Physica A 565 (2021) 125584

Contents lists available at ScienceDirect

Physica A

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

Estimating user influence ranking in independent cascade model

Pei Li [a,*], Ke Liu [a], Keqin Li [b], Jianxun Liu [a], Dong Zhou [a]

[a] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
[b] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Article history:
Received 29 September 2020
Received in revised form 20 November 2020
Available online 28 November 2020

Keywords:
User influence ranking
Independent cascade model
Duplicate forwarding model
Social networks

ABSTRACT

Nowadays, hundreds of millions of people use social networks to express their opinions and communicate with their friends. It is of importance to model and estimate the user influence in social networks. Since most studies perform Monte Carlo simulation to evaluate the user influence in the independent cascade model, which leads to tremendous computational costs, we introduce a duplicate forwarding model to characterize the diffusion process in social networks, and analyze the user influences below and above the diffusion threshold theoretically. After getting the user influence ranking, we propose a Spearman-like correlation coefficient to measure the correlation between two rankings, and find the analysis results from the duplicate forwarding model achieve much better accuracy than the measurements degree, betweenness, k-core and PageRank in estimating the user influence ranking in the independent cascade model. This approach can provide insights in modeling and estimating the influences of social network users, and can be easily extended to estimate the influence ranking for different seed sets in the problem of influence maximization.

1. Introduction

With the rapid development of mobile communication technologies, many people are used to using mobile devices to connect to the Internet, and use social networks like Facebook, Twitter and Weibo to express their opinions and communicate with their friends. Therefore, a large amount of attention from academic and industrial societies has been paid to the studies on social networks [1–3]. Noting that information can be forwarded along the relationships in a social network, which may incur a chain reaction, researchers call this phenomenon word-of-mouth effect [4]. Some advertisers start to deploy advertisements in popular social networks, and hope to spread them to a large number of users, which is regarded as viral marketing [5,6]. An important issue here is to understand the diffusion dynamics and then rank users according to their influences [7–9]. This information will be useful for an advertiser to decide how to deploy an advertisement.

Many efforts have been devoted to the research of diffusion dynamics in the area of epidemiology, and researchers usually use the Susceptible–Infectious–Susceptible (SIS) model and Susceptible–Infectious–Removed (SIR) model [10,11] to investigate the diffusion dynamics. An interesting phenomenon here is the existence of epidemic threshold, above which the epidemic may spread and never terminate. There are some studies which consider the impact of network

* Corresponding author.
E-mail addresses: 8992077@qq.com (P. Li), 1069590903@qq.com (K. Liu), lik@newpaltz.edu (K. Li), ljx529@gmail.com (J. Liu), dongzhou1979@hotmail.com (D. Zhou).

https://doi.org/10.1016/j.physa.2020.125584
0378-4371/

structure on the epidemic threshold [12–15]. However, the diffusion process in social networks is very different from that in epidemiology, and cannot be characterized by these models accurately. Two famous models in the research of social networks are independent cascade model [16,17] and linear threshold model [18]. In the former one, an active node tries to influence its inactive neighbors with given probabilities, and these influence behaviors are independent. In the latter one, each node is assigned with a weight, and an inactive node will be influenced if the sum of its active neighbors' weights is larger than a given threshold. Based on these models, many studies focus on the problem of influence maximization in social networks [19–23].

A fundamental problem in influence maximization is to evaluate the function $\sigma(S)$, which is the expected number of activated users with S being the seed set. However, it has been proved that computing $\sigma(S)$ is #P-hard in both independent cascade model and linear threshold model [24,25]. To overcome the #P-hardness, Kempe et al. [19] use Monte Carlo simulation to evaluate $\sigma(S)$. Unfortunately, this approach needs to generate a large amount of samples to get a good estimation for $\sigma(S)$, which leads to tremendous computational costs. Leskovec et al. [26] propose an early termination heuristic to prune nodes with small influences at subsequent iterations to reduce the number of simulations. Goyal et al. [27] improve the work [26] by avoiding unnecessary simulations, but fail to achieve significant speedups [28]. Zhou et al. [29] use matrix analysis to get an upper bound of influence quickly, and then receive better performance compared with [26] and [27]. Although considerable efforts have been devoted to accelerate the simulation process of estimating $\sigma(S)$, significant computational costs are still required, which has been reported by [28] and [30].

In this paper, we introduce a duplicate forwarding model to characterize the diffusion process in social networks, where messages are forwarded in an all-or-none way and a user can forward multiple messages to a given neighbor. Although there are some differences between the independent cascade model and the duplicate forwarding model, we show that the user influence rankings in these two models are highly positively correlated, and an influential user in one model is very likely to be influential in the other one. Specifically, we adopt generating function [31] to analyze the user influence in the duplicate forwarding model theoretically, and estimate the user influence rankings below and above the diffusion threshold respectively. Besides, we choose 4 real-world networks, and conduct simulations to estimate the user influence ranking in the independent cascade model. We propose a Spearman-like correlation coefficient to measure the correlation between two rankings, and study the accuracies of using the results from the duplicate forwarding model, degree, betweenness [32], k-core [33] and PageRank [34] to estimate the user influence ranking in the independent cascade model. We find that the duplicate forwarding model achieves the best accuracy.

