Exploiting temporal patterns of hot events in Weibo

Jiakun Huang[1], **Kai Niu**[2], **and Zhiqiang He**[2]
[1]Department of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China 100876
huangjiakun1991@126.com, {niukai,hezq}@bupt.edu.cn

Abstract— *With explosive growth of the Internet, microblog has become the largest source of public opinion. The propagation of hot events in microblog has drawn much concern. In this study, we extract 218 time series of hot events in 240 million tweets crawled from Sina-Weibo, the biggest Twitter-like microblog in China, and find that the diffusion process is divided into two step. Furthermore, the patterns can be clustered to several centroids by applying the K-Spectral Centroid (K-SC) clustering algorithm. The centroids are quite qualified for demonstrating the different information propagation features in weibo. We also introduce a modified SpikeM model to fit the centroids. Our results demonstrate that the new model describes all the rise and fall centroids with high accuracy, while SpikeM is only capable of fitting the first spike.*

Keywords: time series analysis, information propagation

1. Introduction

The emergence of microblog has dramatically changed the way people access to information. Due to its convenience and real-time property, people are increasingly engaged in sharing and consuming information in microblog services, which turns microblog to a form of online word of mouth branding [1]. In the case of Sina-Weibo, the biggest Twitter-like microblog of China, there exists 1.3 billion registered users and over 150 million monthly active users. So to some extent, weibo has become the dominate source of public opinion in the new media age.

Hot events reflect social opinion and impact the society both positively and negatively in return. The propagation of hot events has been a hot research topic. However, most of the researches focus on modeling propagation process over graph transmitting information from one node to another [2], [3], which are not suitable for large-scale social networks. Few researchers study the temporal dynamics of hot events. Yang et al. [4] propose a time series cluster algorithm K-Spectral Centroid, and discover six patterns of twitter topics. Yasuko et al. [5] introduce SpikeM, which is based on the so-called 'Susceptible-Infected' (SI) [6] model, performing well on fitting the six patterns. It shows that the temporal dynamics of hot events start with an exponential rise and a power-law decay, which is consistent to our observation in real data. But SpikeM is only applicable for the patterns with one spike or additional periodic tails, since it assumes there exists no 'revive' state in the social network.

As far as we know, the previous literature concentrates on modeling the time series of topic mentions. People participating in the discussion of online topics doesn't mean that they are unknown of the information. The periodicity of temporal dynamics directly owns to users' repeatedly participation of discussion. So topic mentions reflect the popularity of hot events, which is not directly related to the information propagation process. On the other hand, the reposting behavior correctly reflect the dynamics of public awareness over time. When a message is published, all the user's followers will have access to it. Secondary reposting behavior transmits the message to user's followers' followers, forming information cascade between disconnected nodes, which will spread to much more audiences. We focus on modeling reposting behavior to figure out the temporal patterns of information propagation.

The main goal of this paper is to discover how the diffusion process of hot events evolves over time, what kinds of temporal patterns are exhibited by weibo, and how to fit the patterns with high accuracy. First of all, our data set and basic statistical findings are introduced in Section 2. Then in Section 3, we use K-Spectral Centroid algorithm to cluster the time series of hot events, revealing that there exists three representative patterns in Sina-Weibo. In Section 4, a modified SpikeM model is introduced, which performs well as for modeling the diffusion process in Sina-Weibo.

2. Statistical Regularities

2.1 Dataset description

To obtain time series of hot events, we crawled more than 250 million tweets during a three-year period from 2012 to 2014. All the tweets are obtained through Sina Open API. Then 218 hot events are manually extracted from the dataset, according to the monthly reported hot events of Sina Weibo Data Center. Each hot event corresponds to an original tweet, with a retweet list filtered from the whole dataset. For the sake of simplicity, we use symbol consisting of a character "#" and a number to represent specified hot event, such as "#1" which is short for "the disappearing of MH370 on Saturday, 8 March 2014".

