Collective Human Behavior in Cascading System: Discovery, Modeling and Applications

Yunfei Lu
Tsinghua University
luyf16@mails.tsinghua.edu.cn

Linyun Yu
Tsinghua University
yulinyun@bytedance.com

Tianyang Zhang
Tsinghua University
zty@powerlaw.ai

Chengxi Zang
Tsinghua University
zangcx13@mails.tsinghua.edu.cn

Peng Cui
Tsinghua University
cuip@tsinghua.edu.cn

Chaoming Song
University of Miami
c.song@miami.edu

Wenwu Zhu
Tsinghua University
wwzhu@tsinghua.edu.cn

***Abstract*—The collective behavior, describing spontaneously emerging social processes and events, is ubiquitous in both physical society and online social media. The knowledge of collective behavior is critical in understanding and predicting social movements, fads, riots and so on. However, detecting, quantifying and modeling the collective behavior in online social media at large scale are seldom unexplored. In this paper, we examine a real-world online social media with more than 1.7 million information spreading records, which explicitly document the detailed human behavior in this online information cascading system. We observe evident collective behavior in information cascading, and then propose metrics to quantify the collectivity. We find that previous information cascading models cannot capture the collective behavior in the real-world and thus never utilize it. Furthermore, we propose a generative framework with a latent user interest layer to capture the collective behavior in cascading system. Our framework achieves high accuracy in modeling the information cascades with respect to popularity, structure and collectivity. By leveraging the knowledge of collective behavior, our model shows the capability of making predictions without temporal features or early-stage information. Our framework can serve as a more generalized one in modeling cascading system, and, together with empirical discovery and applications, advance our understanding of human behavior.**

***Index Terms*—Collective Human Behavior; Information Cascades; Generative Framework**

I. Introduction

Collective behavior describes the phenomenon that people exhibit same behavior in a spontaneous way which do not reflect existing social structure [1]. The collective behavior lies in various social phenomena, ranging from the worldwide stock crashes in 2018, the popularity of the billion-view-video Ğangnam Style, to inconspicuous ones such as several customers having meals in a restaurant at some point. The collective behavior underlying these phenomena cannot be explained by existing social structure, but indicates that they share some unknown common points, which might be social-economic factors, interests or eating habits, etc. Although different opinions on interpretations, the existence and significance of collective behavior is widely recognized by public. The usefulness of the understanding of collective behavior is further proven by different research topics, such as predicting human mobility [2], analyzing human activity patterns [3] and so on.

However, collective behavior underlying information cascades in online social media is seldom explored. With the rapid growth of various online social media, detailed human behavior is documented at large scale, offering great opportunities to study the collective behavior. The most related works try to model the cascading system in social networks, such as modeling the popularity dynamics [4] or predicting the final size of cascades [5]. However, none of them tries to examine the collective behavior in cascading system which is embodied in the phenomena that a group of users always participate in same cascades collectively.

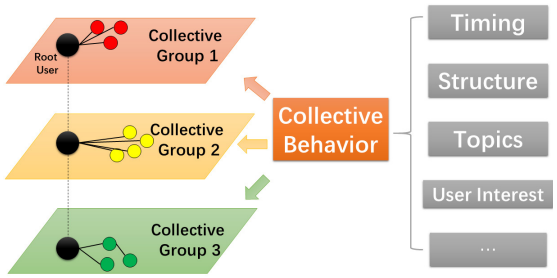


Fig. 1. *Illustration of collective behavior in cascading system.* We illustrate three collective groups of followers of a same root user. Followers in the same collective group tend to participate in the same information cascades, while followers in different collective groups seldom appear together. Possible factors, including timing, structure, topic of posts and user interests, are all responsible for such collective behavior. In this paper, we try to quantify and model such collective behavior in cascading system.

In this paper, we collect more than 1.7 million information spreading records from Tencent Weibo, a Twitter style social media in China, which explicitly document the detailed human behavior in this online information cascading system. We find evident collective behavior in real-world cascading system. In Fig.1 we illustrate three collective groups of followers of a same root user. Followers in the same collective group tend to participate in the same information cascades, while followers in different collective groups seldom appear together. The factors that influence aforementioned collective behavior are complex. First, timing can play an important role in collective behavior due to the shared daily routine of followers in the

same collective group. Second, followers in the same collective group can be embedded in a same tightly-knit community, indicating the impact of structure. Third, followers in the same collective group can possibly share similar interests in specific topics of tweets. Furthermore, all these factors can drive collective behavior simultaneously. Thus, the critical problem is: Can we model the collective behavior in cascading system caused by the complex and mixture factors?

In order to capture the collective behavior, we first provide metrics to quantify the collectivity of behavior, and then prove that previous information cascading models cannot capture the observed collectivity. Furthermore, we propose a generative framework to model the collective behavior in cascading social systems. Our genreative framework is based on point process with a latent user interest layer to capture the collective behavior at behavior level directly in cascading system. With this framework, we not only successfully capture the collective behavior, but also model cascading system more accurately on cascade popularity, structure and their correlation. Besides, our framework shows excellent extensibility and provides unique capability of predicting the popularity and participants of cascade with knowing only the identity of a few randomly selected participants.

In short, we summarize our contributions as follows:

- **Quantification of collectivity:** We discover the ubiquitous collective behavior in cascading system and propose metrics to quantify the collectivity of it.
- **A generative framework:** We propose a generative framework which is based on point process with a latent user interest layer to capture the collective behavior in cascading system. Our framework shows excellent extensibility and unification power.
- **Accuracy and usefulness:** With the knowledge of collectivity, our framework accurately matches real world cascading system in popularity, structure and collectivity, providing capability of making predicitons without temporal features.

The outline of the paper is: survey, method, collective behavior in cascading system, experiments, conclusions and future work.

II. Related Work

As the investigated problem is closely related to collective behavior and information cascades, we mainly review the related works in these two fields.

Collective behavior. Collective human behavior is tightly associated with human culture [6], which has wide applications in social network researches. Lehmann et al. [7] find the dynamic classes of hashtags from the spikes of collective attention in Twitter. Tang et al. [8] propose a scalable learning method for collective behavior based on sparse social dimensions. Banerjee [9] introduce a nature-inspired theory to model collective behavior from the observed data on blogs with aim of prediction. Candia et al. [3] discover the pattern of people calling activity from the collectivity in mobile phone records. However, none has discovered the collective behavior in cascading system and promote cascade modeling and prediciton with this collectivity.

