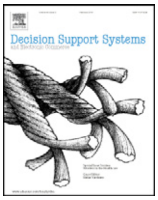
Decision Support Systems 81 (2016) 1–11

Contents lists available at ScienceDirect

Decision Support Systems

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



A model-free scheme for meme ranking in social media

Saike He [a], Xiaolong Zheng [a,*], Daniel Zeng [a,b]

[a] *The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA*

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form 9 July 2015
Accepted 3 October 2015
Available online 13 October 2015

Keywords:
Meme ranking
Model-free scheme
Transfer entropy

ABSTRACT

The prevalence of social media has greatly catalyzed the dissemination and proliferation of online memes (e.g., ideas, topics, melodies, and tags). However, this information abundance is exceeding the capability of online users to consume it. Ranking memes based on their popularities could promote online advertisement and content distribution. Despite such importance, few existing work can solve this problem well. They are either daunted by unpractical assumptions or incapability of characterizing dynamic information. As such, in this paper, we elaborate a model-free scheme to rank online memes in the context of social media. This scheme is capable to characterize the nonlinear interactions of online users, which mark the process of meme diffusion. Empirical studies on two large-scale, real-world datasets (one in English and one in Chinese) demonstrate the effectiveness and robustness of the proposed scheme. In addition, due to its fine-grained modeling of user dynamics, this ranking scheme can also be utilized to explain meme popularity through the lens of social influence.

1. Introduction

Meme (pronounced "meem") was first coined by Richard Dawkin in analogy with gene in genetics four decades ago [1]. It is defined as "unit of conceptual replication" that identifies idea, topic or style that spreads from person to person within a culture. Like the natural selection of genes that confer 'differential reproductivity', memes also compete for our scarce individual and collective attention [2]. During this process, some of them quickly die out of popularity while others persist for a long period of time. In recent years, the advent of various social media platforms has lowered the cost of information generation, boosting the potential reach of each meme among online users. This information abundance is exceeding human capacity to consume it [3–5]. Therefore, an effective ranking scheme is imperative to focus human limited attention on the most important memes. Appropriate solutions for this issue would provide direct implications in refining online advertisement and content distribution. In online advertising, new revenue models could be developed to charge advertisers for the amount of attention that a meme will receive. In media outlets, ranking information can be used to highlight the most popular memes in realtime. These condensed results are especially beneficial in emergent situations where information fragments emerge at random moments, such as social events [6,7], public health [8,9], and political campaigns [10–12]. In light of such importance, meme ranking has attracted considerable research interests in various disciplines [2,13,14].

However, to the best of our knowledge, few existing studies provide an adequate solution for meme ranking task. Traditional bottom-up approaches attempt to construct various diffusion models in analogy to behavior replication [15–18], epidemic contagion [14,19–24], or competitive gaming [25,26]. Although these models can help to track the diffusion process and measure its fact on online users, their computational complexities are often comparatively high, sometimes even NP-hard [27]. Furthermore, oversimplifications made in these models, such as user homogeneity [2,16], static network structure [19], and finite interaction patterns [28,29], can lead to unrealistic or even misleading conclusions. Recently, several top-down approaches have been developed to characterize meme dynamics, which is critical in accessing the evolution and mutation of online memes [4]. These studies mainly focus on quantifying the topological centrality [30,31], content similarity [32], or user behaviors [13] based on large-scale datasets. This line of research has provided significant insights in understanding the trend of the web. Yet, lacking fine-grained modeling of user interactions in meme diffusion, they still cannot characterize meme dynamics well.

To solve the above challenges, in this paper, we elaborate a model-free ranking scheme that characterizes meme dynamics with few assumptions. Different from previous work, our ranking scheme is designed based on information theory and could capture complex meme dynamics without modeling its exact diffusion process. In addition, while most existing studies are concerned with aggregate measures for meme ranking, the scheme presented here allows more fine-grained characterization on information diffusion among online users. This key property enables us not only to rank meme at the macro level, but also to inquiry key factors determining meme popularity at the micro level. For evaluation, we have used two different genres of

[*] Corresponding author.
E-mail addresses: saike.he@ia.ac.cn (S. He), xiaolong.zheng@ia.ac.cn (X. Zheng), dajun.zeng@ia.ac.cn (D. Zeng).

http://dx.doi.org/10.1016/j.dss.2015.10.002

datasets: one from a Chinese microblogging system and the other one from an American political blog forum. Experimental results on these two datasets validate the efficiency and robustness of the scheme compared with several benchmark approaches. By examining two key factors pertaining to meme spreaders, we also uncover several principles governing meme popularity. These findings may provide both academic and industrial implications in understanding other new types of memes such as innovation [15], rumor [19], and viral marketing [14,28].

The remaining parts of the paper are structured as follows. Section 2 reviews existing studies most relevant to our task. In Section 3, the technical details for the proposed meme ranking scheme are represented. Section 4 gives the empirical results of our proposed scheme in comparison with several existing approaches. Finally, Section 5 concludes this paper with a summary and a discussion about future research directions.

2. Literature review

The original work regarding meme traces back to a theory proposed by Dawkin [1], who first coined the concept of meme. This concept is utilized to describe the potential process of information diffusion among online users, in analogy with gene in genetics. In the following part of this section, we will present the existing studies relevant to our work from two perspectives, including meme diffusion and meme ranking.

2.1. Meme diffusion

Existing studies concerning meme diffusion mainly focus on constructing various theoretical models from different views. These models can help us to uncover the potential evolutionary patterns of meme diffusion to a certain extent. Generally, these meme diffusion models can be roughly categorized into three groups, i.e., cascade models, epidemic models, and competitive models respectively.

2.1.1. Cascade models

One of the famous cascade models is proposed by Bikhchandani et al. [16], who explore social changes by assuming all users hold the same belief in behavior making. This assumption clearly does not hold in real-world situations. Kempe et al. [15] then study online innovation diffusion and try to maximize its influence among users by selecting a subset of key nodes. In their cascade model, dynamics of neighbor pairs are considered independently. In fact, user dynamics is highly interwoven. Models for multiple cascades have been studied by extending the existing independent cascade model. These models generally assume that the status of each node keeps intact once influenced by other nodes [17]. Myers and Leskovec [18] further infer social relations based on information propagation in latent social networks. Both the cascades and infections are postulated to be conditionally independent in their propagation model. One common drawback of all these work is that assumptions made in modeling clearly do not hold in real-world practice. In contrast, our model makes no explicit assumptions about information dynamics.

2.1.2. Epidemic models

The epidemical analogy of information to virus has opened a new perspective for investigating meme diffusion and evolution. This, in turn, leads to pervasive applications of compartmental models such as SI, SIR, and SIS [20,21,33]. The spread of rumors and the detection for its source are studied with classical susceptible-infected (SI) model [19]. This model heavily depends on the network structure, which keeps developing and evolving. Some researchers study meme dynamics in the context of personal publishing. Gruhl et al. [23] employ snapshot models to depict topic propagation in blogspace. Their models are designed to characterize dynamics for both the communities and users. Article memes are studied by expressing complex human dynamics in analogy with infection by a virus [22]. These studies often assume the background environment as constant, which is not very practical in real world situations. In another strand of research, Richardson and Domingos [14] seek to optimize viral marketing plans by mining knowledge-sharing websites. In their probabilistic models, only one type of marketing action is considered. This simplicity may run counter to actual marketing scenarios.

2.1.3. Competitive models

To study meme competition among public attention, Weng et al. [2] employ a parsimonious agent-based model. However, their model highly relies on the underlying network structure and does not account for the discrepancy in user interest. Wei et al. [34] try to predict meme prevalence by considering network structural and information propagation at the same time. They assume that all nodes are passive and can be characterized with the same propagation model. Further, mixture of meme effect on individual is forbidden. Such postulation may not reflect the real situation in many circumstances. Goldenberg et al. [28] try to understand personal communications in word-of-mouth marketing. However, their complex system modeling technique could only cope with two types of predefined social interactions.

2.2. Meme ranking

Though there is comprehensive work investigating meme diffusion, to the best of our knowledge, the existing studies concerning meme ranking are comparatively limited. In what follows, we present a brief survey for this line of research.

