International Journal of Automation and Computing

17(3), June 2020, 353-363
DOI: 10.1007/s11633-020-1227-2

The Propagation Background in Social Networks: Simulating and Modeling

Kai Li[1] Tong Xu[1] Shuai Feng[2] Li-Sheng Qiao[1] Hua-Wei Shen[3]
Tian-Yang Lv[2] Xue-Qi Cheng[3] En-Hong Chen[1]

[1] School of Computer Science and Technology, University of Science and Technoloty of China, Hefei 230027, China

[2] IT Center of Chinese National Audit Office, Beijing 100073, China

[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Recent years have witnessed the booming of online social network and social media platforms, which leads to a state of information explosion. Though extensive efforts have been made by publishers to struggle for the limited attention of audiences, still, only a few of information items will be received and digested. Therefore, for simulating the information propagation process, competition among propagating items should be considered, which has been largely ignored by prior works on propagation modeling. One possible reason may be that, it is almost impossible to identify the influence of propagation background from real diffusion data. To that end, in this paper, we design a comprehensive framework to simulate the propagation process with the characteristics of user behaviors and network topology. Specifically, we propose a propagation background simulating (PBS) algorithm to simulate the propagation background by using users′ behavior dynamics and out-degree. Along this line, an IC$_{PB}$ (independent cascade with propagation background) model is adapted to relieve the impact of propagation background by using users′ in-degree. Extensive experiments on kinds of synthetic and real networks have demonstrated the effectiveness of our methods.

Keywords: Social network, information overload, propagation background, simulating, modeling.

1 Introduction

Online social networks play a significant role in our daily lives, especially as a crucial type of information propagation channel. Obviously, understanding the spreading of topics, ideas, and memes in social networks could be helpful to cognize human behaviors and get commercial interests. Thus, large efforts have been made on the propagation related problems[1–6].

In general, the information propagation process reflects users′ decisions on retweeting/sharing the content items they received, which could be affected by several factors, e.g., user preference or information topics. Among them, the competition between items could be an important reason[7]. Usually, people in modern life may be confronted with huge amount of information of any kind. However, they could digest only a little of them due to limited attention capacities. Therefore, users may only select those they like the most, or the most significant information to read, which leads to the fierce competition among information publishers to attract attention[8].

Indeed, prior works have already concerned this phenomenon of information overload which exceeds the attention capacity of users[9]. For instance, as shown in Fig. 1, more than 81% of users of Weibo follow more than 100 others, which leads to hundreds of new tweets every day, while most of them will be ignored. Similar situations could be found during the propagation of video, news and memes, in which popularity of each item will be degraded when a number of competing items are simultaneously available[10–13].

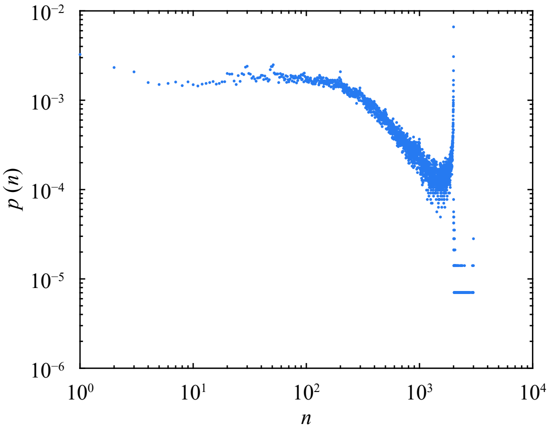


Fig. 1 Following number distribution of Weibo in double logarithmic coordinate. The horizontal axis denotes following number and the vertical axis denotes corresponding ratio.

Research Article
Manuscript received August 9, 2019; accepted February 5, 2020; published online April 5, 2020
Recommended by Editor-in-Chief Guo-Ping Liu

Along this line, two main approaches were designed for describing the propagation process in social networks. On the one hand, the mainstream of propagation modeling neglected the competition among items, for both the propagation/influence maximization problem[14–17] and the retweet prediction task[18, 19]. One possible reason may be that it is almost impossible to differentiate which items were missed or disliked from the item set that a user received but didn't retweet when the real propagation data was given. On the other hand, in a few agent based models, which result from the analysis on the propagation process, a user's memory and attention capacity were considered[20, 21].

However, traditional methods may fail to fully reflect the competition among items. Specifically, the settings of current agent based models are different from the real situation, and they could not be combined with the main stream machine learning based methods. To that end, in this paper, we propose a novel algorithm to simulate the competition and present a way to model it in the main stream methods. Specifically, the scenario we considered is the propagation of one or several items, when they are spreading, there are huge amount of other items which are also diffusing. If the users on the propagating paths of items we considered receive other items, they will compete for the limited attention of these users. And we use the term "propagation background" to represent the items other than the ones being considered and will define it formally in Section 3.

In this study, we concentrate on the impact of propagation background and ignore other factors except the influence from neighbors. For that we cannot extract ground truth from real propagation data, we need to construct the propagation background itself first. Thus we propose a novel propagation background simulating (PBS) algorithm, which can simulate the propagation background of online social networks. In the algorithm, content items are generated by each user, the items a user could deal with at a time is limited, and a user will not conduct activities at every step. Then, we present an IC$_{PB}$ (independent cascade[14] with propagation background) model to relieve the impact of propagation background, which can estimate the content item's diffusion scope more accurately. The model has the same framework with IC model, but before a node decides whether to retweet an item, there is a probability associated with its in-degree to miss the item.

Specifically, our main contributions can be summarized as follows:

1) We propose the PBS algorithm which can simulate the propagation background of online social networks. The algorithm has closer settings to real scenarios than previous agent based models.

2) We present the IC$_{PB}$ model that relieves the propagation background's impact on the diffusion process by improving IC model. Thus, it provides an approach to incorporate the effect of attention competition into existing propagation models.

3) We test our methods on a series of synthetic and real networks and the results demonstrate their effectiveness.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss some related works. Section 3 presents our methods, including PBS algorithm and IC$_{PB}$ model. Next, we report and analyze the experimental results in Section 4. Finally, we conclude this paper in Section 5.

2 Related works

In this section, we briefly review the works that modeled the competition among content items for users' limited attention in social networks.