Table 1 gives several simples of hot events. Every retweet list is sorted by retweet time, but it needs to be quantized to create a time series of the amount of retweets per

Table 1: Four hottest events of Weibo in 2014

Symbol	Description
#1	The disappearing of MH370 on 8 March, 2014
#2	The 2014 Kunming terrorist attack
#3	The famous apology of Wen Zhang over affair
#4	The first Memorial Day of China

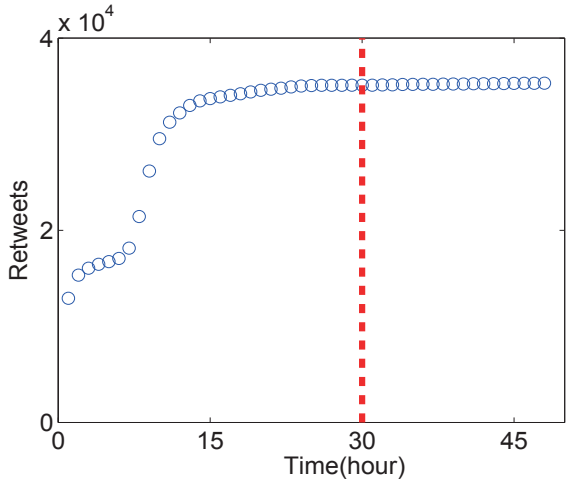

Fig. 1: CDF of #2

unit time interval. In Fig. 1, it shows that the shape of Cumulative Distribution Function has almost no increment after 30 hours, which means the spreading is completed in 30 hours and the subsequent points can be abandoned so as to concentrate on the analysis of preceding variable shapes. Further more, in order to get more variations of the time series shape, we specify the time unit to 10 minutes.

2.2 Findings

Two-stage process. Among the 218 time series, we observe that all these patterns contain two rise and fall spikes peaked at different time points, in which the first one is often much higher than the second one. Fig. 2 shows 4 hottest events in 2013 and 2014, and this unexpected phenomenon is quite different from the finding of six patterns in twitter. In fact, two spikes indicates that the information propagation process is divided to two stages in the life-cycle of hot event. The varying parameter of different hot events is spiking time and the proportion of the first peak and the second peak, indicating that similar diffusion process shares the same temporal pattern.

Causes of the two spikes. For the purpose of figuring out what actually gives rise to the two spikes, we turn to analyze the number of followers in each spike. In weibo social network, the so-called opinion leaders have a major impact on the public opinion, in most cases, and the number of followers is generally a convincing indicator for measuring their significance. In fact, the more followers they have, the more possibility that more people have access to the original message at one point in time. So it is sufficient to just focus on calculating the proportion of users with significant followers. More specially, this proportion is generally very small since the degree distribution social network has a power-law tail, indicating that small changes of the proportion might have huge consequences.

According to the official description, opinion leaders are those who have more than 5 million followers. In order to analysis the different influence of opinion leaders and grass roots, we first divide the number of followers into four levels, 4 to 7, which takes the logarithm base 10, and then calculate the occupation of each level in different spikes. For the most part, as we can see from Table 2, the occupation of users with large number of followers in the first spike, is significantly higher than the second one. As for users with follower count greater than 10^7, who are absolutely authoritative celebrities in Sina-Weibo, there always exists a small proportion in the first stage of diffusion, while in the second one, the proportion is generally zero.

Another special event #4, whose second spike possesses much higher peak value than the first one, is just presenting the opposite case. Furthermore, users with larger number of followers are correspondingly distributed in the higher spike. The above observations are consistent with other hot events in the dataset, strongly suggesting that the first stage of most information propagation process in Sina-Weibo is directly triggered by opinion leader, while the second long-lasting stage is generally caused by the crowd.

This observation is exactly consistent with the so-called Multistep Flow Model [7], which says that most people form their opinions under the influence of opinion leaders, who in turn are influenced by the mass idea. A small fraction of the hot events are exactly the opposite, representing that the information is first introduced by grass roots and propagated in a small scale of the social network, then it is detected by opinion leader which lead to widely spread of the information after several hours. Moreover, the consistency of t_F and t_P also shows that the time of peak point is quite related to the retweet time of users with the largest number of followers, which means opinion leader plays a very important role in the diffusion process.

From the last line of Table 2 we find that for social security events like #1 and #2, the overall retweets in peak 1 is far more than peak2, while entertainment events like #3 and #4 tend to have more retweets in the second peak. This interesting phenomenon indicates that people are more sensitive to events involving social security, and as for entertainment events they tend to have a delayed response.

3. Clustering

In order to figure out typical temporal patterns of hot events in Weibo, we implement the K-Spectral Centroid (K-SC) clustering algorithm to find the clusters.