Ienco et al. [35] and Bonchi et al. [27] initially attempt to construct propagation models to rank memes, but find that these models are pragmatically unfeasible since their computational complexities are NP-hard. Consequently, they turn to employ several heuristic methods. However, these methods cannot distinguish the direction of information flow, which is crucial in determining user importance in meme diffusion. Different from their work, in this paper, we adopt an asymmetric measure that is capable of capturing the direction of information flow among users. In another line of research, Bauckhage [13] ranks memes according to their average daily activity. Since activity level is measured via relative value, the ranking result may be confounded by other memes beyond consideration. Thus, it is highly possible that meme activity increases while its portion drops due to the proliferation of unknown memes.

There also exist other studies trying to rank meme based on topological centrality measures, such as in/out-degree, and number of followers. Gloor [30] measures trends on the web based on betweenness centrality. This measure requires a complete collection of underlying network structure, which is impossible in most scenarios. PageRank is a centrality algorithm that has been used widely in network analysis and ranking related tasks. Rather than prioritizing authoritative blogs, Adar et al. [31] try to rank blogs from the perspective of information diffusion. To this end, they propose an iRank algorithm to rank blogs based on implicit link structure. Their approach requires additional resource to train a link predictor, whose performance highly relies on the quality of this resource. However, such resource is not always available in real world practice, thus limiting its applications on a wide scale. Gordevicius et al. [32] focus on ranking news stories. Instead of using hypertext-links, they construct implicit links based on content similarity. Their algorithm is computational expensive, since it is equivalent to obtaining the stationary distribution of a random walk over a whole graph. Besides, the ranking result varies based on the similarity measure used. In contrast, except for user behaviors, our approach does not need any extra information. In addition, it computational complexity is also acceptable.

Recently, studies on meme ranking turn to explore dynamic information. One of the significant studies is presented by Kwak et al. [36], who attempt to rank trending topics based on singleton, reply, mention, and retweet information. Since such information is highly topic-dependent, a reliable scheme is thus imperative. This constitutes the main motivation for our work in this paper.

3. The meme ranking scheme

This section describes our proposed scheme for meme ranking. We first elaborate the rational for the designed scheme, and then introduce its formulation and computation in turn.

3.1. Scheme construction

Though meme ranking has attracted considerable attention and has been explored in different frameworks, existing ranking criteria are rather task dependent, and there still lack a theoretical guideline to measure meme popularity. Empirical analysis suggests that highly popular memes often associate with influential users spreading them [37]. We thus hope to design a scheme to rank meme based on the influence of users engaged in spreading it. Considering the variation in user volume of different memes, we propose to measure meme popularity based on the average influence of all users for each meme. This averaging manipulation is assumed to partially mitigate some confounding factors caused by external inference [38], unobserved heterogeneity [39], or some other contextual effects [40]. Then, the proposed ranking scheme can be formulated as:

$$Pop_m = \frac{1}{\#U_m}\sum\nolimits_{u \in U_m} Influence(u) \tag{1}$$

where, Pop_m quantifies the popularity of meme m; U_m represents the collection of users participating in spreading it, and the operator '#' measures the volume size of the set next to it; $Influence(u)$ corresponds to the influence of user u.

Given the ranking scheme defined by Eq. (1), we proceed to identify influence for each online user, as discussed next.

3.2. Influence identification

Though influence identification has been explored relatively thoroughly in social dynamics, formulating it in a model-free manner is not done before. To deal with the problems with the previous Mutual Information-based approaches, we adopt the recently developed transfer entropy [41] concept to guide the design of our scheme since this approach is asymmetric and can capture arbitrary nonlinear interactions well. While various kinds of information can be utilized to measure user influence, in this ranking scheme, we choose to use user behaviors during meme diffusion process. This guarantees that the identified user influence is most relevant to meme dynamics. In what follows, we present our influence identification approach, which is a model-free strategy.

3.2.1. Problem formulation

Suppose a pair of users x and y in online social media, whose behavior history can be approximated by two Markov processes $X = x_t$ and $Y = y_t$ (Fig. 1), we define an entropy rate h_1 [42] to measure the amount of additional information required to represent the next behavior x_{t+1} of user x given the historical information of the two users:

$$h_1 = -\sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log p\left(x_{t+1} \middle| \mathbf{x}_t^m, \mathbf{y}_t^n\right) \tag{2}$$

where, $\mathbf{x}_t^m = (x_t, \ldots, x_{t-m+1})$, $\mathbf{y}_t^n = (y_t, \ldots, y_{t-n+1})$; m and n are the orders (memory) of the Markov process X and Y respectively.

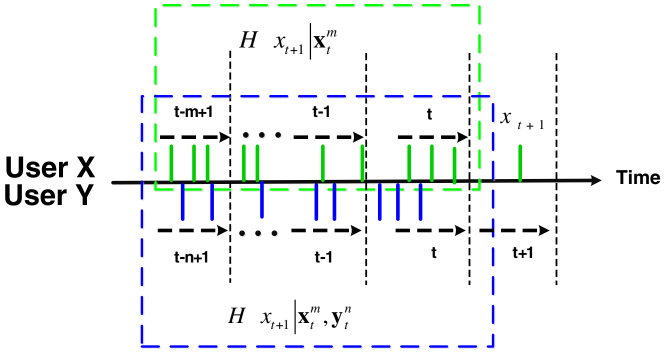

$H \quad x_{t+1}|\mathbf{x}_t^m$
t-m+1
t-1
t
x_{t+1}
User X
User Y
Time
t-n+1
t-1
t
t+1
$H \quad x_{t+1}|\mathbf{x}_t^m, \mathbf{y}_t^n$

Fig. 1. Illustration of transfer entropy. Solid vertical line corresponds to each user behavior launched at timestamp t (green for user x and blue for user y). $H(x_{t+1}|\mathbf{x}_t^m)$ amounts to the uncertainty about user x *(green rectangle)*, $H(x_{t+1}|\mathbf{x}_t^m, \mathbf{y}_t^n)$ amounts to the uncertainty about user x, if we know the behaviors of user y *(blue rectangle)*.

Suppose the status observation x_{t+1} is not dependent on the historical observations $\mathbf{y}_t^n$:

$$h_2 = -\sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log p\left(x_{t+1} \middle| \mathbf{x}_t^m\right). \tag{3}$$

Then, the departure of entropy rate defined by h_1 and h_2 is given by:

$$\begin{aligned} h_2 - h_1 &= -\sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log p\left(x_{t+1} \middle| \mathbf{x}_t^m\right) \\ &\quad + \sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log p\left(x_{t+1} \middle| \mathbf{x}_t^m \mathbf{y}_t^n\right) \\ &= \sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log \left(\frac{p\left(x_{t+1} \middle| \mathbf{x}_t^m, \mathbf{y}_t^n\right)}{p\left(x_{t+1} \middle| \mathbf{x}_t^m\right)} \right). \end{aligned} \tag{4}$$

With substitutions

$$p\left(x_{t+1} \middle| \mathbf{x}_t^m, \mathbf{y}_t^n\right) = p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) / p\left(\mathbf{x}_t^m, \mathbf{y}_t^n\right) \tag{5}$$

$$p\left(x_{t+1} \middle| \mathbf{x}_t^m\right) = p\left(x_{t+1}, \mathbf{x}_t^m\right) / p\left(\mathbf{x}_t^m\right). \tag{6}$$

We obtain:

$$h_2 - h_1 = \sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log \left(\frac{p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \cdot p\left(\mathbf{x}_t^m\right)}{p\left(\mathbf{x}_t^m, \mathbf{y}_t^n\right) \cdot p\left(x_{t+1}, \mathbf{x}_t^m\right)} \right). \tag{7}$$

Eq. (7) captures the transfer entropy from user y to user x, which can be further rewritten into a conditional mutual information:

$$\begin{aligned} TE(\mathrm{Y} \rightarrow \mathrm{X}) = h_2 - h_1 &= -\sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log \left(\frac{p\left(x_{t+1}, \mathbf{x}_t^m\right)}{p\left(\mathbf{x}_t^m\right)} \right) \\ &\quad + \sum_{x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n} p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right) \log \left(\frac{p\left(x_{t+1}, \mathbf{x}_t^m, \mathbf{y}_t^n\right)}{p\left(\mathbf{x}_t^m, \mathbf{y}_t^n\right)} \right) \\ &= H\left(x_{t+1} \middle| \mathbf{x}_t^m\right) - H\left(x_{t+1} \middle| \mathbf{x}_t^m, \mathbf{y}_t^n\right) \end{aligned} \tag{8}$$

where, $H(*)$ calculates entropy over a given distribution. For the sake of simplicity, we take $m = n$ from this point on.

