the activity of the group. The basic model combines two main factors, which can right describe general temporal dynamics.

Though the model considers the generating process of information diffusion, it still cannot bring full expression to the minor variations, which is same as most current modeling methods. More detailed analysis of the re-creating behaviors and re-sharing behaviors is required to improve the accuracy.

B. *Joint Model*

In the following, we present a novel model based on above theory to accurately depict the thread. The model tries to capture the ordinary characteristics of correlated texts.

We show the inner structure of one group in Figure 7. One node associates with an informed user with its posted message, big green nodes for re-creating messages and blue nodes for re-sharing messages. They are linked by two types of main connections. The broken green lines and the full blue lines denote re-creating and re-sharing connections respectively.

In one group, if we ignore the re-creating links, several re-sharing trees are left over. Each node is included in one and only one re-sharing tree, so the volume over time of a thread can be considered as a sum of messages in every re-sharing tree of a group. We can directly represent it as follow:

$$x_t = \sum_{k=0}^{\#rc} rs(k)_t$$

where $\#rc$ denotes the number of re-sharing trees existed at time t, and the formula sums up the re-share volumes. $rs(k)_t$ represents the volume of the k-th re-sharing tree at time t.

Considering the procedure that we build up a group, firstly there always comes re-creating messages. Reading these texts, users may share it if they are interested in. As time goes by, more and more re-sharing messages will emerge. In order to make it simpler at expression and prediction tasks, we have the following two assumptions:

- Re-sharing behaviors in a group follow the same pattern.

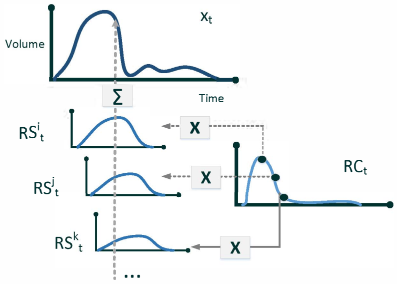


Fig. 8: RRM models a thread over time as a sum of messages at every re-sharing process in a group

- Re-sharing behaviors in a group are distinguished only by the time factor.

So far, re-sharing behaviors can be represented as below:

$$rs_t^{(i)} = \begin{cases} a_s \cdot {rs_{t-1}^{(i)}}^{\alpha} \cdot (t-t_0)^{-1.92} & t > i \\ a_s \cdot (t-t_0)^{-1.92} & t = i \\ 0 & t < i \end{cases}$$

We denote $rs^{(i)}$ as the re-share thread that begins at time i, and $rs_t^{(i)}$ represents the volume of this re-share thread at the time t. At time i, a piece of re-creating message comes, it becomes the root of the re-sharing tree, and users may re-share the re-creating texts at the same time. After that, at every other period of time, $rs_t^{(i)}$ will be updated from $rs_{t-1}^{(i)}$ by the re-sharing behavior, because the new re-sharing nodes in this tree rely on the former re-sharing nodes. Besides, t_0 here acts on all the re-sharing behaviors and the re-creating behaviors, implying the beginning time of the thread.

On the other side, a group might collect plenty of re-creating messages, which means that one group is very like a re-sharing forest. If the number of re-creating messages at every time slice and the varying pattern of re-share threads are known, the thread of the group can be estimated. For the re-creation thread, we try to model just as the basic model does:

$$rc_t = \begin{cases} a_c \cdot (x_{t-1} + b)^{\beta} \cdot (t-t_0)^{-1.24} & t \geq t_0 \\ 0 & t < t_0 \end{cases}$$

rc_t denotes the volume of the re-creation thread at the time t. A re-creating node can be considered as a root node, so it corresponds one re-sharing tree with a series of re-sharing nodes. Users who post re-creating messages are influenced by the non-local sources, i.e., the sematic-similar influence collected from the social network, as well as the unknown influence indicated by a small constant b. Unlike the re-sharing behavior, the re-creating behavior has just one variation curve, implying the incoming number of re-sharing trees.

Integrating re-creating and re-sharing behaviors as above, we propose a joint model, named *Re-creation and Re-share Model (RRM).*

Joint Model (RRM)

$$x_t = \sum_{i=0}^{t} rc_i \cdot rs_t^{(i)}$$

In a conclusion, RRM models a thread by two main steps, shown as Figure 8. The curve on the top represents the thread $\{x_t\}$ over time. The rs_i, rs_j and rs_k depict re-share threads that begin at the time i, j and k respectively. rc_i denotes the number of re-sharing threads rs_i, and therefore two kinds of curves jointly contribute to the thread modeling. Modeling by the RRM, we should fit out the re-creation thread and the re-share threads, so as to induce several possible spreading structures in the group, which can approximately illustrate process of information diffusion.