Table 2: Statistics of the four patterns in Fig. 2. The number of followers is in log scale. $F_1 > 4$:The proportion of users with followers more than 10^4, and so on. P_1:The overall retweets in the first stage. P_2:The overall retweets in the second stage. t_F:The time point when user get the most retweets. t_P:The time point of the maximum peak.

	$F_1 > 4$	$F_2 > 4$	$F_1 > 5$	$F_2 > 5$	$F_1 > 6$	$F_2 > 6$	$F_1 > 7$	$F_2 > 7$	P_1	P_2	t_F	t_P
#1	22.9%	19.9%	4.02%	3.74%	0.71%	0.19%	0.36%	0	63.9%	28.9%	8	8
#2	6.27%	3.58%	1.47%	0.82%	0.35%	0.23%	0.07%	0	62.1%	34.3%	5	5
#3	17.0%	16.5%	2.01%	2.10%	0.16%	0.14%	0.04%	0	41.7%	52.7%	2	3
#4	10.1%	8.57%	1.56%	2.02%	0%	0.48%	0%	0	10.2%	82.8%	57	58

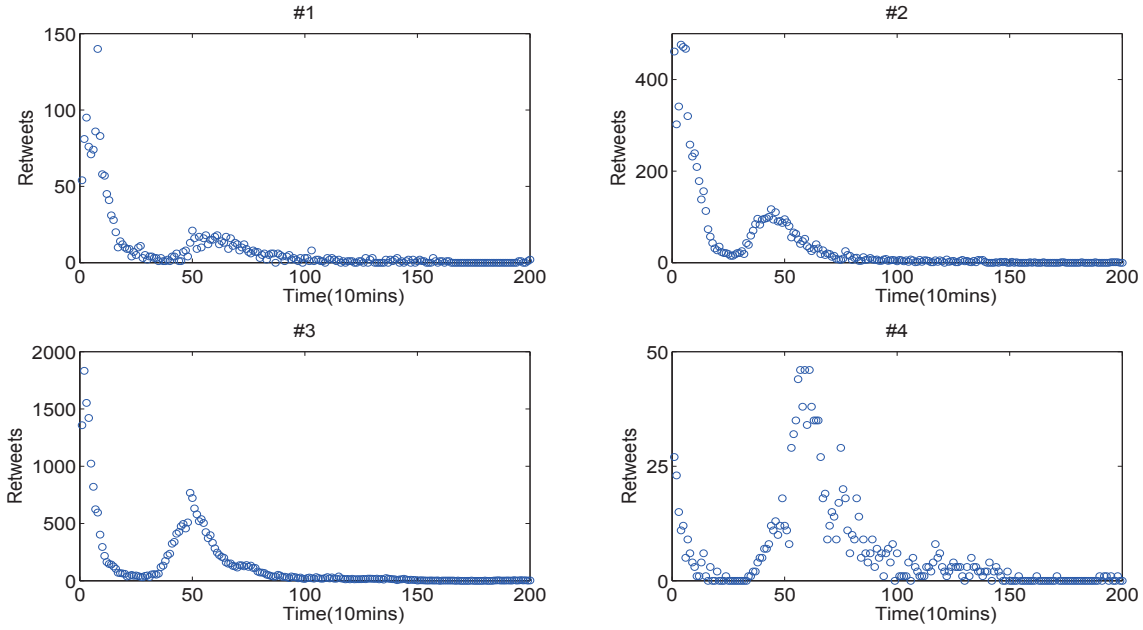

Fig. 2: PDF of the four events in Table 1.

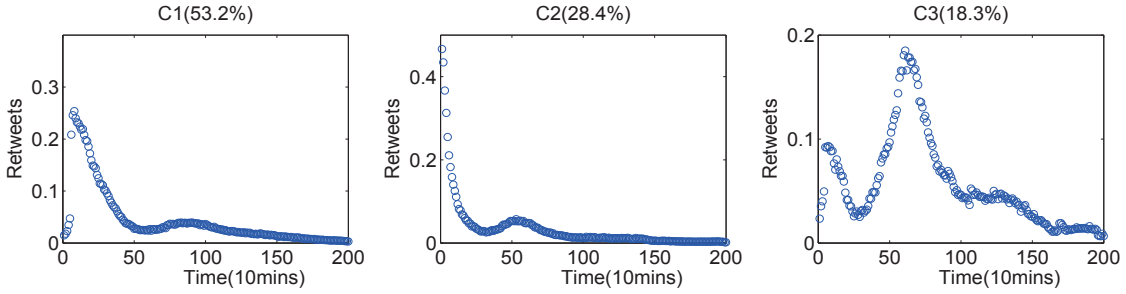

Fig. 3: Clustering results of K-SC. On the top are symbols of each cluster plus percentages of all the time series.