Eq. (8) quantifies the amount of information that can be used to predict the behaviors of user x. This can be considered as a reflection of influence wielded by user y (Fig. 1). As Eq. (8) is defined in an asymmetric manner, it can thus be used to analysis user heterogeneity with regard to personal influence and investigate how it is related to meme popularity. This will be shown in the experiment section.

3.2.2. Influence estimation

We now turn to estimate transfer entropy defined in Eq. (8). Considering finite date samples available, we adopt a bin-based approach used by Kaiser and Schreiber [43]. In our formulation, the behavior history of user x can be recorded in a time series:

$$Sx = \{t_j:\ 0 < t_1 < t_2, \ldots\}. \tag{9}$$

To indicate whether user launches a behavior within a time span, a binned indicator function is introduced:

$$Bx(a,b) = \begin{cases} 1 & if \quad \exists t_j \in Sx \cap (b,a], \\ 0 & \text{otherwise}. \end{cases} \tag{10}$$

Over a long period of observation time interval [δ, T], we can define the probability function of user behavior as:

$$P(Bx(t,t-\delta) = x_t) \equiv \frac{1}{T-\delta}\int_{\delta}^{T} dt[Bx(t,t-\delta) = x_t] \tag{11}$$

where ‘[]’equals to 1 when logical repression enclosed is true and 0 otherwise.

Similarly, a joint probability distribution could be defined over a sequence of adjacent bins:

$$P(Bx(t,t-\delta_0) = x_t, Bx(t-\delta_0, t-\delta_0-\delta_1) = x_{t-1}, \ldots) \tag{12}$$

where, δ_0, δ_1, …, δ_k are bin widths. Without loss of generality, we omit the binned indicator function, then Eq. (12) changes to $P(x_t, x_{t-1}, \ldots, x_{t-k})$, with its most compact formation of $X_t^{(t-k)} \equiv \{x_t, x_{t-1}, \ldots, x_{t-k}\}$.

Then, for two users x and y with respective recorded of time series of Sx and Sy, the joint probability distribution over a common set of bins δ_0, δ_1, …, δ_k can be denoted as $P(\mathbf{x}_t^m, \mathbf{y}_t^n)$.

Given these notations, $TE(Y \rightarrow X)$ defined in Eq. (8) can be readily calculated. Because $TE(Y \rightarrow X)$ only depicts peer influence, the total influence $Influence(Y)$ of user y is obtained by summarizing the total influence he wields on all others in the community.

3.2.3. Bias remediation

Estimation based on limited data will lead to biases and statistical errors [44,45]. These errors mainly come from two sources: systematic deviation and statistical deviation. Systematic deviation is tackled with randomized experiments, as it will be discussed in Section 4. Here, statistical deviation can be eliminated mainly through two methods: ex-ante limitation and ex-post elimination. In the former, statistical deviation can be restrained to an arbitrary level with respect to the given dataset. Ex-post elimination, on the other hand, works by first estimating the bias itself, and then adjusting final result accordingly. Further, this method requires some general a priori knowledge (e.g., Panzeri–Treves bias estimate [46]), and works like a post hoc remedy. Thus, we consider ex-ante limitation more appropriate for our scenario of influence estimation.

In the proposed scheme, we use Simpson's rule [47] to estimate the influence defined in Eq. (8),which is formulated as:

$$\begin{aligned}\int_a^b TE'(t)dt &= \int_a^b \left[\frac{(t-c)(t-b)}{(a-c)(a-b)}TE'(a) + \frac{(t-a)(t-b)}{(c-a)(c-b)}TE'(c) + \frac{(t-a)(t-c)}{(b-a)(b-c)}TE'(b)\right]dt \\ &= \cdots = \frac{b-a}{6}\left[TE'(a) + 4TE'\left(\frac{a+b}{2}\right) + TE'(b)\right]\end{aligned} \tag{13}$$

where, $TE'(t)$ is the derivative of $TE(t)$.

Simpson's method approximates the target function via a “piecewise” quadratic. This means if a function is already quadratic, then the approximation will be exact. This property guarantees an unbiased estimation for user influence. In addition, Simpson's rule is computationally efficient since the computational complexity of estimating Eq. (8) is $O(N\log(N))$. This cost is acceptable for the meme ranking task.

4. Experimental results

In this section, we first introduce the datasets used for evaluation, and then describe the randomized trials used to alleviate systematic deviation. Finally, we present our experimental design and corresponding results.

4.1. Datasets

We evaluate the proposed ranking scheme on two different genres of dataset: Sina Weibo[1] and Daily Kos.[2]

4.1.1. Sina Weibo

Sina Weibo is a Twitter-like microblogging system in China. With more than 40 million active users spreading approximately 100 million messages each day,[3] this system is generally considered an ideal laboratory for investigating information contagion, especially for Chinese content. Of particular interest is the section of Sina Weibo named ‘Hot Topics’. In this section, trending topics are ranked according to their popularity among the public in China. Within each topic, a vast number of messages keep evolving and mutating as the topic flows through the network. In this scenario, a topic is an incarnation of meme, while the messages spreading along different threads are operational proxies to track its dynamics. In what follows, without ambiguity, we will use the term meme to refer to each topic in Sina Weibo.

To evaluate the meme ranking task, we crawled down all 10 memes in the ‘Hot Topics’ section. For each meme, we only collected the top 10 threads. As a huge number of messages are generated in each thread, there already manifests sufficient information about user behavior and its corresponding timestamp. Thus, this dataset (Weibo hereafter) is ideal for evaluating the meme ranking task. Statistical information for each meme in Weibo dataset is shown in Table 1. Table 1 suggests that each meme comprises at least 19,000 messages and users. This is a big enough dataset for our evaluation.

4.1.2. Daily Kos

Daily Kos is an American political blog that enable users to publish news and share opinions liberally. This site was founded in 2002, and soon became the premier online political discussion community with 2.5 million visitors per month.[4] On this blog platform, professional political writers post directly to the front page, while other regular users can post “diaries”. In responses to these front page entries and diaries, users write comments and make recommendations, thus driving topics spreading in the community. Ultimately, fiercely discussed diaries will be assigned a specific tag, and popular tags will be ranked in a ‘HOT TAGS’ section in the front page. In this blog community, we consider the hot tags as popular memes. Hereafter, meme will be used interchangeably with tag in Daily Kos.

To construct the evaluation dataset, we crawled down the top 10 hot tags with the 200 most recent diaries. For each diary, we also collected all the related comments and timestamps. Table 2 summarizes the statistics for Daily Kos dataset (Kos hereafter). Compared with Table 1, we find that users in Kos are more contributive by generating more messages (about 20 times higher). This trait may lead to different ranking results from those in Weibo, as will be investigated in following experiments.

[1] http://www.weibo.com.

[2] http://www.dailykos.com/.

[3] http://en.wikipedia.org/wiki/Sina_Weibo. Accessed on Apr. 1st, 2014.

[4] In 2009, Time magazine readers nominated this site as the second best blog platform.

Table 1
Statistical information for Weibo dataset.

Meme ID	Meme Title	#Messages	#Users
1	年少的爱情 Yong lovers	70,424	69,397
2	我是歌手半决赛 ‘I am a singer’ (Semifinals)	19,901	19,477
3	中国式过马路零容忍 Zero tolerance to Chinese style of crossing roads	88,179	86,200
4	文豪超能力 Literary giant endowed with super power	74,196	72,686
5	SCC 郭美美斗富 Meimei Guo fighting the rich in SCC	220,564	216,234
6	H7N9 禽流感 H7N9 avian influenza	33,092	29,808
7	养老金 Old-age pension	125,200	122,331
8	撒切尔夫人去世 Margaret Thatcher Dies	74,453	58,128
9	明星跳水真人秀 ‘Star in Danger’	36,033	33,723
10	博鳌亚洲论坛 Boao Forum for Asia	53,938	50,425
Total	–	795,980	758,409

Note: Data collected on April 25th, 2013. ‘#’denotes ‘the number of’.
Meme ID corresponds to its ranking position, and ‘–’ means unapplicable.