Weng et al.[20] is the starting point of a series of works, which constructed an agent based model to study whether the competition of different memes may affect their popularity. In this model, each agent has a screen and memory, and both sized by time. The screen acts as receive buffer, and memory records items the agent posted. At each step, a randomly selected agent has a chance to create a new item. For other agents, the selected agent is also included if it doesn't produce a new item, they choose items from their screen according to a predefined probability, then either retweet the chosen items or tweet an item in memory randomly.

Qiu et al.[21] improved Weng's model with more reasonable settings and studied the relationship between the quality of an idea and its likelihood to become prevalent. In the model, the fixed space list of reverse chronological ordering is used as receive buffer. At each step, an agent is randomly selected, then either produces a new item or selects an item from the buffer by the items' quality and retweets it.

Furthermore, Gleeson et al.[22] studied the phenomenon of competition induced criticality based on a simplified version of Weng's model. Fan et al.[23] adopted a model similar to Weng's model to study the emotion contagion in online social networks. And Notarmuzi and Castellano[24] studied the dynamics of Qiu's model by considering some simplification and gave some explicit formulas.

Besides, Su et al.[25] proposed an improved susceptible-infected-recovered (SIR) model which considered the user's incomplete reading behavior. In their model, the rate of reading per unit time was added to the SIR model when describing the process susceptible state translates to infected state.

In general, our PBS algorithm is also an agent based method, but it is based on human dynamics and network topology. The settings of the PBS algorithm are closer to real situations. Specifically, our agents neither keep active every step nor conduct actions in random order, but have their active time sequence based on human dynamics. Furthermore, in PBS, the receive buffer is sized by space which is different from the Weng's model, and each agent could retweet more items when it is active which is different from the Qiu's model. As to Su's model, it considered the analogical problem with us, but that is a macro model while ours is a micro one.

3 Methodology

We will present our PBS algorithm and IC$_{PB}$ model in this section. But first of all, let's define what the propagation background is.

Definition 1. Propagation Background. Suppose there is an online social network N, I^c presents the set of content items that being considered, I_t presents the set of content items that diffuse in N at time stamp t. The set $IB_t^c = I_t - I^c$ is the propagation background of I^c at time stamp t.

It is obvious that the influence IB_t^c impacts on propagation process is competing users' attention with I^c, and I^c is very tiny compared to I_t, so that IB_t^c is almost equal to I_t. Therefore, what we will simulate is I_t, i.e., the propagation background of ϕ.

3.1 Propagation background simulating

Information overload is an ordinary state for users of online social networks. Therefore, the content items that exceed a user's attention capacity won't enter the user's viewing field and will be missed. Hence, the essence that content items compete for a user's limited attention is striving for the position of the user's viewing field. Because users will firstly see the recently posted items of their neighbors in most social networks, a list of reverse chronological ordering will be a reasonable approximation for a user's reading behavior. At any time, the most recent item will occupy the nearest position to the user. Furthermore, the content items will keep coming into the list continuously, so the final result of competing also depends on the time a user login, only the items come at the right time could really hold the good positions.

Thus, to simulate the propagation background, i.e., implementing the PBS algorithm, there are 4 key questions to be answered: Where do the new content items come from? When does each agent conduct activities? What is the size of each agent's attention capacity? And how does each agent act?

First, let's look at how an agent acts overall, that is the framework of PBS algorithm. Each agent has a receive buffer which contains the items it received. If an agent receives too many, the item that came first will be covered. Each agent has a step sequence that denotes at which steps the agent conducts activities. When an agent acts at a step, it will create a new item with a probability, then deal with the items in its receive buffer, and retweet each item with another probability.

Second, the content items are generated by each agent in our background model, which is in accordance to the scenario in real social networks where every user could contribute to the platform's contents. But as everyone knows, there are huge differences among the number of items created by different users. An active user may post tens of items per day, but a user who is very inactive may just post several items for one year or so.

Toubia and Stephen[26] empirically studied the two main source utilities that may motivate users to post content in a microblog: intrinsic utility and image-related utility. They pointed out that the intrinsic utility is derived from posting content to be viewed by many followers, and image-related utility is derived from having many followers. Therefore, the motivations for users to produce contents in social media could be described by their follower numbers. Furthermore, Kwak et al.[27] found there is a positive correlation between follower number and message number, and Li et al.[28] found there is a linear correlation between follower number and activity level for majority of users. Taking into account these cognitions, we employ each agent's out-degree, i.e., follower number, as the indicator of the probability an agent produces a new item and adopt a linear function to approximate it in our model. In addition, the propagation background results from ordinary uses' behaviors. And the time and energy an ordinary user could devote to a social network platform is finite so that there should be a maximum probability of post.

Let α_u denote the probability that an agent u produce a new item when it conduct activities, which could be computed as follows:

$$\alpha_u = \alpha_{\max} \times \frac{OD_u}{Max(OD)} \tag{1}$$

where $\alpha_{\max}$ denotes the maximum probability of all agents, which is a super parameter, OD_u denotes the out-degree of agent u, $Max(OD)$ denotes the maximum out-degree of all agents, and the agent with out-degree $Max(OD)$ generates new item with probability $\alpha_{\max}$.

Third, it is obvious that users of online social network platforms won't always conduct activities. So the assumption that agents act at every step is not in conformity with the truth of human behaviors. Fortunately, research in human dynamics have given us enough understanding of inter-posting time distribution. Besides, if the human dynamics of posting behavior are considered on minute scale, the periodic oscillations resulting from circadian rhythm[29, 30] should not be neglected. To these ends, we consider the posting behavior on day scale.