In our view, the analysis framework and results in this paper are of use to model and estimate the user influence in social networks, and can provide help in solving the influence maximization problem. Note that we can apply this approach to rank seed sets according to their influences, i.e. $\sigma(\cdot)$. For example, given seed sets $S_1, S_2, \ldots$, we can add a user s_i for S_i in the network, and add directed edges with influence probability 1 from s_i to v, where $v \in S_i$. Then we get a new network, and can estimate the user influence ranking for $s_1, s_2, \ldots$, which is actually the influence ranking for the seed sets $S_1, S_2, \ldots$ in the original network.

2. Models

In this section, we first introduce the network model. Then we describe the all-or-none forwarding mechanism which is adopted in many social networks, and introduce the duplicate forwarding model.

2.1. Network model

Nowadays, Facebook is arguably one of the most popular social networks. In Facebook, a relationship is established when a request for friendship is accepted by a user, which adds each other to their contact lists. If one user removes the other, the relationship is broken. That is to say, relationships in Facebook are symmetric.

In this paper, we consider symmetric relationships, and model a social network as an undirected graph, where nodes represent users and edges represent relationships between user pairs. We exclude isolated users, since they will never be involved in a diffusion process. Letting the user number be N, we arrange users and denote the ith user by user i, where $1 \leq i \leq N$. Then the network topology can be represented as an adjacency matrix $\mathbf{A}$, where for each $a_{i,j} \in \mathbf{A}$, $a_{i,j} = 1$ if users i and j are connected, and $a_{i,j} = 0$ if they are not connected.

Note that the adjacency matrix $\mathbf{A}$ is symmetric, since we take undirected networks into account. Actually, a social network with directed relationships can also be considered in this framework, and the corresponding adjacency matrix is asymmetric.

2.2. Duplicate forwarding model

In many social networks, messages are forwarded in an all-or-none way. For example, in Facebook, a user can update its status to broadcast things which happen in its daily life. These messages are pushed to the personal pages of all its neighbors, where messages are arranged in a reverse chronological order. After browsing (i.e., reading) the messages, its neighbors can use buttons such as share to forward them to their neighbors. This phenomenon is shown in Fig. 1. In Fig. 1(a), suppose user 1 generates a message (say M_1), which will be pushed to all its neighbors (i.e., users 2, 3, 4, 5).

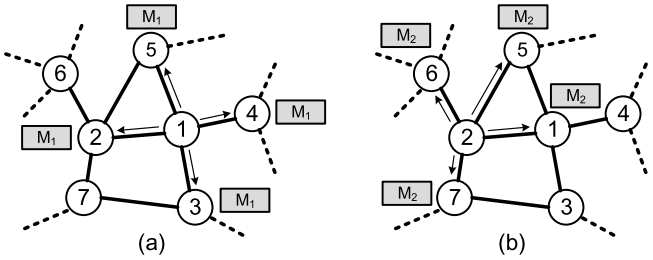

Fig. 1. An illustration for the all-or-none forwarding mechanism.

After receiving this message, user 2 can decide whether to react to it. If it chooses to do so, a message (say M_2) will be forwarded to all its neighbors (i.e., users 1, 5, 6, 7), which is depicted in Fig. 1(b). If user 2 chooses to do nothing, none of its neighbors will receive M_2. Besides, forwarded messages can be further forwarded, which incurs a chain reaction. This is the so-call word-of-mouth effect.

In this paper, we adopt the all-or-none forwarding mechanism in the duplicate forwarding model. Note that the content of M_2 may be different from that of M_1 in Fig. 1. To simplify the model, we ignore the heterogeneity of message contents, and let M_2 be a duplicate of M_1. Then we can focus on the number of messages which are received by users after a user generates a message.

We further let the probability that a user chooses to forward a received message be p. Actually, the probabilities of forwarding a given message will be different for different users, since they may have various personal interests. We assume homogeneous behaviors here to get a simple model, and can consider heterogeneous behaviors by adopting weighted social network models like [35,36] to extend the model.

Note that there are some differences between the independent cascade model and duplicate forwarding model, which are described in the following.

- In the independent cascade model, after getting activated, a user starts to influence its inactive neighbors. If this user fails to activate a neighbor, it cannot influence this neighbor again. That is to say, each edge will be involved at most once during a diffusion process. However, in the duplicate forwarding model, a user can forward multiple messages to a given neighbor during a diffusion process.
- In the independent cascade model, after getting activated, a user seeks to influence its inactive neighbors independently. However, in the duplicate forwarding model, messages will be forwarded in an all-or-none way. If a user chooses to forward a message, all its neighbors will receive it simultaneously.

Although the independent cascade model is a little different from the duplicate forwarding model, we propose a hypothesis (i.e., Hypothesis 1) here to characterize the relationships between the user influence rankings of these two models.

Hypothesis 1. The user influence ranking in the independent cascade model should be highly positively correlated with that in the duplicate forwarding model.