4.2. Randomized trials

To tackle the systematic deviation caused by multiple sources of bias, we design a randomized trial to minimize the potential negative effects. Our strategy is similar to He et al. [48], but comparatively practical and effective: we randomly sample user behaviors with corresponding timestamps for Weibo and Kos respectively.[5] To avoid data sparsity, users who have less than 2 behavior records have been be pruned out.

This procedure brings four main benefits. First, it alleviates the effects of selection bias, such as crawling strategy, and time point for data collection. Second, it guarantees that the sampled data are representative enough for the whole volume. Third, it statistically mitigates the effect of data incompleteness. Finally, it controls the inferences brought about by information leakage [49]. As users may be exposed to multiple memes at one time, randomized manipulation can offset such inferences with systematic expectation.

In the following experiments, we execute 10 independent randomized samplings for each approach, and the experimental results are averaged across all the runs. If not explicitly stated otherwise, we adopt a sampling rate of 5%.

4.3. Experimental design

In the following sections, we investigate meme ranking task by studying three major issues:

Issue 1: ranking performance. We explore how the proposed ranking scheme performs on two large-scale, real-world datasets.
Issue 2: ranking robustness. We examine the robustness of the ranking scheme by testing whether its performance is sensitive to different sampling rates.
Issue 3: popularity factors. We quantify two factors related to user heterogeneous, and dissect how they influence meme popularity.

Table 2
Statistical information for Kos dataset.

Meme ID	Meme title	#Messages	#Users
1	Recommended	372,225	9454
2	Affordable Care Act	180,798	10,855
3	Community	260,810	5506
4	HealthCare	147,686	10,557
5	Elections	148,456	10,851
6	Republicans	157,238	10,049
7	2014	64,318	7043
8	Environment	151,736	7519
9	Economy	148,888	8830
10	Barack Obama	228,088	10,039
Total	–	1,860,243	90,703

Note: Data collected on April 2nd, 2014. ‘#’denotes ‘the number of’.
Meme ID corresponds to its ranking position, and ‘–’ means unapplicable.

4.3.1. Parameter setup

To quantify user influence in the proposed ranking scheme, we employ repost behavior and comment behavior respectively in Weibo and Kos. In parameter settings, for all following experiments, we take $m = n = 3$ in Eq. (8) as the Markovian order for user behaviors. According to our empirical analysis, higher values do not give better performance, yet only improves computational cost. Considering the long tail phenomenon in online social behaviors [50–52], as well as the varying observation periods in different datasets, we divide time bins of user behavior elastically by selecting $\delta_0 = 0.01T$, $\delta_1 = 0.1T$, and $\delta_2 = 0.2T$ respectively,[6] where T is the total observation period in a dataset.

4.3.2. Evaluation method

For an objective evaluation on meme ranking results, we conduct experiments under two criteria: Edit Distance [53] based criterion and Kendall-tau Distance [54] based criterion. Gold standard is chosen as the ranking result of memes from the original website. In the following experiments, we clarify whether we can predict this result merely based on partial dataset available from the website.

4.3.2.1. Kendall tau Distance based criterion. Kendall tau Distance is a ranking metric defined as the number of pairwise disagreements between two rankings [55–57]. Due to its advantages in computability and interpretability, Kendall tau Distance has been used widely in information retrieval to evaluate ranking quality [55,58–60]. Given two lists L_1 and L_2, the Kendall tau Distance between them is:

$$K(\tau_1,\tau_2) = |\{(i,j): i<j, (\tau_1(i)<\tau_1(j) \wedge \tau_2(i)>\tau 2(j)) \vee (\tau_1(i)>\tau_1(j) \wedge \tau_2(i)<\tau 2(j))\}| \tag{14}$$

where, $\tau_1(i)$ and $\tau_2(i)$ are the ranking position of element i in L_1 and L_2. Originally, $K(\tau_1, \tau_2)$ equals to 0 if the two lists are identical and $n(n-1)/2$ (where n is the length of list) if one list is the reverse of the other. For the convenience of comparison, we normalize $K(\tau_1, \tau_2)$ and convert it to a similarity value:

$$\text{Kendall-tau\_Sim}(L_1, \mathrm{L}_2) = 1 - \frac{2K(\tau_1,\tau_2)}{n(n-1)}. \tag{15}$$

Kendall - tau_Sim(src, tar) has an interval [0, 1], where 1 indicates perfect matching of the two lists.

4.3.2.2. Edit Distance based criterion. Edit Distance is an alternative criterion used in evaluating ranking result. It is defined based on the Levenshtein distance, thus sensitive to item positioning in ranking. This trait enables close dissection on the difference between the gold standard and the ranking for evaluations.

[5] In randomized trials, we did not conduct randomized selection on the threads for each meme, for threads ranked higher tend to have more reposts and engaged users. Randomized selection on meme threads may undermine information abundance for tracking memes and measuring their popularity.

[6] In our datasets, this setting for bin widths fixes the finest temporal resolution for recent records of user behaviors and coarser for historical ones.

Also, we convert the original value given by the Edit Distance to a normalized similarity score:

$$\mathrm{Edit\_Sim}(L_1, L_2) = 1 - \frac{edit\_dist(L_1, L_2)}{\mathrm{max}\_length(L_1, L_2)} \tag{16}$$

where, 'L_1' and 'L_2' represent two lists to be measured, $edit\_dist()$ is a function calculating the Edit Distance, and $max\_length()$ denotes the maximum length of the two objects.

The similarity score $\mathrm{Edit\_Sim}(L_1, L_2)$ has a unit value interval of [0, 1], and higher value indicate better ranking result.

4.3.3. Benchmark approaches

To evaluate the comparative performance of the proposed ranking scheme, we also introduce four representative benchmark approaches in the literature, as discussed below.

4.3.3.1. Benchmark 1: followers-centered approach (Follower_Num). Previous studies have employed different topological characteristics of social networks to measure social influence, such as the in/out-degree [61] and PageRank [62]. Here, we use follower number to quantify user influence. Despite its simplicity, follower number is considered as a rational indicator of user influence since following links determine the flow of information [63] and more following links means more opportunities to influence others. More sophisticated algorithms might yield better results, and we will examine one of them later. Given these analytics, for each meme m, its overall popularity score $Pop\_FollowNr_m$ can be defined as:

$$Pop\_FollowNr_m = \frac{1}{\#U_m} \sum\nolimits_{u \in U_m} \#Follower(u) \tag{17}$$

where, U_m represents the collection of users engaged in spreading meme m, $Follower(u)$ is the collection of all followers of user u, and the operator '#' measures the volume size of the set next to it.

4.3.3.2. Benchmark 2: PageRank-based approach (PageRank). PageRank is an alternative metric for quantifying user influence. Apart from the number of links, PageRank also accounts for their qualities. Thus, PageRank is assumed to outperform follower number based approaches. However, PageRank requires explicit knowledge of the underlying network structure as a priori. Actually, an accurate characterization of the network structure is almost impossible as it changes and evolves continuously. In addition, network structure is usually meme independent. Thus, it seems inappropriate to rank online memes by utilizing PageRank directly.

To tackle this issue, we construct two social networks based on user behavior information, which is readily accessible and meme dependent. Taking each user as a node in the network, we consider there is an edge starting from user u to user v if u reposts (in Weibo) or comments (in Kos) a message of v. Statistics of the two constructed networks is shown in Tables 3 and 4 respectively. We notice that network density [64] in Kos dataset is higher (approximately 2 times as that in Weibo). This may be attributed to high level user contribution in meme diffusion.

Table 3
Network properties for memes in Weibo.

Meme ID	#Nodes	#Edges	Density (E−05)
1	32,725	32,514	6.072
2	4958	6776	55.141
3	46,217	108,254	10.236
4	34,750	94,713	15.687
5	117,393	148,111	2.149
6	11,076	14,351	23.398
7	55,996	80,235	5.117
8	29,388	41,001	9.495
9	16,239	29,975	22.735
10	24,159	41,161	14.105
Average	37,290.100	59,709.100	16.403

Table 4
Network properties for memes in Kos.