Cascading system. In social network, a piece of information may get reshared multiple times: one shares the content with friends, several of these friends also share it with their respective sets of friends, then $information\ cascade$ develop and this phenomenon happens all the time, which constitutes $cascading\ system$. In recent years, many methods have been proposed to model cascading system and make prediction on cascades. Some of them focus on predicting the future size of a cascade with topological characteristics of the cascade [10] or dynamic information [11]. Among them Cheng et. al [12] indicate the significance of temporal features like retweeting rate among other features. Other methods, mainly based on point process [13] or survival analysis [14], attempt to model cascade dynamics and predict the evolution of popularity. Shen et. al [4] employ reinforcement poisson process to model cascades. Kobayashi et. al [15] take human circadian nature into account. Mishra et. al [16] combine feature-driven methods with generative modeling approaches.

However, these methods mainly treat each cascade as an independent process rather than considering the whole cascading system as an entirety. Former studies concentrate on popularity but disregard the structure patterns and collective behavior, for which there is still no comprehensive framework that can achieve accuracy in all three aspects of cascades: popularity, structure and collectivity.

III. Collective Behavior in Cascading System

Given a newtowrk $G = (V, E)$, where V indicates set of users and E indicates the relations, supposing a user u usually post different kinds of root tweets $< t_{1,u}, t_{2,u}, ..., t_{n,u} >$, where $t_{i,u}$ indicates the i_{th} kind of information, his followers are ususally only interested in one or a few kinds among them, due to their interest or some other reasons, and thus tend to appear in the cascades triggerred by the corresponding roots. Denoting the cluster of users who frequently retweet $t_{i,u}$ as $C_{i,u}$, if there are many cooccurence for users in same clusters but few cooccurence for users in different clusters, we regard this phenomenon as the collective behavior in cascading system. That means when u triggers a cascade, people involved by retweeting are always those interested in this kind of information, rather than randomly selected from followers or completely determined by social relationship. Users in same cluster share similar interests, which signifies that they are likely to keep this collectivity in future cascades. Also, users with opposite interest present mutual exclusion in their behavior of cascade participation.

An intuitional method to quantify collective behavior is to calculate the correlation of behavior between users. For each user u, we construct his behavior vector $<c_{1,u}, c_{2,u}, ..., c_{n,u}>$ where c_i indicates the i_{th} cascade. Let $c_{i,u} = 1$ if u is involved by retweeting or posting the root in c_i, otherwise $c_{i,u} = 0$. We use classical Pearson correlation coefficient to calculate the correlation between vector x and y as follows:

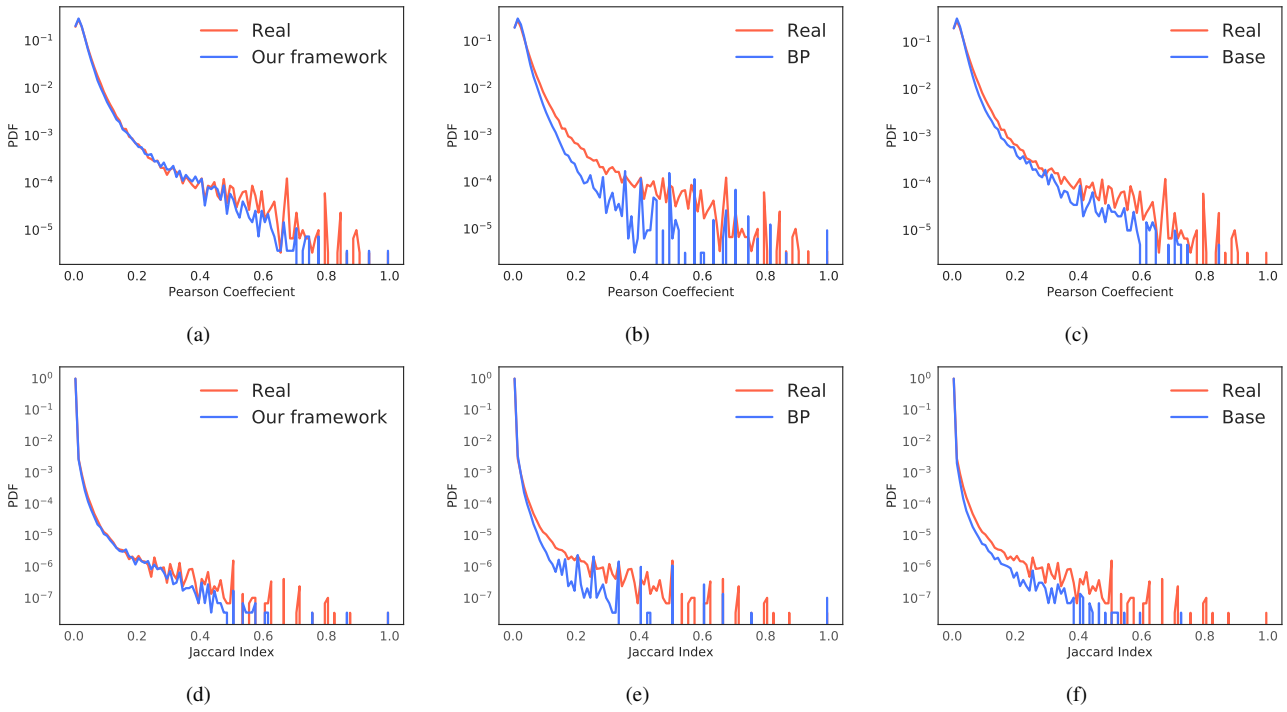


Fig. 2. *We detect the collectivity by calculating the behavior correlation between each pairs of users in network and plot distribution, finding that former methods cannot catch collective behavior in cascading system while our framework can fit very well.* The upper is Pearson Correlation Coefficient and the below is Jaccard Index. We compare the results of simulation data generated by different models with real data at each column. (a)(d) is our framework, choosing participants by inferring latent interest. (b)(e) is branching process, representing models that randomly select participants. (c)(f) is our base model, produced by removing interest layer from our framework, representing models that choose participants according to edge-based parameters.

$$|r_{x,y}| = \frac{|n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i|}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \quad (1)$$

Negative coefficients mean divergence while positive coefficients mean synchronization between two users. We only care about the degree of correlation rather than it is positive or negative, for which we calculate the absolute value.

Another comprehensible metric is the similarity of behavior. Jaccard index is a frequntly-used metric for the overlap ratio. Denote the set of cascades that u has joined by S_u, then we can use Jaccard index for comparing the similarity and diversity of behavior sets from two users:

$$J_{u,v} = \frac{|S_u \cap S_V|}{|S_u \cup S_V|} = \frac{|S_u \cap S_V|}{|S_u| + |S_V| - |S_u \cap S_V|} \quad (2)$$

We calculate Pearson coefficients and Jaccard index between every two users and draw the distribution in Fig 2. Our framework fits the real distribution extremely well. As for previous models that ignore the collectivity, there are mainly two types of methods for them to choose participants. The first type is completely indifferent to the identity of participants and will choose randomly. The second type of model has edge-based parameters which indicate the intensity of relation and are the basis of making choice. We use branching process(BP) and our base model(Base), which belong to the two types above respectively, to examine the ability for these models to capture collectivity. As we can see, collectivity in real data are obviously higher than synthesized data generated by these models, which means previous methods of these two categories fail to model collective behavior.