Fig. 2 presents the inter-tweeting time distribution of Weibo on day scale. Based on the statistical result, the inter activity conducting time distribution of an agent fits power law with slope $\lambda = 1.785$, i.e.,

$$p(\tau) \propto \tau^{-\lambda} \tag{2}$$

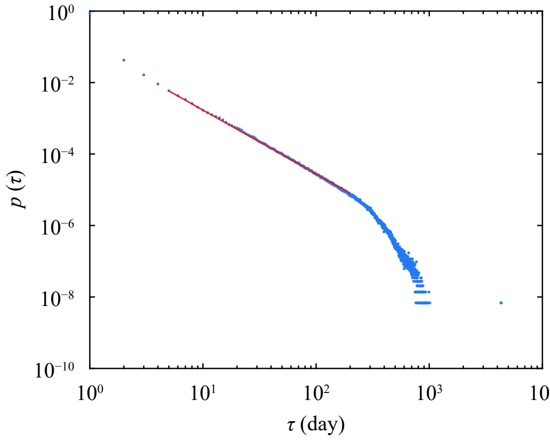


Fig. 2 Inter-tweeting time distribution of Weibo on day scale in double logarithmic coordinate. The distribution fits power law with cutoff, with the slope λ=1.785.

where τ indicates the inter active time.

In this paper, we adopt a power law distribution as approximation to real inter active time distribution and use the tool in [31] to generate the sequence of τ for each agent. Suppose that $\tau_{u1}, \tau_{u2}, ..., \tau_{un}, ...$ is the inter active time of agent u, and $T_{u1}, T_{u2}, ..., T_{uk}, ...$ is the action time of agent u, then we have

$$T_{uk} = \sum_{j=1}^{k} \tau_{uj}. \tag{3}$$

At last, about the size of a user's attention capacity or the buffer size of an agent, it is related to each concrete user and the time user login. It is obvious different users would like to view different number of items, i.e., they have different attention capacity. When a user logins, his mood and state (e.g., if it is leisure time) also affect how many items he'd like to read. Thus, the buffer size should be a function of user and time at least. But note that the items a user received will exceed his attention capacity at most times, so what is most important is to ensure the overflowing of buffer, thus we adopt a fixed size for simplification.

Algorithm 1 shows our method of simulating the propagation background. For convenience, we use an integer to represent an item and equip a receive buffer and a send buffer for each agent. In the algorithm, A is the agent set, and M is the adjacent matrix of a network. $M_{u,v} \neq 0$ means there is an edge from agent u to agent v, i.e., items could be propagated from u to v, and the value of M_{uv} is the propagation probability of the edge. T is the time when agents conduct activities, T_u is u' active time. $iBufSize$ denotes the size of buffer, and $iStep$ denotes the lasting time of agents' activities. The function $random()$ draws a random value from interval (0,1).

Algorithm 1. Propagation background simulating algorithm

```
Require: $A$, $M$, T, $\alpha_{\max}$, $iBufSize$, $iStep$.
Ensure: iMsg // The post number of agents
Create send buffer, receive buffer, posted item set for each agent;
Compute the probability of producing new item for each agent by Equation (1);
$ItemNo = 0$;
for $i$ =1 to $iStep$ do
    for each $u \in A$ do
        if $u$ is not active at current step then
            continue;
        end if
        if $random() < \alpha_u$ then
            $ItemNo++$;
            Put $ItemNo$ into $u$'s send buffer;
            Add $ItemNo$ into $u$'s posted item set;
        end if
        for each $item$ in receive buffer do
            read tuple($v$, $item$) from receive buffer; //$v$ is $u$'s neighbor, the source of $item$
            if $item \in u$'s posted item set then
                continue;
            end if
            if $random() < M_{u,v}$ then
                Put $item$ into $u$'s send buffer;
                Add $item$ into $u$'s posted item set;
            end if
        end for
        clear $u$'s receive buffer;
    end for
    $S = shuffle(A)$;
    for each $u \in S$ do
        if $u$'s send buffer is not null then
            for each $item$ in send buffer do
                Put the tuple ($u$, $item$) into followers' receive buffer;
            end for
        end if
        clear $u$'s send buffer;
    end for
end for
Compute each agent's post number by its posted item set.
```

At each step, for each agent that conducts activities, first, the agent utilizes its chance to produce a new item and deal with items in its receive buffer, all the items it posts will be put into its send buffer. Then, the items in each agent's send buffer will be put into its followers' receive buffer. In the second sub-step, for the sake of limited size of receive buffer, we deal with each agent in random sequence so that any agent will receive items from other agents in random order. And if an agent's receive buffer is full, the earliest one will be overwritten.

3.2 IC with propagation background

With the PBS algorithm, we could estimate a set of items' propagation scope under propagation background. But the computing cost is very expensive, and the receive buffer mechanism is still hard to incorporate into main stream propagation models. Fortunately, there is another choice to consider the impact of propagation background, that is to incorporate the effect of attention competition, not the process of competition, into existing models.

On the other hand, in many scenarios, the huge number of parameters can jeopardize the applicability of propagation models for their gigantic computational burden[17], so we prefer a relatively concise model. To this end, we consider relieving the impact of propagation background based on IC model.

The IC model considers a social network as a weighted graph. Nodes represent users and edges represent social ties. The weight of an edge is the probability that an item spreads from the starting node to end node of the edge. Specifically, at time t, A_t is the set of current active nodes. For node $u \in A_t$, it has only one chance to affect its inactive adjacent node v with the probability w_{uv}. If succeeding, v becomes active and will try to affect its own neighbors in the next time stamp $t+1$, otherwise v keeps inactive and u has no chance to affect v any more.

We have known items compete for users' limited attention while users suffer from information overload in real social networks. As to a specific user, the reason why he misses content items is he follows too many. According to Weng et al.[20], at the competing level of "standard", the probability that a message which a user received could be read is only 0.016.

Based on this cognition, we propose the IC$_{PB}$ model. The model's key point is how to measure the probability that a user received an item and missed it. It is obvious that the more a user follows the more likely the items he received will be omitted. Thus, we use the in-degree, i.e., following number, as an indicator for the probability that items being missed. And a concise function should be chosen according to Occam's Razor. Here, we adopt an exponential function to describe the probability, because the property of exponential function in accordance without understanding of the competing. Let $\beta(u)$ represent the probability that an item being missed by agent u:

$$\beta(u) = 1 - \mathrm{e}^{-\delta \times ID(u)} \tag{4}$$

where $ID(u)$ denotes the in-degree of agent u, δ is an adjust factor that needs to be selected by a concrete network.