Meme ID	#Nodes	#Edges	Density (E−05)
1	6819	6681	28.740
2	7713	7480	25.150
3	4003	3901	48.701
4	7462	7222	25.943
5	7887	7645	24.583
6	6945	6741	27.955
7	4977	4780	38.602
8	5395	5256	36.122
9	6406	6228	30.358
10	7291	7117	26.780
Average	6489.800	6305.100	31.293

Based on the constructed social networks, we then execute the PageRank algorithm. The popularity score of meme m is calculated as the average PageRank value of each user:

$$Pop\_PageRank_m = \frac{1}{\#U_m} \sum\nolimits_{u \in U_m} PageRank(u) \tag{18}$$

where, U_m represents the collection of users engaged in the diffusion process of meme m, $PageRank(u)$ is the corresponding PageRank value of user u, and the operator '#' measures the volume size of the set next to it. Again, a meme's popularity score is moderated by an averaging procedure.

4.3.3.3. Benchmark 3: dynamic information-based approach (Dynamic). Recent studies suggest that static structural measures alone reveal very little about social influence [65,66], while more accurate quantification requires characterizing on dynamic processes [67]. Thus, we design to measure user influence by utilizing the dynamic user behavior information. According to Romero et al. [66], the action rates vary widely across users, and a relatively small portion of them play a key role in meme diffusion. This finding prioritizes the necessity of quantifying user activity level for more accurate influence identification. As such, we use the number of repost (in Weibo) and comment (in Kos) behaviors to measure user activity level.[7] The popularity score of meme m is then formulated as:

$$Pop\_BehaviorNr_m = \frac{1}{\#U_m} \sum\nolimits_{u \in U_m} \#Behavior(u) \tag{19}$$

where, U_m represents the collection of users engaged in the diffusion of meme m, $Behavior(u)$ is the set of a given type of behavior of user u, and the operator '#' measures the volume size of the set next to it.

Apart from user activity, some researchers also employ user passivity [66] (or susceptibility [68] from the opposite point of view) to measure the resistance in influencing others. Aral and Walker [68] suggest that influential users, usually active in information diffusion, are less susceptible to influence. However, user passivity should not be considered as a directly opposite perspective to user activity, as they are totally different metrics for depicting the same user. Further research is needed to determine whether influence based on user passivity will rank memes differently from that based on user activity. This issue is beyond the current research scope and will be considered in our future work.

4.3.3.4. Benchmark 4: diffusion-based approach (Diffusion). Following Bonchi et al. [27] and Goyal et al. [69], we also quantify influence

[7] Within each meme, it is possible for a user to repost or comment multiple times as there are many different variations of one original message.

Table 5
Traits of the benchmark approaches.

Benchmark approach	Information type	Advantages	Disadvantage
Follower_Num	Static information (follower number)	Computational simplicity	Not very accurate for dynamic data
PageRank	Static information (network structure)	Accounting for the number of links and their qualities	Requiring explicit knowledge of the underlying network
Dynamic	Dynamic information (user behaviors)	Revealing social dynamics	Incapable of characterizing social interactions among users
Diffusion	Dynamic information (diffusion potential)	Requiring no explicit causal knowledge about user interactions	Heuristic and cannot capture nonlinear relationships

Influ(u, v) exerted by user u on user v based on the probability that each post generated by u will be further consumed by v.

$$Influ(u,v)=\frac{|\{p\in m|\ \exists t\in T: c(v,u,p,t)\}|}{|\{p\in m|\ \exists t\in T: g(u,p,t)\}|} \tag{20}$$

where, $g(u, p, t)$ represents that user u generates a message p at timestamp t, while $c(v, u, p, t)$ indicates that user v further consumes (repost in Weibo and comment in Kos) message p *generated by* user u at timestamp t.

Then, the popularity score of meme m is given by:

$$Pop\_Influ_m=\frac{1}{\#U_m}\sum\nolimits_{u\in U_m} Influ(u) \tag{21}$$

where, U_m represents the collection of users engaged in the diffusion of meme m, *Influ*(u) is the total influence of user u *wielded on others*, and the operator '#' measures the volume size of the set next to it.

Traits of the above benchmark approaches are summarized in Table 5.

4.4. Results

4.4.1. Issue 1: ranking performance

In designing the meme ranking task, we attempt to evaluate whether the proposed ranking scheme can predict meme popularity (as indicated by the original website) merely based on partial data available. This task is meaningful as it informs the possibility of either ranking memes with limited data samples at a specific timestamp, or predicting future meme popularity based on historical data samples. In this paper, we mainly focus on the former situation and experimental results are shown in Tables 6 and 7.

Primarily, we notice two contradictions to our empirical expectation. First, PageRank group fails to outperform Follower_Num group. Since it is structure dependent, we contemplate this result may be attributed to the reconstructed social network, which is highly biased due to limited data samples. The superiority of Follower_Num over PageRank implies that structural information is more reliable in meme ranking with limited data samples.

Second, Dynamic group fails to perform better than the Follower_Num group in both datasets. This result is inconsistent with previous work that assumes dynamic information is more reliable than structural information in depicting information dynamics [63]. In turn, it indicates that meme popularity, at least in the current situation, is reflected more by the long-term status of a user's social network geometry [70], while less by their short-term communication relationship. This superficially uncanny fact can be understood by delving into the underlying interaction network. An individual's social network mainly constitutes of two parts, i.e., explicit following relationships and implicit communication avenues [71]. Explicit social relation reflects one's long-term status in a community, and implicit communication avenues reflects his short-term prestige, which is highly event dependent (e.g., spreading a specific meme). In the situation where data samples are incomplete, dynamic information turns to be unreliable while explicit social networks become more predictable for social influence, and ultimately for meme popularity.

Apart from the above contradictions, we also notice that the proposed scheme outperforms the four benchmark approaches under both evaluation criteria. Specially, we notice that most of the benchmark approaches do not consider the meme "Yong lovers" (ID = 1) as the most popular meme in Weibo dataset (Table 1). This meme corresponds to an untimely death of a lovelorn teenager. Though the absolute number of messages generated in diffusing this meme is not the highest, it does exert intense influence among online users, where fierce discussions develop. While the benchmark approaches merely rely on network structure (constructed either statically or dynamically) for ranking, our proposed scheme quantifies influence among users in a model-free manner with high accuracy. This capability enables our scheme to captures complex nonlinear social interactions through both explicit social networks and implicit communication avenues.

Table 6
Performance of different ranking approaches on Weibo.

Meme ID	Follower_Num	PageRank (E − 6)	Dynamic	Diffusion	MF
1	100.325	30.492	2.128	0.393	3.701
2	154.308	201.803	2.814	0.256	3.002
3	102.148	216.381	2.140	0.136	2.736
4	122.958	28.868	2.146	0.129	2.500
5	110.127	8.433	2.210	0.263	1.956
6	139.055	90.148	2.119	0.244	1.851
7	120.585	17.688	2.263	0.168	1.116
8	77.262	33.883	4.242	0.193	3.395
9	122.821	61.771	2.352	0.099	2.813
10	85.099	41.573	2.638	0.169	2.473
Kendall-tau_Sim	0.578	0.533	0.289	0.689	0.689
Edit_Sim	0.200	0.100	0.100	0.200	0.400

Note: Meme ID is consistent with meme position in gold standard; 'MF' indicates the experiment group of our model free ranking scheme. 'Kendall-tau_Sim' and 'Edit_Sim' correspond to evaluate criteria based on Kendall tau Distance and Edit Distance respectively.

Table 7
Performance of different ranking approaches on Kos.

Meme ID	Follower_Num	PageRank (E − 6)	Dynamic	Diffusion	MF
1	31.543	146.904	39.372	0.080	148.582
2	30.378	129.571	16.655	0.067	12.622
3	48.671	249.691	47.368	0.068	23.089
4	34.453	133.842	13.989	0.073	20.162
5	29.653	126.569	13.681	0.079	5.558
6	32.749	144.133	15.647	0.068	10.449
7	42.962	200.624	9.132	0.125	12.495
8	41.082	185.536	20.18	0.045	6.324
9	37.226	156.337	16.861	0.071	8.129
10	33.785	137.092	22.72	0.052	2.186
Kendall-tau_Sim	0.578	0.467	0.533	0.578	0.822
Edit_Sim	0.1	0.1	0.1	0.2	0.5

Note: Meme ID is consistent with meme position in gold standard; 'MF' indicates the experiment group of our model free ranking scheme. 'Kendall-tau_Sim' and 'Edit_Sim' correspond to evaluate criteria based on Kendall tau Distance and Edit Distance respectively.