One shortcoming for the edge-based metrics is that they are easily deliquated by huge quantity of irrelevant point pairs in social network, for which we define a new point-based collectivity measurement Col. Assuming u has m followers and has posted n tweets, the collectivity measurement Col on u is defined by:

$$Col_u = \frac{\sum_{(x,y)\in N_u} |r_{x,y}|}{\binom{k}{2}} \quad (3)$$

where N_u is the tweet set of u and k is the number of followers. Col is the average absolute value of Pearson correlation coefficient among all pairs of tweet vectors. High coefficient between two tweets means that they are retweeted by nearly same users or completely different users, both of which are manifestation of collectivity. Followers show strong collectivity toward his tweets if a user has a high value of Col. We discard Jaccard index here because the average value for it is meaningless, since high positive correlation(value towards 1) and high negative correlation(value towards 0) will produce low relation in average.

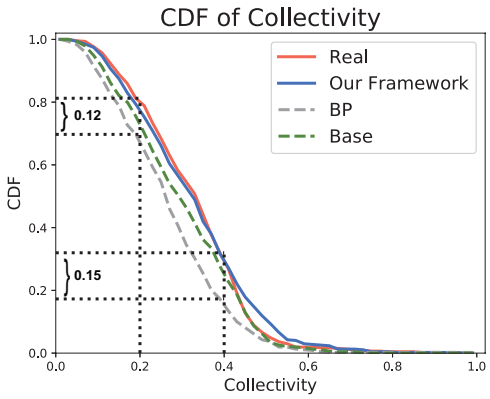


Fig. 3. *Our model match the reality well and other methods lose lots of collectivity (0.12 among 0.81 for weak part and 0.15 among 0.31 for medium part) on the cumulative distribution function of Col.* We can also judge that the collective behavior is considerably general in cascading system.

From Fig.3 we can see, more than 80% users in real data have average Pearson coefficient bigger than 0.2, and over 31% users bigger than 0.4. Generally, Pearson coefficient of 0.2 means relatively obvious relation (although weak), and value of 0.4 means medium relation. Considering the large quantities of tweet vector pairs, this is a strong evidence for the existence of collectivity. However, previous methods ignored the knowledge behind collective behavior and can only generate behaviors of much lower collectivity.

IV. Proposed Method

In this section, we present our framework and generator in detail, introducing how to capture collectivity with latent user interest layer and make generation with the proposed method.

A. Framework intuition

Our framework based on network level, which means we take the whole cascading system as an entirety with behaviors mutually connected for training and generation. For generating the microscopic forming process of collective behavior, we adopt point process as the basis of our framework. Inspired by [17], we also use poisson process as the assumption of temporal part to control the rate of posting root tweets and retweeting. Since the formation of collective behavior is diverse and can be very complicated, we aim to capture the collectivity at the behavior level directly with a latent variable layer. We call this part user interest layer where the concept of $interest$ is expanded to represent any internal motivation for collectivity, which is learned from history behavior.

B. Our framework

For every user u in the network, the action of generating root tweets can be regarded as following a poisson process of intensity λ_u, which indicates the frequency of posting. Same user may post tweets based on different latent topics, following a topic distribution Φ_u. When other users gets access to a tweet tw generated by u, their actions can be separated into two steps. At first they need to see the tweet, and the view action of a user can also be regarded as following a poisson process approximately. Second, at the time of seeing they will make decisions whether to retweet it and get involved in the cascade. The decisions depend on following three factors:

i) **User interest ϕ.** Users have different interests to different latent topics. A tweet with their preferential topic will be more likely to be retweeted.

ii) **Relationship intensity π.** If u is retweeted frequently by one of his followers v, it means that the relation between u and v is strong and u may be important to v, for which v has more probability to retweet u in sebsequent cascades.

iii) **Position power r and decay factor x.** Position power determines the level of a user in a cascade. The author of root tweet has important influence on decisions of whether to retweet it for others, so we denote whether this user is the author by r. Besides, in reality, a cascade cannot spread endlessly and we find there is an exponential decay of retweet rate along with the depth of tweet in cascade increasing, for which we amend the rate of retweeting a tweet with depth d by multiplying x^d on it, where global decay factor $x \in$ (0,1).

With all the aforementioned, the rate for u to retweet a tweet with topic τ and depth d posted by v is $\pi_{u,v,r}\phi_{v,\tau}x^d$. The users who decide to retweet tw will get involved in the cascade with a new tweet tw_r. Iteratively, some users will see tw_r and decide to join or not based on the factors above. This process continues until no more new users join the cascade.

According to the process described above, we propose our user interest framework by steps. At first, consider the likelihood function of one cascade. As the author of root tweet, the probability of u_c generating a information at time t is denoted by $Q_{u_c}(t)$. For every users u who have joined the cascade by retweeting, the probability of u retweets a tweet tw at t is denoted by $P_{tw,v}(t)$. And for other user who already get access to c but have not been involved until the end time of observation, we denote the probability of u have not retweeted to tw until T_e by $R_{tw,u,T}$. Assuming that Q, P and R are mutually independent, the likelihood function of cascade c is:

$$\begin{aligned} L_c &= P(Q_{u_c}(t) \cap \left(\bigcap_{(tw,v)\in RS_c} P_{tw,v}(t) \right) \cap \left(\bigcap_{(tw,v)\in RS_c} R_{tw,u,T} \right)) \\ &= Q_{u_c}(t) \prod_{(tw,v)\in RS_c} P_{tw,v}(t) \prod_{(tw,v)\in NS_c} R_{tw,u,T} \end{aligned} \tag{4}$$

Considering that generating root tweets follows a poisson process of intensity λ_u, and that the latent interest distribution of u on root tweets is $\Phi_{u,\tau}$, we get:

$$Q_{u_c,\tau}(t) = \lambda_{u_c}\Phi_{u_c,\tau}e^{-\lambda_{u_c}(t-t_l)} \tag{5}$$

where t_l is the last time u generates an information. If there is no former tweets, t_l is set to be the start time of observation.

The probability for u to see tw at t is $\omega_u e^{-\omega_u(t-t_{tw})}$, where ω_u is poisson intensity for seeing and t_{tw} is when tw posted.