Then, we get IC$_{PB}$ by adding the impact of propagation background to IC model. In IC$_{PB}$, the probability an item propagated from agent v to u is

$$p_{vu} = w_{vu} \times (1 - \beta(u)). \tag{5}$$

Thus, the probability that v activates u is the product of v's influence and the chance u sees the item.

So far, we have gotten the IC$_{PB}$ model. It just models the ideal case, and looks quite simple. Of course, further work is needed to describe the complex real scenarios, for instance, network embedding and neural networks could be used to replace β, but we only explain the basic idea here because it is helpful to understand the essence.

Finally, let's consider the influence function, denoted by $f(A)$, which is defined as the expected number of active nodes at the end of propagation process, given that A is the initial active set A_0. Then, similar to the IC model, the influence function of IC$_{PB}$ model is also satisfying the properties of non-negative, monotone and submodular, because the propagation probability of each edge is stable if the network structure remains unchanged. That is the basic assumption of most works today. So the greedy algorithm could still be utilized for influence/propagation maximization problem based on IC$_{PB}$ model.

4 Experiments

In this section, we demonstrate the effectiveness of our methods on four synthetic networks and two real-world networks. First, we show that our PBS algorithm could well simulate the propagation background of real social network. Then, we show that the IC$_{PB}$ model could relieve the impact on propagation process from the background.

4.1 Experimental setup

Data. In order to test our methods, we adopted networks of several types of topology, including ER (Erdos-Renyi) random network, WS (Watts-Strogatz) small world network and BA (Barabási Albert) scale free network[32], GC (benchmark Graph with Community) network[33], and two real networks, i.e., ego-Facebook and email-Eu-Core[1]. Besides, each synthetic network has 512 nodes, while ego-Facebook and email-Eu-Core have 1005 and 4039 nodes, respectively.

Evaluated methods. The PBS algorithm is utilized to simulate the propagation background. And the post number distribution is a direct consequence of propagation background. We would evaluate it by checking the post number distribution result of PBS and comparing it with the statistical distributions of models in [20, 21].

As to the IC$_{PB}$ model, we compared its simulated propagation scope to other models. All the scopes were computed by Monte Carlo simulation[2] and we just counted the retweet times of the given items. First, on the synthetic networks, we constructed the ground truth by simulating a set of items' propagation scopes while the propagation background was considered. In the diffusion process, agents would act the same way as in PBS algorithm. We compared the diffusion result of IC$_{PB}$ with

[1] http://snap.stanford.edu/data/

results of the IC model. Second, on the real networks, we compared IC_{PB}'s propagation scope to IC, IC_{ND}[16] and RAIN[34, 35] without ground truth.

Parameter setting. We set each edge equal propagation probability because it is better than setting arbitrary random values when there is no real data to deduce them. When considering an item propagation in a network, each item's source node number was sampling from the power law distribution with a slope 2.5[36], and its source nodes were selected randomly. The active time sequences of agents were created based on the method in Section 3.1.

The $\alpha_{\max}$ and $iBufSize$ are core parameters of PBS, their values were selected ensuring 2 things: 1) The distribution of agents' item number follows power law. 2) Information overload occurs on most agents. We adopted the grid search method on a GC graph to determine the values, and set $\alpha_{\max} = 0.12$, $iBufSize = 32$ in PBS algorithm. The corresponding item number distribution looks like Fig. 3 (d), and information overload occurs on over 70% nodes. Because of too many distribution figures and information overload rate pairs, we don't list them here.

The value of δ in IC_{PB} reflects the competing level, or on the contrary, the probability that an item existing in receive buffer will be read. In [20], the probability was set to 0.205, 0.016 and 0.001 corresponding to weak, standard and strong competitions, respectively. Here we set follower number 100 for the case of low competition.

The decay rate of the IC_{ND} model was set based on [10], we adopted the default value, i.e., 0.3^n, where n is the times that a user's neighbors had posted/reposted an item.

Enlightened by Yang et al.[35], in RAIN model, we considered 10% of nodes with the highest PageRank scores[37] to be opinion leaders, 10% of users with the lowest network constraint scores[38] to be structural hole spanners, and the remaining as ordinary nodes.

4.2 Experimental results

Background simulating. Research on human dynamics have found that the post number distribution of online social networks fits heavy tail or power law[7, 29].

Fig. 3 presents the simulating results of propagation background. Each method just runs once at each network. The parameters of the Weng's model and Qiu's model were the same as their original papers. PBS and Weng's methods run 4 096 steps, and Qiu's model runs 100 000 steps.

As seen from Fig. 3, the Qiu's model does not result in the expected long tail distribution on real-world networks, so it is not very suitable for simulating the propagation background. In addition, the idea that agents conduct activities in random order also makes it difficult to combine the Qiu's model with IC and other models. Though the distributions of PBS are somewhat more like to be power law than Weng's model, it is difficult to say which method is better to simulate propagation background simply based on their message number distributions. However, there are two points to note here. 1) PBS does not allow an agent to post the same content item many times, and the memory mechanism of the Weng's model determines that an agent could post an item many times. 2) PBS's buffering mechanism, activity mechanism and new item generation mechanism are closer to the real situation than Weng's method. Besides, the time complexity of PBS is lower than Weng's model, that is $O(\eta kn)$ verse $O(kn)$ given that k and n denote step number and agent number respectively, and $\eta \leq 1$. Therefore, PBS should be the best way to simulate the propagation background at present.

Propagation scope simulating. To show the effect of propagation estimating, we computed a set of items' diffusion scope under different methods by Monte Carlo simulation. The experiments were conducted under the same settings, i.e., we employed equal propagation probability for each edge and same source nodes for each network.

On each synthetic network, we conducted three experiments, i.e., computing the propagation scopes of a set of items with PBS, IC and IC_{PB}, respectively. As mentioned in Section 4.1, the propagation result of PBS was viewed as ground truth because it was the closest to actual circumstance. The results are shown in Table 1.

In the experiments, we employed a series of propagation probabilities, such as 0.25, 0.2, 0.15, 0.10, 0.05, 0.01, 0.005, the computed propagation scopes varied greatly when different propagation probabilities were used, but the order of propagation scopes of three methods was stable. We reported only the results of one propagation probability in Table 1. Furthermore, the parameter of equation (4) should be different for each network. But we didn't adjust it to make the results very beautiful and just show its effectiveness.