Table 8
Meme ranking performance with different sampling rates for Weibo.

Meme ID	SR = 1%	SR = 5%	SR = 10%	SR = 15%	SR = 20%
1	3.613	3.701	2.943	3.786	2.654
2	3.310	3.002	2.796	3.38	2.466
3	2.517	2.736	2.813	3.224	2.509
4	2.466	2.500	2.515	2.961	2.218
5	1.751	1.956	2.511	2.623	2.267
6	1.316	1.851	2.192	2.630	2.192
7	1.483	1.116	2.105	3.537	2.186
8	4.191	3.395	2.733	3.504	2.609
9	3.180	2.813	2.671	2.297	2.011
10	2.397	2.473	2.578	2.239	2.384
Kendall-tau_Sim	0.644	0.689	0.711	0.756	0.711
Edit_Sim	0.3	0.4	0.3	0.5	0.4

Note: Meme ID is consistent with meme position in gold standard; The popularity score of each meme is calculated according to (1); 'SR' represents 'sampling rate'; 'Kendall-tau_Sim' and 'Edit_Sim' correspond to evaluate criteria based on Kendall tau Distance and Edit Distance respectively.

Similar situation occurs in the meme "Margaret Thatcher Dies" (ID = 8) in Weibo dataset (Table 1). Though this meme receives a low ranking position from both the official website and the four benchmark approaches, it is found to has been gaining momentum in the following several days when other memes already lose popularity. Our ranking scheme 'predicts' this meme's future popularity, while other approaches only rank meme temporarily without any forecast. Similar findings are also found in Kos dataset, thus further validating our conclusions.

Finally, there is one point worth mentioning for the evaluation criteria. Though Kendall-tau_Sim generally gives higher scores ($p < 0.01$ according to a two-tailed paired t-test), Edit_Sim is more sensitive to meme positioning in ranking. This fact is indicated by the higher relative error between the maximal and the minimal ranking scores given by Edit_Sim (about 2 times higher than that given by Kendall-tau_Sim). This disparity enables us to investigate meme ranking task from different perspectives. The consistency in ranking results under these two criteria strengthens the soundness of previous conclusions derived.

4.4.2. Issue 2: ranking robustness

As sampling manipulation has been adopted in randomized trials, we next explore whether our scheme is robust with respect to the sampling rate. By comparing ranking performance at different sampling rates, we can validate the possibility of predicting meme popularity with only partial data available. In this subsection, we run several simulations with distinct sampling rates, and record the corresponding performance (Tables 8 and 9). The maximum sample rate in this experiment is set to 20%. We believe this is the upper allowable bound for sampling as larger values may exacerbate the representativeness of the sampled dataset.

Table 9
Meme ranking performance with different sampling rates for Kos.

Meme ID	SR = 1%	SR = 5%	SR = 10%	SR = 15%	SR = 20%
1	117.813	148.582	91.534	106.756	79.638
2	14.754	12.622	24.231	18.521	21.879
3	26.230	23.089	17.264	9.524	34.477
4	5.682	20.162	15.876	13.207	18.934
5	14.948	5.558	19.906	6.733	21.171
6	5.796	10.449	13.147	12.738	9.222
7	3.156	12.495	7.850	15.616	13.180
8	12.131	6.324	9.387	5.621	16.815
9	6.742	8.129	4.295	7.529	3.631
10	3.667	2.186	8.432	5.440	3.130
Kendall-tau_Sim	0.755	0.822	0.888	0.8	0.888
Edit_Sim	0.3	0.5	0.4	0.4	0.5

Note: Meme ID is consistent with meme position in gold standard; The popularity score of each meme is calculated according to (1); 'SR' represents 'sampling rate'; 'Kendall-tau_Sim' and 'Edit_Sim' correspond to evaluate criteria based on Kendall tau Distance and Edit Distance respectively.

Table 10
Statistical results for Weibo.

Meme ID	MF	α_m (%)	β_m (%)
1	3.701	27.184	1.942
2	3.002	27.692	6.154
3	2.736	4.108	2.528
4	2.500	44.204	2.180
5	1.956	9.614	6.801
6	1.851	2.666	8.984
7	1.116	17.038	7.268
8	3.395	45.127	0.479
9	2.813	54.946	0.860
10	2.473	42.804	0.747

Note: the second column represents results given by our scheme (sampling rate = 5%).

Results from Tables 8 and 9 suggest that the ranking results are insensitive to different sampling rates. In Weibo dataset, the average scores of Kendall-tau_Sim and Edit_Sim are 0.702 (±0.041) and 0.380 (±0.084) respectively. In Kos dataset, the standard deviation is only slightly larger, with average Edit_Sim 0.420 (±0.083) and Kendall-tau_Sim 0.830 (±0.057). These results prove the robustness of our ranking scheme by insensitive to changes in the underlying network structure. This trait distinguishes our scheme from benchmark approaches that are structure-dependent. Also, the results validate the possibility of ranking memes reliably with only finite data samples.

4.4.3. Issue 3: popularity factors

Popular memes are assumed to possess certain competitive advantages. In this subsection, we attempt to explore two factors contributing to meme popularity.

Bakshyet et al. [37] allege that information diffusion is driven by influential users. Thus, we first study how such users contribute to a meme's popularity. Specifically, we are interested in the portion of users (α_m) whose influence as a whole exceeds a given threshold θ of the total influence, which is defined as:

$$\alpha_m = \frac{N \left| \min_N \left(\left(\sum_{i=1, Influ(u_i) \geq Influ(u_{i+1})}^{N} Influ(u_i) \right) \geq \theta \sum_{u \in U_m} Influ(u) \right) \right.}{\# U_m} \tag{22}$$

where, U_m represents the collection of users engaged in the diffusion of meme m, $Influ(u)$ is the total influence of user u *wielded on others*, and the operator '#' measures the volume size of the set next to it.

In the results presented here, we use a value of $\theta = 70\%$. We have also experimented with other values of θ and observed similar qualitative results.

Table 11
Statistical results for Kos.

Meme ID	MF	α_m (%)	β_m (%)
1	148.582	47.761	0.021
2	12.622	40.247	0.087
3	23.089	40.702	0.409
4	20.162	41.953	0.373
5	5.558	32.919	0.149
6	10.449	42.699	0.262
7	12.495	42.447	0.848
8	6.324	40.368	0.549
9	8.129	44.726	6.04
10	2.186	40.491	2.578

Note: the second column represents results given by our scheme (sampling rate = 5%).

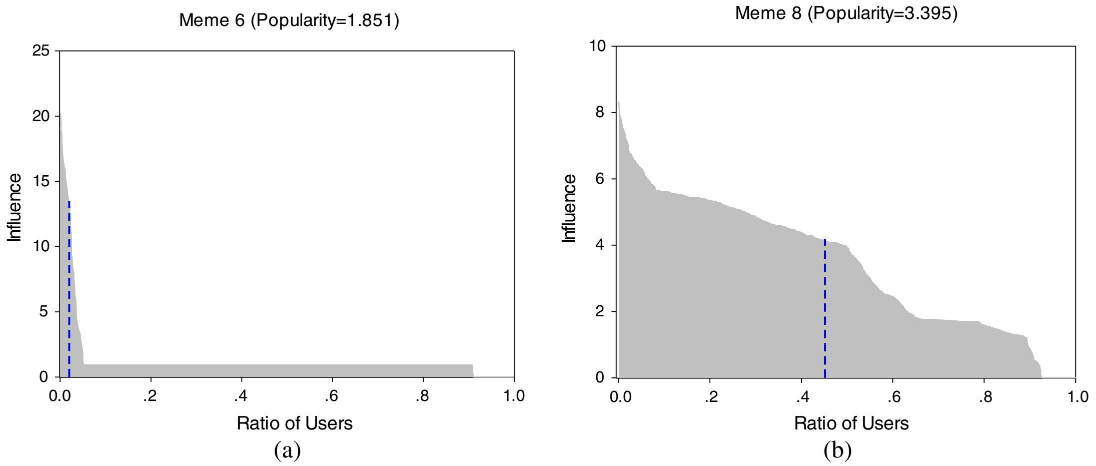


Fig. 2. Influence distribution of users over different memes in Weibo. Note: User influence is listed in descending order, and the left side of the blue line corresponds to 70% of the total influence.