TABLE I
SYMBOLS AND DEFINITIONS

Symbols	Definitions
u_c	The sender of the root tweet in cascade c
r_{tw}	The position power, indicating whether a tweet tw is a root tweet. $r_{tw} = 1$ for root tw and $r_{tw} = 0$ otherwise
T_s, T_e	The start time and end time of the observation window
τ	The latent topic variable. We set the number of topic categories to be 5 in experiments.
$\Phi_{u,\tau}$	The distribution on latent user interest for posting root tweet, $\Sigma_\tau \phi_{u,r,\tau}$ = 1, $0 \leq \phi_{u,\tau} \leq 1$.
$\phi_{u,\tau}$	The distribution on latent user interest for retweeting, $\Sigma_\tau \phi_{u,r,\tau}$ = 1, $0 \leq \phi_{u,\tau} \leq 1$.
λ_u	The rate of generating a root tweet of a user, $\lambda_u \geq 0$
ω_u	The rate of viewing a tweet of a user, $\omega_u \geq 0$
$\pi_{u,v,r_{tw}}$	The relationship intensity of user v to u on a r-based tweet $\omega_u \geq 0$
x	The global decay factor to control the depth of generating cascades, $0 < x < 1$
PS_c	The set of (tweet, user) pairs who can get access to at least one tweet in cascade c from users they follow and possibly join cascade c
RS_c	The set of (tweet, user) pairs who are in PS_c and did get involved in cascade c
NS_c	The set of (tweet, user) pairs who are in PS_c but have not joined the cascade

After seeing a tweet, if u decide to retweet this tweet with topic τ and depth d, we get:

$$\begin{aligned} P_{tw,u,\tau}(t) &= P_{saw}(t - t_{tw}) P_{retweet} \\ &= \omega_u e^{-\omega_u (t - t_{tw})} \pi_{u_{tw},u,r_{tw}} \phi_{u,\tau} x^{d_{tw}} \end{aligned} \tag{6}$$

where ϕ_u is the latent interest distribution on retweeting tweets, distinguished with Φ_u. For experiments in Sec. V, we assume there are 5 latent interest variables in total.

For u who has not retweeted an accessible tw, there are two possible situation: u has not seen tw till T_e, or u has seen it but decide not to reply. So we get:

$$\begin{aligned} R_{tw,u,T,\tau} &= (1 - P_{saw}) + P_{saw}(1 - P_{retweet}) \\ &= 1 + \pi_{u_{tw},u,r_{tw}} \phi_{u,\tau} x^{d_{tw}} (e^{-\omega_u (T_e - t_{tw})} - 1) \end{aligned} \tag{7}$$

Substitute equation 5 6 7 into equation 4 and summate on all interest variables. The likelihood function of cascade c is:

$$\begin{aligned} L_c = & \sum_\tau \lambda_{u_c} \Phi_{u_c,\tau} e^{-\lambda_{u_c}(t - t_l)} \\ & \prod_{(tw,v) \in RS_c} \omega_u e^{-\omega_u (t - t_{tw})} \pi_{u_{tw},u,r_{tw}} \phi_{u,\tau} x^{d_{tw}} \\ & \prod_{(tw,v) \in NS_c} (1 + \pi_{u_{tw},u,r_{tw}} \phi_{u,\tau} x^{d_{tw}} (e^{-\omega_u (T_e - t_{tw})} - 1)) \end{aligned} \tag{8}$$

A whole cascading system S contains all cascades generated by all users. The set of cascades generated by u is denoted by C_u. Thus the likelihood function for S is:

$$\begin{aligned} L_{cs} &= \prod_{u \in V} \prod_{c \in C_u} L_c \\ &= \prod_{u \in V} e^{-\lambda_u T_{se}} \prod_{c \in C_u} (\sum_\tau \lambda_u \Phi_{u,\tau} \\ &\quad \prod_{(tw,v) \in RS_c} \omega_v e^{-\omega_v (t - t_{tw})} \pi_{u_{tw},v,r_{tw}} \phi_{v,\tau} x^{d_{tw}} \\ &\quad \prod_{(tw,v) \in NS_c} (1 + \pi_{u_{tw},v,r_{tw}} \phi_{v,\tau} x^{d_{tw}} (e^{-\omega_v (T - t_{tw})} - 1)) \end{aligned} \tag{9}$$

where $T_{se} = T_e - T_s$ is the length of the observation.

Justification of the model:

Timing. λ and ω control the temporal part of cascading system, indicating the frequency of posting roots and retweeting respectively. With user-specific λ and ω, our framework can generate root tweets and retweets with consistent time distribution with empirical data, standing out from the models with no dynamics part. Our framework can handle more complicated dynamic patterns by replacing poisson process with more complex dynamic model.

Structure. π indicates the hetergeneous influence of information sender to his followers, independent with the content. Big π means that this relation is impactive and the receiver often retweet information from the sender. If π is small, the edge between sender and receiver is weak and interactions seldom happen on it. Combining the following two layers with π, our framework can make accurate generations on both popularity and structure and their correlation.

Other factors of collective behavior. Φ and ϕ make up the latent user interest layer, indicating the distributions of topics and user interests respectively, making it possible to catch collective behavior without content. Furthermore, Φ and ϕ are learned from history behavior, for which they can capture the complicated collectivity directly at behavior level. Inhomogeneous distribution means this user has strong preference for specific topic, and his behavior on these categories of information is very different from other kinds. Homogeneous distribution indicates that the behavior of user seldom changes when faced with information on various topics. Users with similar interest tend to appear in same cascade more frequently while users with unmatched interest have few cooccurence, both of which reflect the collectivity in cascading system.

C. Parameter estimation

Our parameter set consists of $\{\lambda, \Phi, \phi, \omega, \pi, x\}$. Since λ is only related to root tweets, we calculate λ directly from the times of generating root tweets in observation firstly. For other parameters, considering that there is unobservable latent variable distribution parameter Φ and ϕ, we learn the parameters by Expectation Maximization(EM) Algorithm [18] rather than maximizing likelihood function directly.