3.1 K-SC

K-SC is an algorithm similar to the classical K-means clustering algorithm, which is mainly comprised of similarity metric and calculation of clustering center. The basic idea of K-SC is iterating a two step procedure, the assignment and the refinement step. In the assignment step, every time series is assigned to the closest cluster by computing the distance between presenting time series and cluster center. In the refinement step, the cluster centroids are then updated. The similarity metric is only related to the shapes of time series by applying scaling and translation. Given two time series x and y, the similarity metric $d(x, y)$ is defined as follows:

$$d(x, y) = \min_{\alpha, q} \frac{\left| x - \alpha y_{(q)} \right|}{|x|} \tag{1}$$

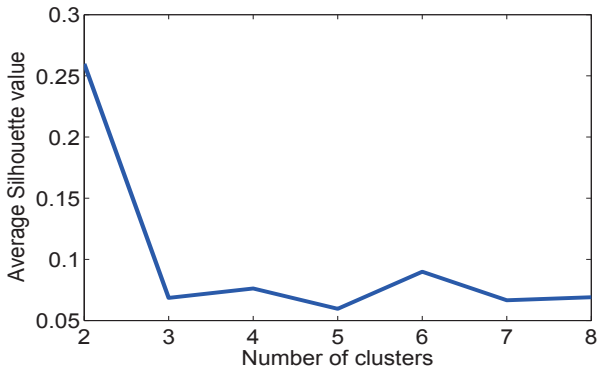

Fig. 4: Average Silhouette of different number of clusters.

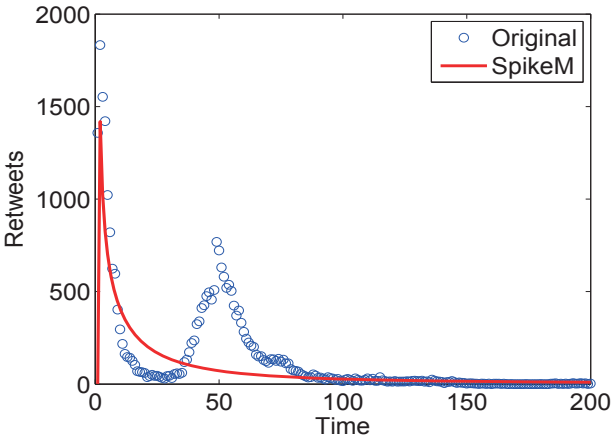

Fig. 5: SpikeM fitting result of #1 with *RMSE=189*.

where $y_{(q)}$ is the variation of time series y by shifting q time units and $|\cdot|$ is the l_2 norm. On the other hand, the new cluster center μ_k^* is updated by calculating virtual center of the cluster C_k, rather than simply averaging every member. It should be the minimizer of the sum of $d(x_i, \mu_k)^2$ over all $x_i \in C_k$:

$$\mu_k^* = \arg\min_{\mu} \sum_{x_i \in C_k} d(x_i, \mu_k)^2 \tag{2}$$

3.2 Experimental Results

As other variants of K-means algorithm, K-SC is also sensitive to the initially specified cluster centers. We use evaluation method Average Silhouette to determine the best number of cluster. Fig. 4 suggests that the Average Silhouette value keeps fluctuating across a fixed value when cluster count is bigger than 3. Empirically we find that the cluster centers are quite stable while setting the number of clusters from 3 to 8. Hence we choose 3 as the best number of clusters.