Besides, we notice that there are users whose total influence is zero. Then, we intend to examine how the portion of such users (β_m) affects meme popularity:

$$\beta_m = \frac{N \left| \max_N \left(\left(\sum_{i=1, Influ(u_i) \le Influ(u_{i+1})}^{N} Influ(u_i) \right) \le 0 \right) \right.}{\#U_m} \tag{23}$$

where, U_m represents the collection of users engaged in diffusing of meme m, $Influ(u)$ is the total influence of user u *wielded on others*, and the operator '#' measures the volume size of the set next to it.

In following experiments, we analyze the respective correlation of α_m and β_m between meme popularity. Experimental results are shown in Tables 10 and 11.

It turns out that memes with higher popularity scores tend to associate with larger α_m. Pearson's Correlation Coefficient between them is 0.453 in Weibo and 0.609 in Kos.[8] This means the popularity of a meme is achieved by involving a wide scale of users with moderate influence. Put another way, just a small amount of high influential users is insufficient to guarantee a meme's prevalence. Taking meme 6 and meme 8 in Weibo dataset as an example,[9] meme 6 (Fig. 2-a) comprises a portion of high influential users, with the highest influence exceeding 20. However, this portion is relatively small ($\alpha_m = 2.666\%$), and there is a sharp drop in user influence for the remainder of the distribution. On the other hand, the influence distribution of meme 8 (Fig. 2-b) is much smoother ($\alpha_m = 45.127\%$), with few users possessing prohibitively high influence. This difference makes meme 8 more popular among the public. Thus, users are more likely to be exposed to and influenced by it, and are more willing to spread it. This finding (finding 1) coincides with previous postulation that information diffusion can also be realized by moderate or less influential users; indeed, sometimes they even do a better job [37].

By comparing columns 2 and 4 in Table 10, we observe that popular memes are likely to contain fewer zero-influence users (finding 2). Pearson's Correlation Coefficient between them is –0.687, which suggests a relatively strong negative relationship. This phenomenon can be explained partly by the concept of re-diffusion intention [72] in marketing. Under its theoretical framework, if a meme is really popular, then the re-diffusion intention of users is high, and other users are more likely to be exposed to this meme. Hence, the probability that users release no influence by spreading the meme is low.

Likewise, we find a similar phenomenon in Table 11, but Pearson's Correlation Coefficient is comparatively low, only −0.254. As memes in Kos generally have much higher influence level than that in Weibo, the existence of relative small portion of zero-influence users might not affect meme popularity much. Thus, the correlation between meme popularity and ratio of zero-influence users is low.

The above findings are obtained because of our scheme's fine-grained modeling of user dynamics. These results may not be found if we only utilize the existing approaches. As the scheme makes no domain-specific assumptions, our work can be readily generalized to analyze other new types of memes, such as innovation [15], rumor [19], and viral marketing [14,28].

5. Conclusions and future work

In this paper, we proposed a novel model-free scheme for meme ranking. Empirical studies on two large-scale real world datasets validate its efficiency and robustness. This scheme can provide us significant insights into understanding the meme popularity due to its fine-grained modeling of user dynamics. By analyzing two key factors regarding the user influence, we uncover two significant findings: (1) the meme popularity is achieved by a wide scale of users on its diffusion trace, while just a small amount of high influential users is insufficient; (2) more popular memes are prone to contain less zero-influence users.

In our future work, we intend to investigate whether other types of dynamic information also contribute to meme ranking task, such as user passivity and adoption rate. As only user behaviors are considered in quantifying user dynamics, we wonder whether the patterns of user behaviors bear some relationship with meme ranking results. Further, we will investigate whether involving contextual information and the behavioral and cognitive factors of online users would predict the meme popularity more accurately.

Acknowledgement

We would like to thank each member of the SMILES group in the Institute of Automation, Chinese Academy of Sciences, especially Kainan Cui, Zhu Zhang, and Chuan Luo for their useful discussions. We also appreciate the kind help from Yufang Wu and Kainan Cui for the data collection. This work was supported in part by the following grants: the National Natural Science Foundation of China under Grant Nos. 71402177, 71472175, and 71103180; the National Institutes of Health

[8] Considering the limited data samples we used, these correlation values are relatively high.

[9] We also experimented with Kos data, and observed similar results.

(NIH) of USA under Grant No. 1R01DA037378-01; and the Ministry of Health under Grant No. 2013ZX10004218.