Model Parameter Estimation Due to multiple-step modeling procedure, general optimization techniques, such as *Levenberg-Marguard* (LM) [15], cannot be used to estimate parameters directly. Therefore we propose an efficient algorithm to accurately estimate parameters. Here, the joint model consists a set of parameters: $\theta = \{a_c, a_s, b, \alpha, \beta, t_0\}$. Considering a real thread that talks about one topic, $y_1, y_2, ..., y_T$, as well as the known re-creation thread, $c_1, c_2, ..., c_T$, we try to give a sequence $x_1, x_2, ..., x_T$, agreeing with the characteristics of the topic according to our model. Here, we try to extend LM algorithm to minimize the sum of errors: $D(x, \theta) = \sum_{t=0}^{T}(x_t - y_t)^2$, and as far as possibly minimize $P(rc, \theta) = \sum_{t=0}^{T}(rc_t - c_t)^2$.

Therefore, like classical optimization methods, we first need some initial seeds for subsequent LM operations. Next, there are two steps when update new values interatively: 1) determine the number of re-creating nodes at every period by fitting known re-creation thread, and 2) model the thread while learning re-sharing behaviors. In every iteration, we update new parameters and new fitting sequences, and let them become the seed of the next iteration. The optimization will be stopped until updated values converge.

V. Experimental Evaluation

Based on the groups of correlated messages derived from the Sina Weibo dataset, we carry out extensive experiments to evaluate the performance of our model in the following tasks associated with information diffusion: (1) to match the threads of several groups; (2) to make tail-part predictions.

A. Thread Fitting

This experiment studies the ability of our model (RRM) in thread fitting, and we first construct threads from the collected groups of messages. For comparison, we implement a state-of-the-art information diffusion model, SpikeM [9]. It is a kind of analytical model over the process of information diffusion, avoiding several problems that prior models may have, as well as following the power-law fall pattern and periodicity. It can match real data, and deal with some meaningful tasks, such as predictions. Moreover, it supports the most similar applications as our model.

To quantitatively evaluate the performance of our model and SpikeM, we adopt the classic metric *RMSE* (i.e., the root mean square error) that is widely used to measure the difference between the true values and fitting values. In our experiment, RMSE is computed as follows:

$$RMSE = \sqrt{\frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} (x_t - y_t)^2} \tag{2}$$

where t_1 and t_2 represent the starting and ending time of the test thread, respectively.

To make an overall evaluation of RRM and SpikeM, we first test them on all the threads. The RMSEs of RRM and SpikeM are 15.26 and 18.42, respectively. Our model archives approximately 20% relative improvement over SpikeM. Although both SpikeM and RRM are based on the law of information diffusion, RRM distinguishes users' re-sharing behaviors from re-creating behaviors, which enables us to model and capture the process of information diffusion in micro-blog platforms more accurately.

To further analyze why our model RRM performs better than SpikeM in an intuitive way, we choose three test threads for case study in Figures 9(a-c). Table II shows the fitting accuracy of both models. From the results, we observe that RRM achieves higher accuracy although SpikeM can also capture the basic temporal trends of the threads. Specifically, RRM is better at processing small fluctuates before t_0, and SpikeM might be affected more by the beginning part especially in last two cases. Case (b) is a tough thread, the curve of which is difficult to fit since it varies irregularly. Our RRM can still better fit this case, showing the advantage of distinguishing users' re-sharing behaviors from re-creating behaviors, while SpikeM can only fit the cases with periodic patterns.

TABLE II: Fitting Accuracy of the Models (RMSE)

	Group a	Group b	Group c
RRM	13.47	10.48	14.86
SpikeM	16.32	12.29	15.19

B. Tail-part Prediction

This experiment studies the performance of our model in the prediction task. Being different from the thread fitting task, we divides each thread into the training part and test part in this task. We first train our model RRM and SpikeM in the training part, and then use the estimated model parameters to predict the test part, i.e., the future variation trend of the curve. The metric RMSE is used to evaluate the prediction accuracy.

We first compared the overall performance of both models, 13.33 for RRM and 26.72 for SpikeM. Then, we choose three cases to perform further analysis, and Figure 10 shows the results of the three cases. We use a vertical line to separate the training and prediction parts. Table III present the prediction accuracy of the three cases. From the results, we observe that our proposed RRM achieves more accurate predictions because RRM can capture and model the process of information diffusion in much finer detail. In other words, RRM benefits from distinguishing users' re-sharing behaviors from re-creating behaviors. Specifically, from Figure 10(a) we observe that RRM can more accurately predict the rising trends of threads, and Figure 10(b) shows that our RRM performs better than SpikeM in the case that predicts the declining trends of threads. The case in Figure 10(b) is difficult for SpikeM to accurately predict because of its complicated variation trend, while our RRM still achieves good performance in these complicated cases. In other words, SpikeM can only predict the simple variation trends and fails in the relatively complicated ones because SpikeM cannot capture the weak signal of variation.