1) Optimize with EM algorithm: Let $\sum_\tau q_{c,\tau} = 1$, the log-likelihood function of cascading system S is:

$$
\begin{aligned}
lnL_c = & \sum_{c\in S} ln(\sum_{\tau} q_{c,\tau} \frac{\lambda_{u_c}\Phi_{u_c,\tau}}{q_{c,\tau}} \cdot \\
& \prod_{(tw,v)\in RS_c} \omega_v e^{-\omega_v(t-t_{tw})}\pi_{u_{tw},v,r_{tw}}\phi_{v,\tau}x^{d_{tw}} \\
& \prod_{(tw,v)\in NS_c} (1+\pi_{u_{tw},v,r_{tw}}\phi_{v,\tau}x^{d_{tw}}(e^{-\omega_v(T-t_{tw})}-1))) \\
& - \sum_{u\in V}\sum_{\tau}\lambda_u T_{se} d\tau
\end{aligned} \tag{10}
$$

With Jensen inequality we get:

$$
\begin{aligned}
lnL_c \geq & \sum_{c\in S}\sum_{\tau} q_{c,\tau}(ln\lambda_{u_c} + ln\Phi_{u_c,\tau} + \\
& \sum_{(tw,v)\in RS_c} (ln\omega_v - \omega_v(t-t_{tw})) + \\
& \sum_{(tw,v)\in RS_c} (ln\pi_{u_{tw},v,r_{tw}} + ln\phi_{v,\tau} + d_{tw}lnx) + \\
& \sum_{(tw,v)\in NS_c} ln(1+\pi_{u_{tw},v,r_{tw}}\phi_{v,\tau}x^{d_{tw}}(e^{-\omega_v(T-t_{tw})}-1))) \\
& - \sum_{c\in S}\sum_{\tau} q_{c,\tau} lnq_{c,\tau} - \sum_{u\in V}\sum_{\tau}\lambda_u T_{se} d\tau \\
& = F_{CS}
\end{aligned} \tag{11}
$$

With EM algorithm, we update the value of $q_{c,\tau}$ with current values of parameters at E-step, and estimate parameters by maximizing F_{CS}, the lower bound of joint likelihood, at M-step. By alternately iterating E- and M-steps until the value of F_{CS} converges, we infer the value of $\{\lambda,\ \Phi,\ \phi,\ \omega,\ \pi,\ x\}$ from empirical data. In experiments we use gradient descent algorithm as optimizer in M-step. For the purpose of finding a good region of parameter space and getting faster convergence, we set the initial value of parameters using some prior knowledge from empirical data.

D. Generator

We design the generator by simulating the process of cascading system formation. During generation, any event that happens after the end of observation will be discarded, along with any potential subsequent events activated by it. In poisson process, we know:

$$
\begin{aligned}
P(X \leq t) &= 1 - e^{-\lambda t} \\
t &= -\frac{ln(1-P(X\leq t))}{\lambda}
\end{aligned} \tag{12}
$$

For the process of generating root tweets and seeing tweets, we first sample p from $u \in V$ as $P(X \leq t)$, then use equation 12 to infer t as interevent time. We present the generation process in Algorithm 1.

Algorithm 1: Generating Process

Input : Network structure $G = (V, E)$, observation window (T_s,T_e), parameter set $\{\lambda,\ \Phi,\ \phi,\ \omega,\ \pi,\ x\}$
Output: Behavior logs of cascades system during observation window

Collect F_u, the set of followers of each user $u \in V$, from E;
Set $tweet\_queue$ to be an empty queue;
Set $cascades$ to be an empty list;
for $u \in V$ **do**
  $t = T_s$;
  while $t \leq T_e$ **do**
    Sample $p \sim Uniform([0,1])$;
    $iet = -\frac{ln(1-p)}{\lambda_u}$;
    Sample $\tau \sim \Phi_u$;
    $t = t + iet$;
    $r = 1$; // Is a root
    $d = 0$; // Define depth of root to be 0
    $a = NULL$; // Root has no ancestor
    Place tweet (u, t, τ, r, d, a) into $tweet\_queue$;
  end
end
while *tweet_queue is not empty* **do**
  $tw = tweet\_queue.pop()$;
  if $t_{tw} < T_e$ **then**
    Place tw into $cascades$;
    for $v \in F_{u_{tw}}$ **do**
      $threshold = \pi_{u_{tw},v,r_{tw}}\phi_{v,\tau_{tw}}x^{d_{tw}}$;
      Sample $p \sim Uniform([0,1])$;
      if $p \leq threshold$ **then**
        $iet = -\frac{ln(1-p)}{\lambda_v}$;
        $t = t_{tw} + iet$;
        $r = 0$;
        $d = d_{tw} + 1$; $a = tw$; Place tweet (u, t, τ, r, d, a) into $tweet\_queue$;
      end
    end
  end
end
Make $cascade$ sorted and return;

V. Experiments

In this section, we evaluate the effectiveness of our user interest framework on the real data. We first introduce the dataset and then show that our model is more accurate than existing methods on matching real data in three aspects at macroscopic level. We also demonstrate the unique predicting power of our model on popularity and participants of cascades with utilization of collective behavior.

A. Datasets

Our experiments are conducted on an online information diffusion dataset [19] from Tencent Weibo[1], a Twitter-style social platform in China. It includes all cascades generated in the 10 days between Nov 15 and Nov 25 2011 on a sub-network. For each tweet, there is a triad $< u, t, a >$ to respectively represent the sender of tweet, sending time of tweet, and the tweet it reply to, from which we can easily construct the whole cascading system with dynamics. The underlying social network is reconstructed based on the observed retweeting activities. In this network there are 7625 users, 59828 directed edges. The whole cascading system contains 447453 cascades and 598169 tweets.

B. Accuracy

We validate the accuracy of our framework by answering whether the simulation data generated from our framework

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

can match real data on distribution of various metrics. The experiments are conducted by validating the accuracy of following three aspects: *(i)* the timing in cascading system, *(ii)* the popularity and structure of cascades, and *(iii)* the colletive behavior existing in real world. We train models with all 10 days data and then generate behavior logs of same length of time with modeling parameters.

1) Baselines for experiments: Our framework, based on point process, takes the whole cascading system as an entirety with behaviors mutually connected for training and generation. The generation is based on history behavior rather than early-stage information, which makes it different from most cascading models. Here We consider following representative generative models for comparison:

i) Branching process (BP) [20]: BP assumes that each user forwards the information to a set of offspring neighbors, whose size is determined by the offspring size distribution of this user in real data. Since our framework distinguish roots from retweets, we also reinforce BP by learning two offspring size distribution for each user, one for roots and one for others.

ii) Epidemic model (EP) [21]: EP regards the spread of information as a process of contagion between users through their relationship. When infected with the information, the user may recover with certain probability, or spread the information to his followers on the edge-based infectious rate. We also extend the epidemic model by learning two infectious rate for each edge, distinguishing whether the information is root.

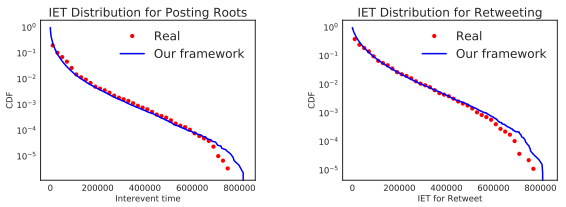
iii) Our framework without interest layer (Base): We remove latent interest distribution parameters Φ and ϕ from our framework and employ the rest part as our third baseline. From comparing the results with and without interest layer, the effectiveness of considering collectivity is more outstanding.