It could be seen that different propagation probabilities were utilized for different networks. This is because the buffer mechanism will become invalid when the propagation probability of network is below some thresholds. And the thresholds are different for each network.

The results show that IC model tends to overestimate the propagation scope. And IC_{PB} could get more reasonable results in general. Specifically, IC_{pb}'s results on ER and WS are just between IC's and PBS', they are as our expectation. IC_{pb}'s results of item1 on GC are lower than PBS', but the difference is very small. This is because the source node just lies in the fringe of the network and the impact from the buffer mechanism is lower than the effect from equation (4). Most of the IC_{pb}'s results on BA are lower than PBS'. The reason is that the topology of BA network is more sensitive compared to others, and

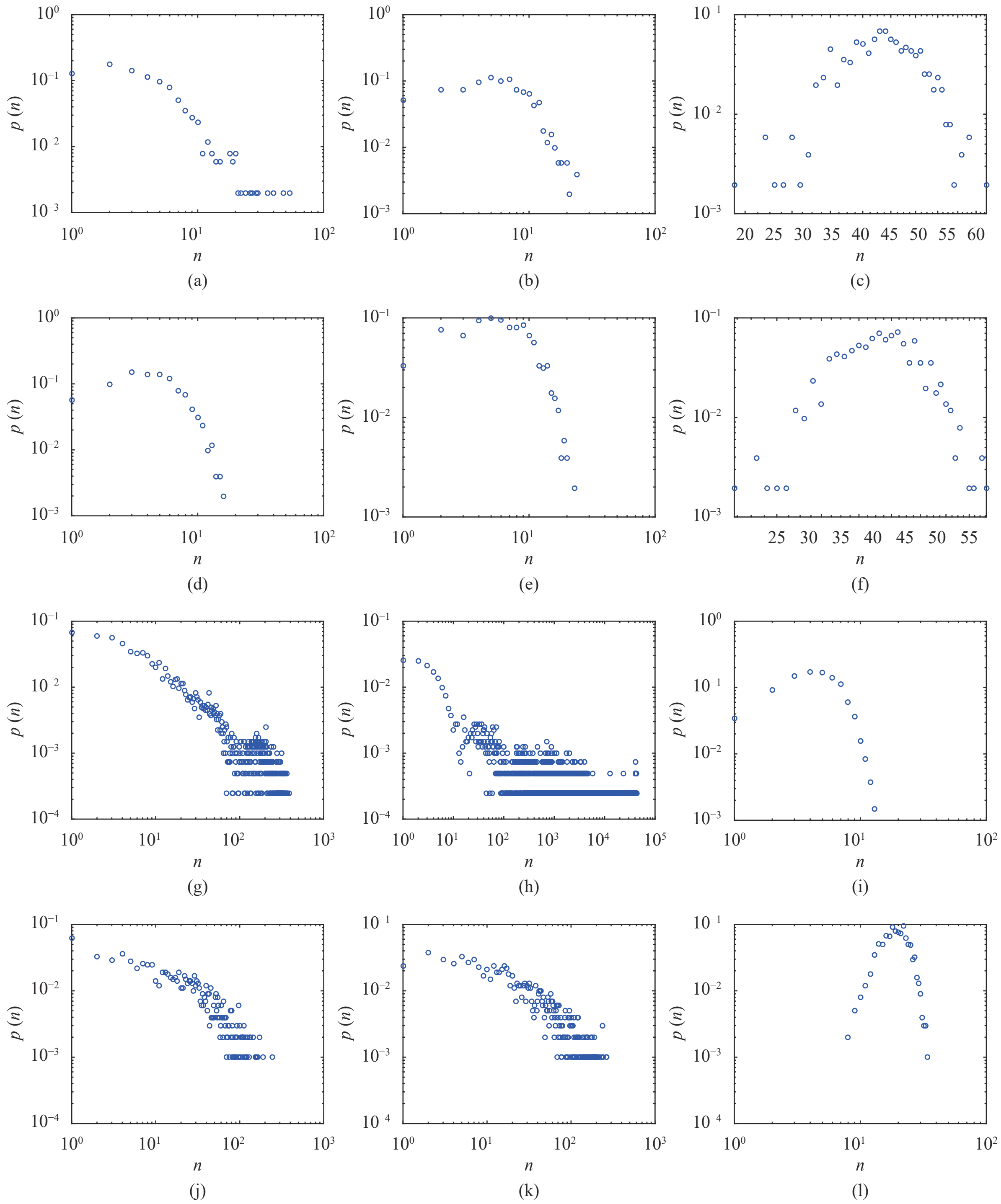


Fig. 3 Propagation background simulating results. The first, second and third columns are results of PBS algorithm, Weng's model and Qiu's model, respectively. The first, second, third and forth rows are results on BA network, GC network, ego-Facebook network and email-Eu-core network respectively.

the proper parameter of equation (4) for BA network is different from others. And we verified it in experiments.

On the two real networks, we conducted five experiments, i.e., computing the propagation scopes of a set of items with PBS, IC, IC$_{PB}$, IC$_{ND}$ and RAIN respectively by Monte Carlo simulation. The results are shown in Table 2.

If we consider only the results of PBS, IC and IC$_{PB}$, it is obvious that our statement on results of synthetic networks also holds. But there are also the results of IC$_{ND}$ and RAIN here. Thus, we couldn't decide whose results acted as ground truth. Therefore, let's look at them another way.