References

[1] R. Dawkin, The Selfish Gene, Oxford University Press, New York City, 1976.
[2] L. Weng, A. Flammini, A. Vespignani, F. Menczer, Competition among memes in a world with limited attention, Sci. Rep. 2 (2012) (03/29/online).
[3] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. 105 (41) (2008) 15649–15653.
[4] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France 2009, pp. 497–506.
[5] K. Lerman, R. Ghosh, Information contagion: an empirical study of the spread of news on Digg and Twitter social networks, ICWSM 10 (2010) 90–97.
[6] M. De Choudhury, A. Monroy-Hernandez, G. Mark, Narco emotions: affect and desensitization in social media during the mexican drug war, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2014, pp. 3563–3572.
[7] Y. Qu, C. Huang, P. Zhang, J. Zhang, Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake, Proceedings of the ACM 2011 conference on Computer supported cooperative work 2011, pp. 25–34.
[8] M. Salathé, D.Q. Vu, S. Khandelwal, D.R. Hunter, The dynamics of health behavior sentiments on a large online social network, EPJ Data Sci. 2 (1) (2013) 1–12.
[9] Y. Liang, X. Zhou, D.D. Zeng, B. Guo, X. Zheng, Z. Yu, An Integrated Approach of Sensing Tobacco-oriented Activities in Online Participatory Media, 2014.
[10] M. Choy, M. Cheong, M.N. Laik, K.P. Shung, US presidential election 2012 prediction using census corrected twitter model, arXiv Preprint arXiv:1211.09382012.
[11] J. Pasek, K. Kenski, D. Romer, K.H. Jamieson, America's youth and community engagement. How use of mass media is related to civic activity and political awareness in 14-to 22-year-olds, Commun. Res. 33 (3) (2006) 115–135.
[12] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with Twitter: what 140 characters reveal about political sentiment, ICWSM 10 (2010) 178–185.
[13] C. Bauckhage, Insights Into Internet Memes, ICWSM, 2011.
[14] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002, pp. 61–70.
[15] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2003, pp. 137–146.
[16] S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades, J. Polit. Econ. (1992) 992–1026.
[17] C. Budak, D. Agrawal, A.E. Abbadi, Limiting the spread of misinformation in social networks, Proceedings of the 20th International Conference on World wide Web, Hyderabad, India 2011, pp. 665–674.
[18] S.A. Myers, J. Leskovec, On the convexity of latent social network inference, arXiv preprint arXiv:1010.55042010.
[19] D. Shah, T. Zaman, Rumors in a network: who's the culprit? IEEE Trans. Inf. Theory 57 (8) (2011) 5163–5181.
[20] N.T. Bailey, The Mathematical Theory of Infectious Diseases and Its Applications, Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE, 1975.
[21] R.M. Anderson, R.M. May, B. Anderson, Infectious Diseases of Humans: Dynamics and Control, Wiley Online Library, 1992.
[22] M. Kubo, K. Naruse, H. Sato, T. Matubara, The possibility of an epidemic meme analogy for web community population analysis, Intelligent Data Engineering and Automated Learning—IDEAL 2007. , Springer, 2007 1073–1080.
[23] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, Proceedings of the 13th International Conference on World Wide Web 2004, pp. 491–501.
[24] Y.-M. Li, Y.-L. Shiu, A diffusion mechanism for social advertising over microblogs, Decis. Support. Syst. 54 (1) (2012) 9–22.
[25] L. Shifman, M. Thelwall, Assessing global diffusion with Web memetics: the spread and evolution of a popular joke, J. Am. Soc. Inf. Sci. Technol. 60 (12) (2009) 2567–2576.
[26] Q.H. Nguyen, Y.S. Ong, M.H. Lim, Non-genetic transmission of memes by diffusion, Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation 2008, pp. 1017–1024.
[27] F. Bonchi, C. Castillo, D. Ienco, Meme ranking to maximize posts virality in microblogging platforms, J. Intell. Inf. Syst. 40 (2) (2013) 211–239 (2013/04/01).
[28] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (3) (2001) 211–223.
[29] C.M. Cheung, M.K. Lee, What drives consumers to spread electronic word of mouth in online consumer-opinion platforms, Decis. Support Syst. 53 (1) (2012) 218–225.
[30] P.A. Gloor, Coolhunting for trends on the web, Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on 2007, pp. 1–8.
[31] E. Adar, L. Zhang, L.A. Adamic, R.M. Lukose, Implicit structure and the dynamics of blogspace, Workshop on the Weblogging Ecosystem, 2004.
[32] J. Gordevicius, F.J. Estrada, H.C. Lee, P. Andritsos, J. Gamper, Ranking of evolving stories through meta-aggregation, Proceedings of the 19th ACM International Conference on Information and Knowledge Management 2010, pp. 1909–1912.
[33] L.J. Allen, Some discrete-time SI, SIR, and SIS epidemic models, Math. Biosci. 124 (1) (1994) 83–105.
[34] W. Xuetao, N.C. Valler, B.A. Prakash, I. Neamtiu, M. Faloutsos, C. Faloutsos, Competing memes propagation on networks: a network science perspective, IEEE J. Sel. Areas Commun. 31 (6) (2013) 1049–1060.
[35] D. Ienco, F. Bonchi, C. Castillo, The meme ranking problem: maximizing microblogging virality, 2010 IEEE International Conference on Data Mining Workshops 2010, pp. 328–335.
[36] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web 2010, pp. 591–600.
[37] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on twitter, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining 2011, pp. 65–74.
[38] S.A. Myers, C. Zhu, J. Leskovec, Information diffusion and external influence in networks, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012, pp. 33–41.
[39] C. Van den Bulte, G.L. Lilien, Medical innovation revisited: social contagion versus marketing effort, Am. J. Sociol. 106 (5) (2001) 1409–1435.
[40] C.F. Manski, Identification of endogenous social effects: the reflection problem, Rev. Econ. Stud. 60 (3) (1993) 531–542.
[41] Y. Saito, H. Harashima, Tracking of information within multichannel {EEG} record causal analysis in {EEG}, in: N. Yamaguchi, K. Fujisawa (Eds.), Recent Advances in {EEG} and {EMG} Data Processing, Elsevier 1981, pp. 133–146.
[42] V. Latora, M. Baranger, Kolmogorov–Sinai entropy rate versus physical entropy, Phys. Rev. Lett. 82 (3) (1999) 520.
[43] A. Kaiser, T. Schreiber, Information transfer in continuous processes, Phys. D Nonlinear Phenom. 166 (1) (2002) 43–62.
[44] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis, Phys. Rep. 441 (1) (2007) 1–46.
[45] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[46] J.D. Victor, Approaches to information—theoretic analysis of neural activity, Biol. Theory 1 (3) (2006) 302.
[47] F.B. Hildebrand, Introduction to Numerical Analysis, Courier Corporation, 1987.
[48] S. He, X. Zheng, D. Zeng, K. Cui, Z. Zhang, C. Luo, Identifying peer influence in online social networks using transfer entropy, Intelligence and Security Informatics. Springer, 2013 47–61.
[49] S. Aral, D. Walker, Creating social contagion through viral product design: a randomized trial of peer influence in networks, Manag. Sci. 57 (9) (2011) 1623–1639.
[50] Z. Xiang, U. Gretzel, Role of social media in online travel information search, Tour. Manag. 31 (2) (2010) 179–188.
[51] X. Cheng, C. Dale, J. Liu, Statistics and social network of YouTube videos, Quality of Service, 2008. IWQoS 2008. 16th International Workshop on 2008, pp. 229–238.
[52] T. Hogg, K. Lerman, L.M. Smith, Using stochastic models to predict user response in social media, Social Computing (SocialCom), 2013 International Conference on 2013, pp. 63–68.
[53] M.J. Atallah, Algorithms and Theory of Computation Handbook, CRC Press, 2002.
[54] L.J. Hubert, R.G. Golledge, Measuring association between spatially defined variables: Tjøstheim's index and some extensions, Geogr. Anal. 14 (3) (1982) 273–278.
[55] J. Teevan, S.T. Dumais, E. Horvitz, Personalizing search via automated analysis of interests and activities, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2005, pp. 449–456.
[56] J. Teevan, S.T. Dumais, E. Horvitz, Characterizing the value of personalizing search, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2007, pp. 757–758.
[57] R. Fagin, R. Kumar, D. Sivakumar, Efficient similarity search and classification via rank aggregation, Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data 2003, pp. 301–312.
[58] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, E. Vee, Comparing and aggregating rankings with ties, Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems 2004, pp. 47–58.
[59] P. Godfrey-Smith, M. Martínez, Communication and common interest, PLoS Comput. Biol. 9 (11) (2013) e1003282.
[60] A. Barg, A. Mazumdar, Codes in permutations and error correction for rank modulation, IEEE Trans. Inf. Theory 56 (7) (2010) 3158–3165.
[61] J. Goldenberg, S. Han, D.R. Lehmann, J.W. Hong, The role of hubs in the adoption process, J. Mark. 73 (2) (2009) 1–13.
[62] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 1999.
[63] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in Twitter: the million follower fallacy, 4th International AAAI Conference on Weblogs and Social Media (ICWSM) 2010, p. 8.
[64] M. Newman, A.-L. Barabasi, D.J. Watts, The Structure and Dynamics of Networks, Princeton University Press, 2006.
[65] M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, Measuring user influence in Twitter: the million follower fallacy, ICWSM 10 (10–17) (2010) 30.
[66] D. Romero, W. Galuba, S. Asur, B. Huberman, Influence and passivity in social media, Machine Learning and Knowledge Discovery in Databases2011 18–33.
[67] R. Ghosh, K. Lerman, Predicting influential users in online social networks, arXiv Preprint arXiv:1005.48822010.
[68] S. Aral, D. Walker, Identifying influential and susceptible members of social networks, Science 337 (6092) (2012) 337–341.
[69] A. Goyal, F. Bonchi, L.V. Lakshmanan, Learning influence probabilities in social networks, Proceedings of the Third ACM International Conference on Web Search and Data Mining 2010, pp. 241–250.

[70] H.P. Young, The diffusion of innovations in social networks, Economy as an Evolving Complex System. Proceedings Volume in the Santa Fe Institute Studies in the Sciences of Complexity, vol. 3 2002, pp. 267–282.

[71] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, R. Merom, "Suggesting friends using the implicit social graph," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010 233–242.

[72] T. Min, Influence of eWOM message on receivers' re-diffusion intention [J], J. Intell. 4 (2012) 028.

Saike He received a B.S. degree in automation and M.S. degree in computer science in 2007 and 2010, respectively, from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently pursuing a Ph.D. degree in Computer Science at the Institute of Automation, Chinese Academy of Sciences.

His research interest includes sentiment analysis, behavior modeling, information diffusion, and synchronization in complex networks.

Xiaolong Zheng is currently an Associate Professor at the Institute of Automation, Chinese Academy of Sciences. He got Ph.D. from the Institute of Automation, Chinese Academy of Sciences in 2009, M.S. from Beijing Jiaotong University in 2006, and B.S. from China Jiliang University in 2003. Xiaolong Zheng's current research interests include social dynamics modeling, social and behavior computing, big data analytics and prediction. Xiaolong Zheng has served as the Program Co-chair of the International Conference of Smart Health (ICSH2014), Pacific Asia Workshop on Intelligence and Security Informatics 2013 (PAISI2013), and Pacific Asia Workshop on Intelligence and Security Informatics 2011 (PAISI2011). He is also the Academic Secretary of ACM Social and Economic Computing Chapter from 2012–2013.

Daniel Zeng (M′04–SM′07) received the Ph.D. degree in industrial administration from Carnegie Mellon University in 1998.

He is a Research Professor at the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include software agents and multi-agent systems, intelligence and security informatics, social computing, and recommendation systems.