Since BP and EP are unable to generate the process of posting root tweets, we let each user to generate root tweets as much as the number he posted in real data for these models in experiments.

2) Accuracy of timing: We evaluate the effect of our temporal part by comparing the interevent time(IET) distribution of generated data with reality, which is a significant metric of information spreading dynamics [22]. Note that BP and EP have no dynamic property and cannot generate tweets with timestamps, not to mention IET distribution.

As shown in Fig. 4, the temporal layer of our framework fits the reality pretty well on the interevent time distribution of posting roots and retweets. We also notice that there is a small divergence between our framework and reality for big IET. It is caused by the non-poisson nature in human behavior and can be solved by extending our framework with more complicated dynamic model. See Sec. VI.

3) Accuracy of popularity and structure: We evaluate the accuracy on not only popularity but also structure, which gets few attention in most cascading prediction method. Popularity n measures the total number of tweets involved in a cascade. Regarding a cascade as a treelike structure, we choose the following metrics for our experiments which are representative



(a) (b)

Fig. 4. *Our model fits reality well on the interevent time cumulative probability distribution(CDF) of posting roots (a) and retweets (b).* For better visual effect on the tail, we plot the curve at semilog scale.

enough to describe and quantify structural patterns of information cascades [23]:

i) **Depth** measures the largest distance from root to other tweets in cascades.

ii) **Width** is the largest number of offsprings diverged from a same tweet in the cascade.

iii) **Wiener Index** is defined as the sum of the lengths of the shortest paths between all pairs of tweets in the cascade.

iiii) **Diameter** is the length of the longest path between two tweets in the cascade. Since the positive correlation between wiener index and popularity is too strong, we use diameter as replacement for the experiments of 2-dimensional distribution.

Since generation for once is of high contingency, to recede the instability, we partition x-axis into bins with equal width at log scale, and the real width is denoted by w_i. To avoid the result being dominated by first few data points, we evaluate the popularity and structure accuracy between the empirical distributions $\{x_1, x_2, ..., x_k\}$ and the estimated distributions $\{\hat{x_1}, \hat{x_2}, ..., \hat{x_k}\}$ by considering logarithmic average mean error. We calculate $logMAE$ on cumulative distribution function(CDF) to refrain 0 from appearing in logarithm. The CDF is given by $X_i = \sum_{j \leqslant i}^{k} p_{x_j}$ and denoting the numbers of estimation by k, $logMAE$ is calculated as following:

$$logMAE = \frac{\sum_{i=1}^{k} |logX_i - log\hat{X}_i|}{k} \tag{13}$$

The physical meaning of $logMAE$ is the area between two distribution curves on a log-scale plot.

From Fig.5 we can see our framework fits the reality much better than all baseline models. Table II presents the $logMAE$ on various metrics. Our framework wins with obvious superiority on all metrics except width. We are 32.68% better on popularity, 12.73% better on depth, 36.44% better on diameter and 32.35% better on wiener index than the best baseline. The theory of BP is based on the distribution of offsprings number, which is identical to the definition of width, for which BP achieves smallest error, though our framework also perform well on this metric. Note that baseline models may achieve good fitting on a portion of metrics, however, only our framework match reality in all metrics, which show its excellent comprehensiveness and unification power.

Right distribution for one metric separately not always means right cascade generation on all of them. The correlation

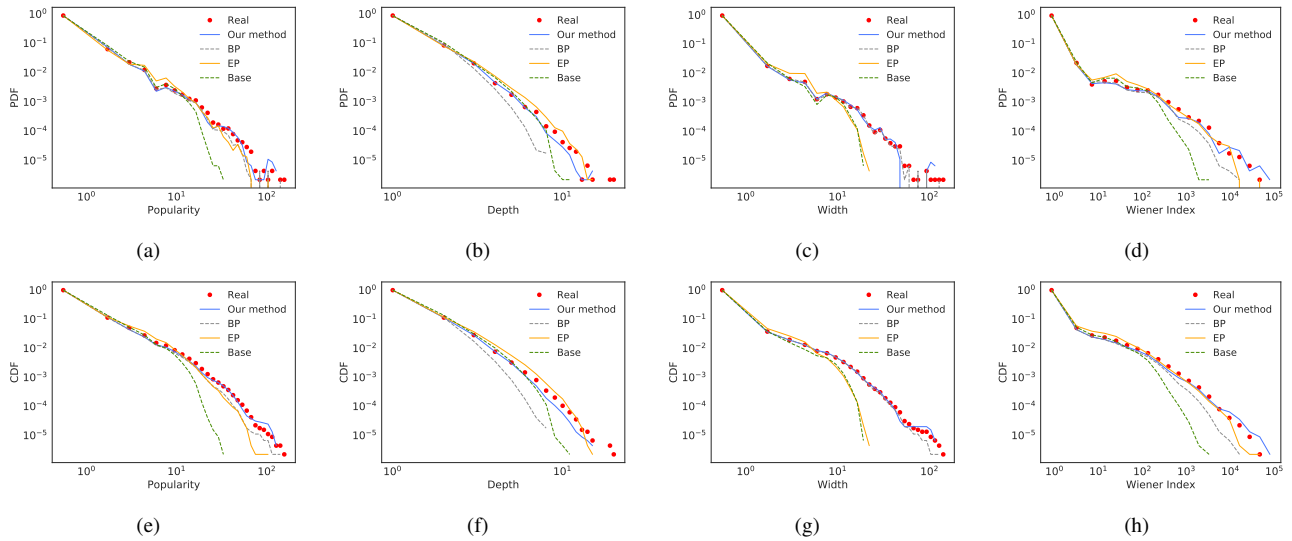


Fig. 5. *Our framework fits perfectly on all metrics above while other baselines can only handle part of them.* The upper is probability distribution function and the below is cumulative distribution function. From left to right is (a)(e) $Popularity$, (b)(f) $Depth$, (c)(g) $Width$, (d)(h) $WienerIndex$.

between metrics are also important. The two-dimensional distribution in Fig.6 show how well our framework captures the internal relation between these metrics, which means our framework is comprehensively accurate on popularity and structure. The fitting of BP is obviously worse and we omit other baselines for two-dimensional distribution since BP is the best baseline in Fig.5.

TABLE II
LOG MAE ON VARIOUS METRICS. WINNER IN BOLD.

	log MAE			
$\times e^{-3}$	Our framework	BP	EP	Base
Popularity	**2.088**	3.102	4.447	9.813
Depth	**1.801**	3.037	2.064	4.082
Width	0.044	**0.018**	2.164	1.881
Diameter	**1.158**	3.007	2.106	1.822
Wiener Index	**0.975**	1.442	1.889	1.675

4) Accuracy of collective behavior: We have demonstrated that our framework accurately captures collective behavior in cascading systems while baseline models are incapable in Section 3. As shown in Fig.3, the method that randomly choose participants generates only 69% users with Col more than 0.2, 12% lower than reality; and only 15% users has Col bigger than 0.4, nearly half of real distribution. However, our framework achieve high accuracy on fitting collectivity.