First, we could sort the simulated propagation scopes

Table 1 Simulated propagation scope on synthetic networks

Net	P	Method	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
ER	0.1	PBS	20.95	70.08	23.66	55.82	20.10	25.24	46.18	77.54	20.56	46.95
		IC	150.73	290.64	166.00	274.25	184.56	193.13	249.38	296.40	170.95	263.01
		IC$_{PB}$	48.35	136.10	52.37	114.61	59.24	63.97	95.69	147.29	54.60	105.90
WS	0.25	PBS	6.43	19.45	6.03	20.05	4.85	6.54	12.42	26.32	6.14	13.83
		IC	7.95	28.96	8.78	26.53	8.17	7.70	17.01	34.76	8.46	16.71
		IC$_{PB}$	6.76	25.35	7.38	22.15	7.18	6.74	14.69	29.77	7.15	14.18
BA	0.15	PBS	5.33	94.12	16.39	46.42	10.68	6.72	24.97	62.49	17.00	20.60
		IC	22.94	122.66	27.83	78.37	18.63	12.28	41.63	96.88	31.42	37.78
		IC$_{PB}$	10.73	85.87	12.69	43.15	8.98	5.55	20.93	55.73	15.36	19.24
GC	0.1	PBS	3.85	26.90	3.70	18.66	6.59	5.97	13.74	31.85	6.73	8.47
		IC	5.19	55.90	6.29	36.89	13.93	13.80	31.09	61.56	15.26	17.45
		IC$_{PB}$	3.83	33.56	4.57	21.82	7.62	7.79	17.84	37.38	8.63	11.06

The "P" in table means the propagation probability on each edge.

Table 2 Simulated propagation scope on real networks

Net	P	Method	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
ego	0.1	PBS	74.35	542.33	744.53	421.75	498.90	712.46	657.18	481.09	951.14	226.51
		IC	711.93	2891.82	2091.09	2069.88	2901.93	2901.21	2862.23	2889.50	2903.03	2251.94
		IC$_{PB}$	77.06	731.34	1416.48	739.98	1202.65	1541.30	1145.33	1045.41	1676.94	586.22
		IC$_{ND}$	33.27	159.86	195.29	128.48	133.50	198.59	191.52	136.23	260.99	148.71
		RAIN	215.59	2108.89	2320.66	1773.28	2373.25	2375.47	1785.37	2188.05	2392.95	1033.79
email	0.1	PBS	349.59	63.01	224.28	335.98	210.97	259.72	133.08	244.18	325.42	60.74
		IC	657.29	208.76	619.29	657.66	591.56	640.01	468.32	653.84	658.06	214.46
		IC$_{PB}$	565.86	74.01	340.10	549.13	316.16	392.87	200.66	460.86	517.90	86.20
		IC$_{ND}$	92.01	28.78	82.98	90.37	77.86	84.52	62.84	87.59	89.95	29.08
		RAIN	533.31	73.11	360.35	530.84	299.51	399.20	211.20	440.42	517.36	74.67

The "P" in table means the propagation probability on each edge. For RAIN, we set the propagation probability from opinion leader to other nodes 1.5P, the propagation probability from structural hole to other nodes P, and the propagation probability between ordinary nodes 0.5P.

of former four methods, i.e., IC$_{ND}$ < PBS < IC$_{PB}$ < IC. The order is very stable. Second, the results of RAIN fluctuate significantly compared to other methods. When in valley, it is less than IC$_{PB}$, and when at peak, it exceeds IC. This is because the source nodes belong to different types, say, opinion leader, structural hole or ordinary user. Third, IC$_{ND}$'s inhibitory effect on IC is very striking. It seems the inhibition is something more than necessary. Finally, from the perspective of mechanism, IC$_{PB}$, IC$_{ND}$ and RAIN model the different aspect of propagation process, and we are certain that IC$_{PB}$ does works, it grasps the underlying essence of propagation background.

After showing the overall effectiveness of IC$_{PB}$, we would like to demonstrate some details of the results in Tables 1 and 2. Fig. 4 shows error bar diagram of the effect of IC$_{PB}$, where horizontal axis denotes the number of the source node number of items, and the vertical axis denotes the proportion of diffusion scope reducing relative to IC model.

The effect of IC$_{PB}$ model was relatively stable on the

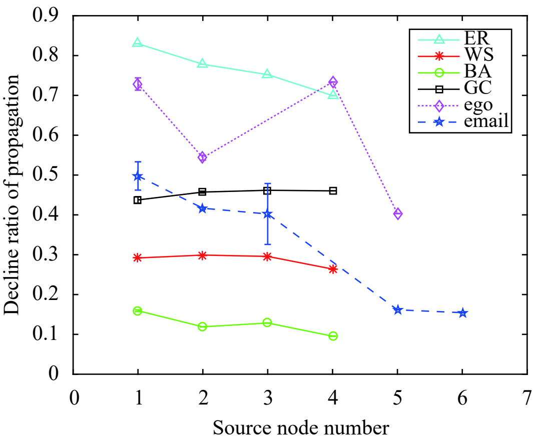


Fig. 4 Effect of IC$_{PB}$ comparing to IC

four synthetic networks. We think it is because they have a single topology structure. Nevertheless, the effects on real networks had obvious fluctuation. It maybe results from their hybrid structure, they are scale-free, small world and with community, etc. So deeper understanding of influence on propagation from the network topology structure is necessary.

4.3 Discussions

We would like to discuss two problems that are caused by the neglect of propagation background.

First, in online business practices, often the effect of word of mouth needs to be estimated. The problem is known as influence/propagation maximization aimed to select seed sets to maximize the effect of word of mouth. But we believe that the expected effect could not be reached if the propagation background was omitted.

Second, many tasks need to learn the propagation probability from real data by machine learning methods. In such cases, usually the items a user retweeted were considered as positive samples and others as negative samples. Thus, the items that a user missed were considered as those the user didn't like. So the learned propagation probabilities were underestimated, we believe.

5 Conclusions

In this paper, we proposed an agent-based algorithm to simulate the propagation background in online social networks. According to the validation performance, it could be among the best methods to simulate propagation background, and further establish a foundation for future research on social propagation. Also, we presented the IC_{PB} model by considering the effect of propagation background for better describing the propagation process. Extensive experiments have demonstrated that our method could effectively relieve the impact from propagation background on the diffusion process.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 91546110, 61703386, 61727809 and U1605251).

References

[1] J. J. Li, L. R. Wu, J. Y. Qi, Q. Yan. Research on information dissemination in online social network based on human dynamics. *Journal of Electronics & Information Technology*, vol. 39, no. 4, pp. 785–793, 2017. DOI: 10.11999/JEIT160940. (in Chinese)

[2] B. Chang, T. Xu, Q. Liu, E. H. Chen. Study on information diffusion analysis in social networks and its applications. *International Journal of Automation and Computing*, vol. 15, no. 4, pp. 377–401, 2018. DOI: 10.1007/s11633-018-1124-0.