Fig. 3 shows the three cluster centers, which are represented by C_1 to C_3. As discussed in Section I, the patterns have no periodic trailing, which is totally different from the six patterns in tweeter. Note that the ordinate values are normalized by scaling. Occupying almost half(53.2%) of all the time series, C_1 is supposed to be the most common pattern. Its shape is also a compromised of the three cluster centers, confirming that C_1 is a very typical temporal pattern of hot events in Weibo. It has a brief rising period before reaching the peak, and then follows a pow law decay after peak point. The overall period around the second peak is very long, which means that it takes much long time to get the information widely adopted by the crowd. This matches the reality because in most cases the hot event is initially exposed to a small slice of users, then it is well adopted by the general public through the opinion leaders' significant influence which is corresponding to the rapidly rising period. Soon the propagation process experiences a descending period after the effect of opinion leader, stepping into the second stage. It rises and falls much more gently in the second stage since most users are grass roots with few followers.

C_2 is quite different from C_1 both in the first stage and in the second stage. It doesn't experience a rising period in the first stage. This significantly indicates that the information is directly published by opinion leader. When confronting celebrity gossips, the general public seems to be much more sensitive than usual. So the second peak reaches more quickly and greatly than C_1, and the second stage lasts a shorter period implying high-volatility property.

C_3 is entirely different from the above cluster centers. It represents rare circumstances of diffusion process, possessing only a proportion of 18.3%. According to C_3, the second stage plays a leading role. It has a much higher peak and a much longer lasting period than the first stage. In this case, the original source of information is generally grass roots. The information is initially spread in their small social network, soon it is adopted by opinion leaders due to the increasingly popularity, which in turn creates trend in the entire network.

4. Modeling the Shapes

4.1 SpikeM

SpikeM is a variation of 'Susceptible-Infected' (SI) model, which is the most basic epidemic model. On one hand, it assumes that the infectivity f of a node decays with pow-law distribution:

$$f(\tau) = \beta * \tau^{-1.5} \tag{3}$$

where τ corresponds to the time. Our observation is concordant with this assumption. In Fig. 4 we can see that every pattern has two power-law fall periods. On the other hand, it conditions that the total population of the social network is finite so as to avoid the divergence to infinity. The base

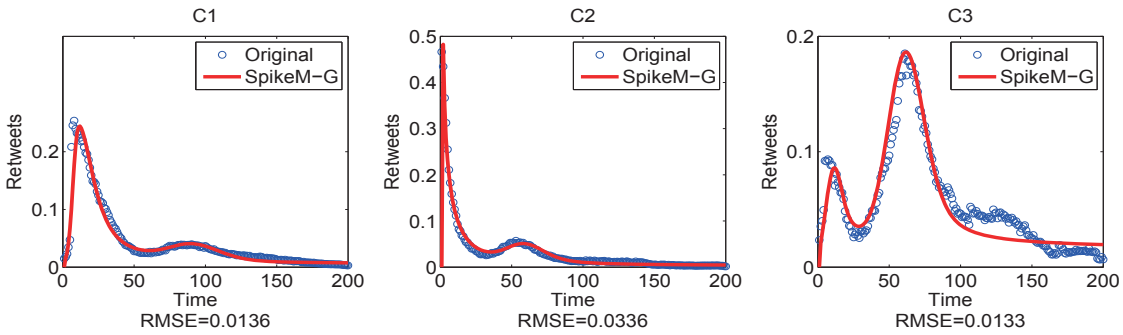

Fig. 6: Fitting results of SpikeM-G. On the bottom of each figure is the RMSE of fitting result.

model is defined by the following equations:

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \epsilon \quad (4)$$

$$\Delta U(n+1) = U(n) - \Delta B(n+1) \quad (5)$$

where $\Delta B(n)$ is the number of retweets at time n, $U(n)$ is the count of un-informed nodes, $S(n)$ is an external shock generated at birth-time n_b.

Although SpikeM model correctly captures the exponential rising period and the power-law decay period, it is only appropriate for the patterns with one spike. Because SpikeM assumes that the diffusion process is consist of only one stage. According to SpikeM, the number of un-informed nodes in the social network keeps dropping after the first peak, neglecting that the general public are unresponsive which will generate the second gently "shock" after several hours. Fig. 5 provides the fitting result of #1. Note that SpikeM model successfully captures the first rise and fall pattern, while the fitting result keep descending at the second period. We also evaluate the fitting accuracy by using the root mean square error (RMSE) metric between estimated values and real values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{1}^{n} (X_{model} - X_{real})^2} \quad (6)$$

As expected, the RMSE of SpikeM is 189, indicating a poor fitting.