C. Predictions with collectivity

Most of previous models for cascading prediction depend on temporal features and early stage information, which demand trace of cascade evolution at global environment. But in reality, we generally only know what happen around us and temporal information along the forward chain is also hard to trace. With only local information and no temporal features, can we still make predictions? Our framework provides a new method for cascading prediction under this situation.

We train our framework and baseline models with the first 5 days data. All the testing data is chosen from last 5 days and we removed the users with hugely different behavior in this two periods. Our prediction are based on the participants knowledge rather than temporal features. Note that BP always randomly choose offsprings and participants knowledge is meaningless to it, so we focus on comparing with other baselines for following prediction.

1) Popularity Prediction: Former work has demonstrated the importance of first layer in popularity prediction [24], [25], i.e., the followers of the user posting root. Thus the problem we are about to solve is pretty practical: given the root user and several participants of a cascade without any temporal information, can we predict the cascade popularity on the first layer, namely, how many followers of root author will be involved in this cascade?

We select all the cascades with popularity more than 5, and randomly sample 4 participants together with the author of root as known information. For our framework there are two useful clues: *(i)* the popularity already reaches 5. We can infer an interest distribution Φ for this cascade by looking back the likelihood of former cascades with popularity larger than 5 for this user. *(ii)* this 4 participants joined this cascade. We can calculate the probability for these 4 users to retweet together on each topic variable, based on the interest distribution of the users learned before. After normalization on these probabilities, we get the inference of interest distribution ϕ.

Thus, the popularity predicted by our framework is:

$$E_{pop} = \sum_{i=1}^{k} \frac{\Phi_i \phi_i}{\sum_{i=1}^{k} \Phi_i \phi_i} E_i \tag{14}$$

where V is the set of followers and E_i is the expected value of retweets from followers for a root tweet with latent interest variable i. For EP and Base, they just calculate expectation popularity based on relation intensity or infectious rate.

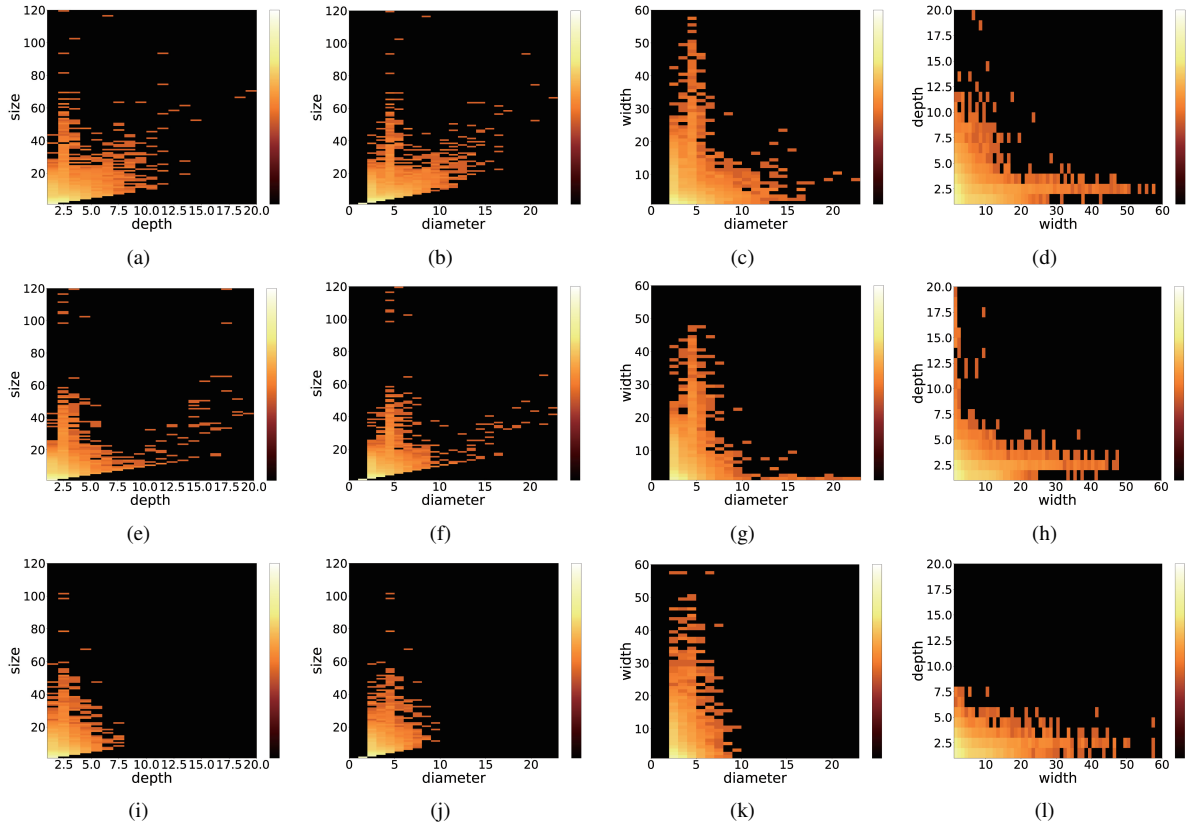


Fig. 6. *Our framework match the shape of two-dimensional distribution which reflects the correlation between metrics and whether models produce identical cascades to real data.* The upper is realdata, the middle is our framework and the below is BP. (a)(e)(i) $Popularity$ vs. $Depth$, (b)(f)(j) $Popularity$ vs. $Diameter$, (c)(g)(k) $Width$ vs. $Diameter$, (d)(h)(l) $Depth$ vs. $Width$.

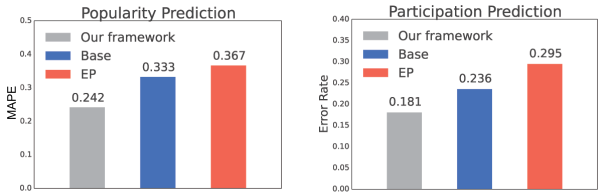


Fig. 7. *The mean average percentage error(MAPE) for popularity prediction(a) and the error rate for participation prediction(b).* We achieve 0.24 of MAPE and 0.18 of error rate, surpassing all the baselines.

We use mean average percentage error(MAPE) as metric. Fig.7(a) shows that as an expectation based method, we achieve MAPE of 0.24 with obvious advantage over baselines, which means latent interest information from known participants is significant and useful.