[3] A. Tselykh, V. Vasilev, L. Tselykh. Management of control impacts based on maximizing the spread of influence. *International Journal of Automation and Computing*, vol. 16, no. 3, pp. 341–353, 2019. DOI: 10.1007/s11633-018-1167-2.

[4] T. Y. Jin, T. Xu, H. Zhong, E. H. Chen, Z. F. Wang, Q. Liu. Maximizing the effect of information adoption: A general framework. In *Proceedings of SIAM International Conference on Data Mining*, San Diego, USA, pp. 693–701, 2018. DOI: 10.1137/1.9781611975321.78.

[5] B. Chang, E. H. Chen, F. D. Zhu, Q. Liu, T. Xu, Z. F. Wang. Maximum a posteriori > estimation for information source detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, to be published. DOI: 10.1109/TSMC.2018.2811410.

[6] T. Xu, H. S. Zhu, H. Zhong, G. N. Liu, H. Xiong, E. H. Chen. Exploiting the dynamic mutual influence for predicting social event participation. *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1122–1135, 2019. DOI: 10.1109/TKDE.2018.2851222.

[7] K. Li, G.Y. Lv, Z. F. Wang, Q. Liu, E. H. Chen, L. S. Qiao. Understanding the mechanism of social tie in the propagation process of social network with communication channel. *Frontiers of Computer Science*, vol. 13, no. 6, pp. 1296–1308, 2019. DOI: 10.1007/s11704-018-7453-x.

[8] C. Orellana-Rodriguez, M. K. Keane. Attention to news and its dissemination on Twitter: A survey. *Computer Science Review*, vol. 29, pp. 74–94, 2018. DOI: 10.1016/j.cosrev.2018.07.001.

[9] H. A. Simon. Designing organizations for an information-rich world. *Communication, and the Public Interest*, M. Greenberger, Ed., Baltimore, MD, USA: The Johns Hopkins Press, pp. 40–41, 1971.

[10] F. Wu, B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 45, pp. 17599–17601, 2007. DOI: 10.1073/pnas.0704916104.

[11] R. Crane, D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 41, pp. 15649–15653, 2008. DOI: 10.1073/pnas.0803685105.

[12] J. Leskovec, L. Backstrom, J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Paris, France, pp. 497–506, 2009. DOI: 10.1145/1557019.1557077.

[13] S. Asur, B. A. Huberman, G. Szabo, C. Y. Wang. Trends in social media: Persistence and decay. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, pp. 434–437, 2011.

[14] J. Goldenberg, B. Libai, E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, vol. 12, no. 3, pp. 211–223, 2001. DOI: 10.1023/A:1011122126881.

[15] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978. DOI: 10.1086/226707.

[16] S. S. Feng, X. F. Chen, G. Cong, Y. F. Zeng, Y. M. Chee, Y. P. Xiang. Influence maximization with novelty decay in social networks. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, pp. 37–43, 2014.

[17] N. Barbieri, F. Bonchi, G. Manco. Topic-aware social influence propagation models. *Knowledge and Information*

Systems, vol. 37, no. 3, pp. 555–584, 2013. DOI: 10.1007/s10115-013-0646-6.

[18] S. Petrovic, M. Osborne, V. Lavrenko. RT to win! predicting message propagation in twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, pp. 586–589, 2011.

[19] M. M. Wang, W. L. Zuo, Y, Wang. A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. *Mathematical Problems in Engineering*, vol. 2015, Article number 936397, 2015. DOI: 10.1155/2015/936397.

[20] L. Weng, A. Flammini, A. Vespignani, F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, vol. 2, Article number 335, 2012. DOI: 10.1038/srep00335.

[21] X. Y. Qiu, D. F. M. Oliveira, A. S. Shirazi, A. Flammini, F. Menczer. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, vol. 1, no. 7, Article number 0132, 2017. DOI: 10.1038/s41562-017-0132.

[22] J. P. Gleeson, J. A. Ward, K. P. O′Sullivan, W. T. Lee. Competition-induced criticality in a model of meme popularity. *Physical Review Letters*, vol. 112, no. 4, Article number 048701, 2014. DOI: 10.1103/PhysRevLett.112.048701.

[23] R. Fan, K. Xu, J. C. Zhao. An agent-based model for emotion contagion and competition in online social media. *Physica A: Statistical Mechanics and its Applications*, vol. 495, pp. 245–259, 2018. DOI: 10.1016/j.physa.2017.12.086.

[24] D. Notarmuzi, C. Castellano. Analytical study of quality-biased competition dynamics for memes in social media. *EPL (Europhysics Letters)*, vol. 122, no. 2, Article number 28002, 2018. DOI: 10.1209/0295-5075/122/28002.

[25] Q. Su, J. J. Huang, X. D. Zhao. An information propagation model considering incomplete reading behavior in microblog. *Physica A: Statistical Mechanics and its Applications*, vol. 419, pp. 55–63, 2015. DOI: 10.1016/j.physa.2014.10.042.

[26] O. Toubia, A. T. Stephen. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, vol. 32, no. 3, pp. 368–392, 2013. DOI: 10.1287/mksc.2013.0773.

[27] H. Kwak, C. Lee, H. Park, S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, ACM, New York, USA, pp. 591–600. 2010. DOI: 10.1145/1772690.1772751.

[28] K. Li, T. Y. Lv, H. W. Shen, L. S. Qiao, E. H. Chen, X. Q. Cheng, Z. Sun. An empirical analysis on the behavioral differentia of the "Elite-Civilian" users in Sina microblog. *Physica A: Statistical Mechanics and its Applications*, vol. 539, Article number 122974, 2020. DOI: 10.1016/j.physa.2019.122974.

[29] Q. Yan, L. R. Wu, L. Zheng. Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 7, pp. 1712–1723, 2013. DOI: 10.1016/j.physa.2012.12.008.