VI. Related Work

TABLE III: Prediction Accuracy of Models (RMSE)

	Group a	Group b	Group c
RRM	6.65	24.22	14.86
SpikeM	10.98	63.86	15.19

The emergence of social stream has generated huge online contents, and attracted much interest to investigate its mechanism and characteristics [9], [16], [17], [18].

Correlated Texts As a kind of correlated texts, memes were broadly studied and applied. These works find robust ways of extracting and identifying the mutational variants of distinctive phrases, such as phrase graph [6] or provenance-based solution [11]. Current related works pay more attentions to detecting and tracking problems [6], [7], [11], [19], but there is rear rise and fall analysis of correlated texts on micro-blog. They open an opportunity to analyze and model information diffusion in more detail.

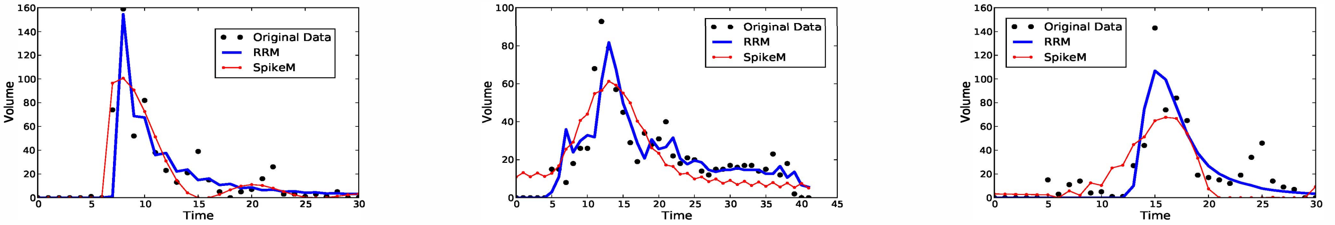
Characteristics of Temporal Dynamics The statistical mechanics of complex network had been discussed for a long time. Growth and preferential attachment are presented as two important ingredients for generating social network [13]. The evolvement of social media led to deeply analyze the topics of retweet cascades and following behaviors [2], [3], [4], [18], [20], [21], [22], [23], [24]. After proposing that information diffusion also depending on its age [14], the power-law decays of influence in the different situations have been reported, such as in blogs with -1.5, and -1 for tweets and the human response time [12], [22].

Modeling Temporal Dynamics Traditional approaches were widely applied in the area of modeling and predicting temporal dynamic, including some regression and classification methods [8]. These methods are ease of use, but lack deep understanding of information diffusion. Inferring model by experiments or mathematic theories is more difficult [1], [5], [9], [21], [25], [26], [27], [28], [29], [30]. Besides, some probability based methods were applied to explain information diffusion, such as modeled with exposures [1] and social selection [31].

After years of efforts, these coarse-grained ways for modeling temporal variations had done well. In our paper, we focus on modeling with the re-sharing and re-creating components and put forward a new angle to improve effectiveness of modeling information diffusion.

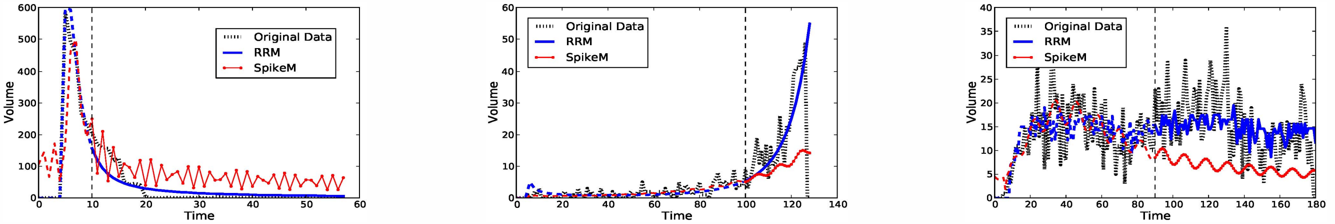
VII. Conclusions

In this paper, we first designed a tracking framework to collect groups of correlated messages from a real micro-blog system - Sina Weibo, and then divided the messages into re-sharing ones and re-creating ones, which enables the study of the effects of underlying social networks and non-local information sources in information diffusion. We found that 69% of the information volume in Sina Weibo can be attributed to network diffusion, and the remaining is due to the influences of other non-local information sources. Moreover, we observed that the re-sharing and re-creating messages have different temporal characteristics after our detailed analysis and investigations. Based on the analysis results, we proposed a novel model to capture the process of information diffusion in micro-blog platforms. The experimental results showed the superiority of our model and verified the advantages of distinguishing users' re-sharing behaviors from re-creating behaviors in the information diffusion.