2) Participants prediction: After predicting how many followers will retweet root, we predict who will retweet among all the followers with knowing several participants and the popularity on first layer as input. Participants prediction can lead to promising applications such as forecasting whether influential users will join the cascade. Existing studies for this problem most set on diffusion content and user profiles [26] [27], or rely on temporal and early adopter features [28]. Different from them, our framework only need a randomly selected subset of participants.

We still select the cascades larger than 5 and randomly select 4 participants but avoid selecting followers of root author u. For all the followers of u, who really retweet him is regarded as positive example and others as negative sample. However, since u usually has lots of followers, there are much more negative examples than positive ones. To construct an unbiased set, we randomly choose equal number of negative examples with positive ones from followers for each prediction, and predict which of them will retweet.

The prediction process is similar to equation 14. First we use same method to infer the latent interest distribution of this cascade. Then we calculate likelihood for each follower to retweet this root based on interest distribution, together with other parameters. At last we rank the likelihood and pick first k followers as our result.

As shown in Fig.7(b), our error rate is 18.14%, superior to baselines which can only capture the frequency of interaction and rank followers according to the relation intensity. Our framework can infer interests from a portion of participants and estimate the probability for other users to join this cascade, for which we improve the accuracy.

VI. Conclusions & Future Work

In this paper, we study the collective behavior underlying in cascading system and design the metrics for quantifying collectivity, with which we prove that existing information

cascading models ignore the knowledge behind collective behavior and cannot capture the collectivity in real-world. For solving this problem, we propose a generative framework which is based on point process with a latent interest variable layer to capture the collectivity. We successfully explore and utilize the information behind collective behavior, with which our framework achieves high accuracy on cascades modeling in popularity, structure and collectivity, and also provides capability of making predictions without temporal features.

While the main superiority of our framework is the ability to capture and utilize collective human behavior, we also show its comprehensiveness in other aspects of cascading system. However, more efforts can be made by following our framework to get more accurate model. For example, we use poisson process to control the temporal part, but former works have found and captured some non-poisson nature of the information spreading dynamics [29]. It is promising to replace poisson process with these advanced dynamic model in our framework to match more complicated dynamic patterns, such as circadian rhythm [30]. With excellent extensibility and comprehensiveness, our framework can serve as a more generalized one in modeling information cascading system, and, together with empirical discovery and applications, advance our understanding of human behavior.

ACKNOWLEDGEMENT

The authors thank anonymous reviewers for many useful discussions and insightful suggestions. This work was supported in part by National Program on Key Basic Research Project No. 2015CB352300, National Natural Science Foundation of China Major Project No. U1611461; National Natural Science Foundation of China No. 61772304, 61521002, 61531006, 61702296. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, and the Young Elite Scientist Sponsorship Program by CAST. Song was partly supported by the National Science Foundation (IBSS-L-1620294).

REFERENCES

[1] R. H. Turner, L. M. Killian *et al.*, *Collective behavior*. Prentice-Hall Englewood Cliffs, NJ, 1957.

[2] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 2010, pp. 312–317.

[3] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of physics A: mathematical and theoretical*, vol. 41, no. 22, p. 224015, 2008.

[4] H. W. Shen, D. Wang, C. Song, and A. Barabási, "Modeling and predicting popularity dynamics via reinforced poisson processes," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 291–297.

[5] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1513–1522.

[6] W. C. Mcgrew, "Culture in nonhuman primates?" *Annual Review of Anthropology*, vol. 27, no. 1, pp. 301–328, 1998.

[7] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in twitter," *Computer Science*, pp. 251–260, 2011.

[8] L. Tang and H. Liu, "Scalable learning of collective behavior based on sparse social dimensions," in *ACM Conference on Information and Knowledge Management*, 2009, pp. 1107–1116.

[9] S. Banerjee and N. Agarwal, "Analyzing collective behavior from blogs using swarm intelligence," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 523–547, 2012.

[10] R. Cohen, S. Havlin, and D. Ben-Avraham, "Efficient immunization strategies for computer networks and populations," *Physical Review Letters*, vol. 91, no. 24, p. 247901, 2002.

[11] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, and S. Yang, "Cascading outbreak prediction in networks:a data-driven approach," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 901–909.

[12] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" pp. 925–936, 2014.

[13] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proc Natl Acad Sci U S A*, vol. 105, no. 41, pp. 15 649–53, 2008.

[14] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang, "From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics," in *IEEE International Conference on Data Mining*, 2015, pp. 559–568.

[15] R. Kobayashi and R. Lambiotte, "Tideh: Time-dependent hawkes process for predicting retweet dynamics," 2016.

[16] S. Mishra, M. A. Rizoiu, and L. Xie, "Feature driven and point process approaches for popularity prediction," in *ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1069–1078.

[17] T. Iwata, A. Shah, and Z. Ghahramani, "Discovering latent influence in online social activities via shared cascade poisson processes," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 266–274.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[19] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang, "Uncovering and predicting the dynamic process of information cascades with survival model," *Knowledge Information Systems*, vol. 50, no. 2, pp. 1–27, 2017.

[20] T. E. Harris, *THE THEORY OF BRANCHING PROCESSES*. Springer, 1963.

[21] C. Moore and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Physical Review E Statistical Physics Plasmas Fluids Related Interdisciplinary Topics*, vol. 61, no. 5 Pt B, p. 5678, 2000.

[22] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," pp. 561–568, 2011.

[23] C. Zang, P. Cui, C. Song, C. Faloutsos, and W. Zhu, "Quantifying structural patterns of information cascades," in *International Conference on World Wide Web Companion*, 2017, pp. 867–868.

[24] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs." *WOSN*, vol. 10, pp. 3–11, 2010.

[25] N. Sastry, E. Yoneki, and J. Crowcroft, "Buzztraq: predicting geographical access patterns of social cascades using social networks," *Acm Eurosys Sns Workshop*, pp. 39–45, 2009.

[26] J. Bian, Y. Yang, and T. S. Chua, "Predicting trending messages and diffusion participants in microblogging network," pp. 537–546, 2014.

[27] Q. Zhang, Y. Gong, Y. Guo, and X. Huang, "Retweet behavior prediction using hierarchical dirichlet process," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 403–409.

[28] C. T. Li, Y. J. Lin, and M. Y. Yeh, "Forecasting participants of information diffusion on social networks with its applications," *Information Sciences*, vol. 422, 2017.

[29] M. G. Rodriguez, J. Leskovec, and B. Schoelkopf, "Modeling information propagation with survival theory," *Philosophical Magazine Letters*, vol. 95, no. 2, pp. 85–91, 2013.

[30] L. Yu, P. Cui, C. Song, T. Zhang, and S. Yang, "A temporally heterogeneous survival framework with application to social behavior dynamics," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1295–1304.

[31] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabasi, "Human mobility, social ties, and link prediction," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1100–1108.