[30] C. X. Wang, X. H. Guan, T. Qin, T. Yang. Modeling heterogeneous and correlated human dynamics of online activities with double Pareto distributions. *Information Sciences*, vol. 330, pp. 186–198, 2016. DOI: 10.1016/j.ins.2015.09.016.

[31] A. Clauset, C. R. Shalizi, M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009. DOI: 10.1137/070710111.

[32] T. G. Lewis. *Network Science: Theory and Applications*, Hoboken, USA: John Wiley & Sons, pp. 97–216, 2011.

[33] A. Lancichinetti, S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, vol. 80, no. 1, Article number 016118, 2009. DOI: 10.1103/PhysRevE.80.016118.

[34] Y. Yang, J. Tang, C. W. K. Leung, Y. Z. Sun, Q. C. Chen, J. Z. Li, Q. Yang. RAIN: Social role-aware information diffusion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, USA, pp. 367–373, 2015.

[35] Y. Yang, J. Jia, B. Y. Wu, J. Tang. Social role-aware emotion contagion in image social networks. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, USA, pp. 65–71, 2016.

[36] N. Du, Y. Y. Liang, M. F. Balcan, L. Song. Influence function learning in information diffusion networks. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, pp. 2016–2024, 2014.

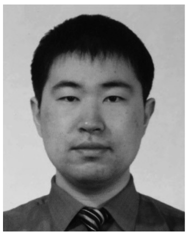
[37] S. Brin, L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998. DOI: 10.1016/S0169-7552(98)00110-X.

[38] R. S. Burt. *Structural Holes: The Social Structure of Competition*, Cambridge, Massachusetts, USA: Harvard University Press, pp. 51–64, 1992.

Kai Li received the B. Sc. and M. Sc. degrees in computer science from the YanShan University and Ji-Lin University, China in 2000 and 2003, respectively. He is now a Ph. D. degree candidate at School of Computer Science and Technology, University of Science and Technology of China (USTC), under the supervision of professor En-Hong Chen. He also visited the Institute of Computing Technology, Chinese Academy of Sciences, China, as a research assistant under the supervision of professor Xue-Qi Cheng from August 2017 to December 2019.

His research interests include social network analysis and human dynamics.

E-mail: marvin77@mail.ustc.edu.cn
ORCID iD: 0000-0002-6227-7158



Tong Xu received the Ph. D. degree in computer science from University of Science and Technology of China, China in 2016. He is currently working as an associate researcher of the Anhui Province Key Laboratory of Big Data analysis and Application, USTC. He has authored more than 40 journal and conference papers in the fields of social network and social media analysis, including *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Mobile Computing*, KDD, AAAI, ICDM, SDM, etc. He was a recipient of the ACM(Hefei) Doctoral Dissertation Award, 2016.

His research interests include social network analysis and data mining.

E-mail: tongxu@ustc.edu.cn

Shuai Feng received the M.Sc. degree in computer science from the Northeast University, China in 2014. He is currently an engineer of the IT center of Chinese National Audit Office.

His research interests includes complex network and audit technology.

E-mail: 18301638625@163.com

Li-Sheng Qiao received the M.Sc. degrees in electric power system & automation from Southwest Jiaotong University, China in 2009. He is now a Ph.D. degree candidate at School of Computer Science and Technology, University of Science and Technology of China, under the supervision of professor En-Hong Chen.

His research interests include deep learning and data mining.

E-mail: lsqiaoa@mail.ustc.edu.cn

Hua-Wei Shen received the B.Sc. degree in electronic information from the Xi'an Institute of Posts and Telecommunications, China in 2003, and the Ph.D. degree in information security from the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS), China in 2010. He is currently a professor in ICT-CAS. He has published more than 20 papers in prestigious journals and top international conferences, including *Physical Review E*, *Journal of Statistical Mechanics*, *Physica A*, WWW, CIKM, and IJCAI. He is a member of the Association of Innovation Promotion for Youth of CAS. He received the Top 100 Doctoral Thesis Award of CAS in 2011 and the Grand Scholarship of the President of CAS in 2010.

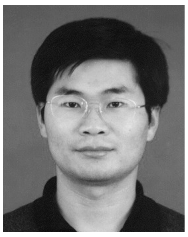
His research interests include network science, information recommendation, user behaviour analysis, machine learning, and social network.

E-mail: shenhuawei@ict.ac.cn

Tian-Yang Lv received the Ph.D. degree in computer science from Jilin University, China in 2007. He is currently a senior engineer of IT center of Chinese National Audit Office.

His research interests include complex networks and audit technology.

E-mail: raynor1979@163.com



Xue-Qi Cheng received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, China in 2006. He is a professor in ICT-CAS, and the director of the Research Center of Web Data Science & Engineering (WDSE), ICT-CAS. He has published more than 100 publications in prestigious journals and conferences, including the *IEEE Transactions on Information Theory*, *IEEE Transactions on Knowledge and Data Engineering*, *Journal of Statistics Mechanics: Theory and Experiment*, *Physical Review E.*, ACM SIGIR, WWW, ACM CIKM, WSDM, IJCAI, ICDM, He has won the Best Paper Award in CIKM (2011) and the Best Student Paper Award in SIGIR (2012). He is currently serving on the editorial board of *Journal of Computer Science and Technology*, *Journal of Computer*, etc. He received the China Youth Science and Technology Award, 2011, the Young Scientist Award of Chinese Academy of Sciences, 2010, CVIC Software Engineering Award, 2008, the second prize for the National Science and Technology Progress, 2004, etc. He is a member of the IEEE.

His research interests include network science, web search and data mining, big data processing and distributed computing architecture.

E-mail: cxq@ict.ac.cn

En-Hong Chen received the Ph.D. degree from University of Science and Technology of China, China in 1996. He is a professor and vice dean of School of Computer Science, USTC. He has published more than 150 papers in refereed conferences and journals, including *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Industrial Electronics*, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences including KDD, ICDM, and SDM. He received the Best Application Paper Award on KDD, 2008, the Best Research Paper Award on ICDM, 2011, and the Best of SDM, 2015. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is a senior member of the IEEE.

His research interests include data mining and machine learning, social network analysis, and recommend systems.

E-mail: cheneh@ustc.edu.cn (Corresponding author)

ORCID iD: 0000-0002-4835-4102