(a) Group #342679 (b) Group #335717 (c) Group #342958

Fig. 9: Results of Model Fitting on Three Threads



(a) Group #336939 (b) Group #339427 (c) Group #336855

Fig. 10: Results of Tail-part Prediction

Acknowledgment

The research is supported by the National Natural Science Foundation of China under Grant No. 61272155 and 61272340, 973 program under No. 2014CB340405, and partially supported by the Australian Research Council (Grants No. DP110103423 and No. DP120102829).

References

[1] S. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in network," in *Proc. of KDD*, 2012.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proc. of WWW*, 2010.

[3] 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 *Proc. of SocialCom*, 2010.

[4] S. Wu, J. M.Hofman, W. A.Mason, and D. J.Watts, "Who says what to whom on twitter," in *Proc. of WWW*, 2011, pp. 705–714.

[5] J. Yang and J. Leskovec, "Modeling information diffusion in implicit network," in *Proc. of ICDM*, 2010.

[6] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. of KDD*, 2009, pp. 497–506.

[7] C. Suen, S. Huang, C. Eksombatchai, R. Sosic, and J. Leskovec, "Nifty: a system for large scale information flow tracking and clustering," in *Proc. of WWW*, 2013, pp. 1237–1248.

[8] M. Gupta, J. Gao, C. Zhai, and J. Han, "Predicting future popularity trend of events in microblogging platforms," in *Proc. of ASIST*, 2012.

[9] Y. Matsubara, Y. Sakurai, B. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: Model and implications," in *Proc. of KDD*, 2012.

[10] N. I.Sapankevych and R. Sankar, "Time series preditcion using support vector machines: A survey," *Computaitonal Intelligence Magazine, IEEE*, vol. 4, no. 2, pp. 24–38, May 2009.

[11] J. Yao, B. Cui, Z. Xue, and Q. Liu, "Provenance-based indexing support in micro-blog platforms," in *Proc. of ICDE*, 2012, pp. 558–569.

[12] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, pp. 207–211, May 2005.

[13] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.

[14] S. Dorogovtsev and J. Mendes, "Evolution of reference networks with aging," *Physical Review E*, vol. 68, no. 1842, 2000.

[15] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[16] B. Cui, H. Mei, and B. C. Ooi, "Big data: the driver for innovation in databases," *National Science Review*, vol. 1, no. 1, pp. 27–30, 2014.

[17] X. Wang, L. Yu, J. Yao, and B. Cui, "A multiple feature integration model to infer occupation from social media records," in *Prof. WISE*, 2013.

[18] J. Cheng, L. A. Adamic, P. A. Dow, J. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. of WWW*, 2014.

[19] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen, "Lcars: A location-content-aware recommender system," in *Proc. of KDD*, 2013.

[20] E. Baskshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Identifying influencers on twitter," in *Proc. of WSDM*, 2011.

[21] D. Budak and A. El Abbadi, "Information diffusion in social networks: Observing and influencing societal interests," *PVLDB*, vol. 4, no. 12, pp. 1–5, 2011.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. Flance, and M. Hurst, "Cascading behavior in large blog graphs," in *Proc. of SDM*, 2007.

[23] G. Miller, "Social scientists wade into the tweet stream," *Science*, vol. 333, no. 6051, pp. 1814–1815, 2011.

[24] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, "A unified model for stable and temporal topic detection from social media data," in *Proc. of ICDE*, 2013.

[25] H. Yin, B. Cui, L. Chen, Z. Hu, and X. Zhou, "Dynamic user modeling in social media systems," in *ACM Transaction on Information Systems*, 2015.

[26] K. Radinsky, K. Svore, and S. Dumais, "Modeling and predicting behavioral dynamics on the web," in *Proc. of WWW*, 2012.

[27] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proc. of KDD*, 2010.

[28] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern, "Predicting information spreading in twitter," in *Proc. of Computational Social Science and the Wisdom of Crowds Workshop, NIPS*, 2010.

[29] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang, "A temporal context-aware model for user behavior modeling in social media systems," in *Prof. SIGMOD*, 2014.

[30] H. Yin, B. Cui, L. Chen, Z. Hu, and C. Zhang, "Modeling location-based user rating profiles for personalized recommendation," in *ACM Trans. Knowl. Discov. Data.*, 2015.

[31] K. Lewisa, M. Gonzaleza, and J. Kaufmanb, "Social selection and peer influence in an online social network," in *Proc. of PNAS*, 2012.