4.2 Modified SpikeM

As for the above shortcomings of SpikeM, we propose a modified SpikeM model named SpikeM-G(short for SpikeM with Gaussian Function) based on the following assumptions:

The information propagation of hot events in Weibo is consist of two stages. The first stage generally experiences a short and rapidly spreading period, resulting in a much spiky pattern. Then after a peak-to-trough period, the "sleeping" nodes of the social network begin to "wake up", stepping into the second propagation stage. The accumulation of these nodes relatively generates another external shock. But in this stage the propagation process is much more gently and has a long-lasting period since the "waking up" time of each node is usually not the same.

Macroscopically speaking, there are two cases of the two-stage diffusion process of in Weibo. In the first case, which is more generally, information is propagated from the opinion leader to the crowd. The former plays an important role, while the latter act as audiences. The second case is just the opposite, where information is first published by the general public, and then it is spread to the whole network under the leadership of opinion leaders.

SpikeM only models the first stage of the diffusion process with external shock $S(n)$, so it needs another external shock at the second stage. We use gaussian function here since the rise and fall pattern of around the second peak is much gentle. Our modified model SpikeM-G is governed by the equations:

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t) + G(t)) \cdot f(n+1-t) + \epsilon \quad (7)$$

$$\Delta U(n+1) = U(n) - \Delta B(n+1) \quad (8)$$

and $G(t)$ is defined as:

$$G(t) = a \cdot e^{-w(t-t_p)} \quad (9)$$

where a is the volume of the second peak, t_p is time point of the second peak. The term $G(t)$ is very important. It models both the overall volume and the lasting period of the second stage. It also ensures the power-law decaying pattern, since it is multiplied by the infectivity function $f(\tau)$ outside the brackets.

Fig. 6 describes the results of SpikeM-G fitting on the three typical clustered temporal patterns. On the bottom of Fig. 6 displays the RMSE of each fitting result. In this figure, we can see that SpikeM-G is quite consistent with the previous two assumptions. Firstly, it successfully characterizes the two stages of information propagation process in Weibo where SpikeM model fails. On the other hand, whether information is propageted from opinion leader to

the crowd or the opposite, SpikeM-G is capable of correctly capturing the temporal pattern.

5. Conclusions

In this paper, we study the temporal patterns of hot events in three steps. Firstly, we analysis the statistics of all the temporal patterns, figuring out two basic fundamentals. On one hand, the information propagation of hot events in Weibo is comprised of two stages. On the other hand, we find out who actually contributes to the spike of each stage by analyzing the number of followers. Then the three typical temporal patterns of hot events are uncovered by implementing the KSC clustering algorithm. Finally, we introduce SpikeM-G which is based on SpikeM to get better fittings of the patterns. The experimental results show that our method performs well as for capturing the shape and achieving high accuracy.

This study helps to figure out who actually promotes information diffusion process in social network, which will contribute to the effectiveness of viral marketing. In order to produce increases in brand awareness, the viral campaign can be divided to opinion leader advertising stage and grass roots advertising stage. What's more, the study also provides a new access to public opinion monitoring since the spreading process is predictable.

References

[1] B.J.Jansen, M.Zhang, K.Sobel and A.Chowdury, *QMicro-blogging as Online Word of Mouth Branding*, CHI EA '09, pp. 3859–3864, Apr. 1977.

[2] T.Lou and J.Tang, *Mining Structural Hole Spanners Through Information Diffusion in Social Networks*, WWW'13, pp. 825–836, May. 2013.

[3] Z.Yin and W.Chen, *Discovering Patterns of Advertisement Propagation in Sina-Microblog*, ADKDD'12, Aug. 2012.

[4] J.Yang and J.Leskovec, *Patterns of Temporal Variation in Online Media*, WSDM'11, pp. 177–186, Feb. 2011.

[5] Y.Matsubara and Y.Sakurai and B.A.Prakash, *Rise and Fall Patterns of Information Diffusion: Model and Implications*, KDD'12, pp. 6–14, Aug. 2012.

[6] M.E.J.Newman, *The structure and function of complex networks*, KDD'13, pp. 6–14, Mar. 2013.

[7] Elihu Katz, *The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis*, Public Opinion Quarterly, vol. 21(1), pp. 61–78, 1957.