That is to say, if user i is more influential than user j in the independent cascade model, it is very likely that user i is also more influential than user j in the duplicate forwarding model. This hypothesis will be validated through experiments in the following.

3. User influence in the duplicate forwarding model

In the duplicate forwarding model, we take into account the mean number of messages which are received by users after a user (say user i) generates a message, and denote it by u_i. Intuitively, we know the value of u_i will grow and approach infinity with increasing p. So we denote the diffusion threshold by ρ, which is a critical value for p. If $p \geq \rho$, the diffusion process may never terminate, and infinite messages may be received by users.

In the following, we first analyze u_i and ρ theoretically, and get the user influence ranking in the duplicate forwarding model below the diffusion threshold. Since u_i makes no sense when the diffusion threshold is exceeded, we introduce and calculate q_i, which is the probability that infinite messages will be received after user i generates a message, to quantify the user influence when $p \geq \rho$, and then get the user influence ranking above the diffusion threshold.

3.1. User influence below diffusion threshold

Letting the probability of k messages are received after user i generates a message be $g_{i,k}$, we can get a generating function

$$G_i(x) = \sum_k g_{i,k} x^k. \tag{1}$$

Then we can obtain the user influence u_i by

$$u_i = \sum_k k g_{i,k} = G_i'(1). \tag{2}$$

According to the properties of generating function [37], we have

$$G_i(x) = \prod_{j \in \mathcal{N}_i} (1-p)x + pxG_j(x), \tag{3}$$

where $\mathcal{N}_i$ is the neighbor set of user i, the term $(1-p)x$ accounts for the situation that user j receives but does not forward the message, and the term $pxG_j(x)$ stands for the situation that user j receives and then forwards the message. Noting

$$G_j(1) = \sum_k g_{j,k} = 1, \tag{4}$$

we know

$$(1-p)x + pxG_j(x)\Big|_{x=1} = 1 \tag{5}$$

for any user j. Then from Eq. (3), we can get

$$\begin{aligned} G_i'(1) &= \sum_{j \in \mathcal{N}_i} \frac{\mathrm{d}\left[(1-p)x + pxG_j(x)\right]}{\mathrm{d}x} \prod_{j' \in \mathcal{N}_i \setminus j} (1-p)x + pxG_{j'}(x)\Big|_{x=1} \\ &= \sum_{j \in \mathcal{N}_i} (1-p) + pG_j(x) + pxG_j'(x)\Big|_{x=1} \\ &= \sum_{j \in \mathcal{N}_i} 1 + pG_j'(1) \\ &= |\mathcal{N}_i| + p \sum_{j \in \mathcal{N}_i} G_j'(1). \end{aligned} \tag{6}$$

We write Eq. (6) in matrix form, and have

$$\mathbf{G}'(1) = \mathbf{A1} + p\mathbf{A}\mathbf{G}'(1), \tag{7}$$

where $\mathbf{A}$ is the adjacency matrix, and

$$\begin{aligned} \mathbf{G}'(1) &= \left(G_1'(1), G_2'(1), \ldots, G_N'(1)\right)^\top, \\ \mathbf{1} &= (1, 1, \ldots, 1)^\top. \end{aligned}$$

So we get

$$\mathbf{G}'(1) = (\mathbf{I} - p\mathbf{A})^{-1}\mathbf{A1}, \tag{8}$$

from which we know the elements in $\mathbf{G}'(1)$ will grow and then approach infinity when the determinant of $\mathbf{I} - p\mathbf{A}$ arrives at the first 0. Among the eigenvalues of $\mathbf{A}$, let the largest one be $\lambda_{\mathbf{A}}$. Then the diffusion threshold ρ can be got by

$$\rho = \frac{1}{\lambda_{\mathbf{A}}}, \tag{9}$$

and the elements in $\mathbf{G}'(1)$ will be infinite if $p \geq \rho$.

Therefore, if $p < \rho$, we can calculate the user influence u_i from Eqs. (2) and (8) for each user, and then get the user influence ranking in the duplicate forwarding model accordingly.

3.2. User influence above diffusion threshold

Note that even if $p \geq \rho$, a diffusion process may be terminated due to the diffusion fluctuations. Here we introduce q_i, which is the probability that infinite messages will be received after user i generates a message, to quantify the user influence in the duplicate forwarding model when $p \geq \rho$, and try to calculate q_i in the following.

According to the properties of generating function [37], we have

$$G_i(1) = \sum_k g_{i,k} < 1 \tag{10}$$

if $p \geq \rho$, and $G_i(1)$ is actually the probability that finite messages will be received after user i generates a message. Therefore, we obtain

$$q_i = 1 - G_i(1), \tag{11}$$

Table 1
Properties of networks adopted in simulations.

Network	# of node	# of edges	Description
email-Eu-core	986	16,064	E-mail network
CollegeMsg	1893	13,835	Messages on a Facebook-like platform at UC-Irvine
soc-sign-bitcoin-alpha	3775	14,120	Bitcoin Alpha web of trust network
p2p-Gnutella08	6299	20,776	Gnutella peer to peer network from August 8 2002

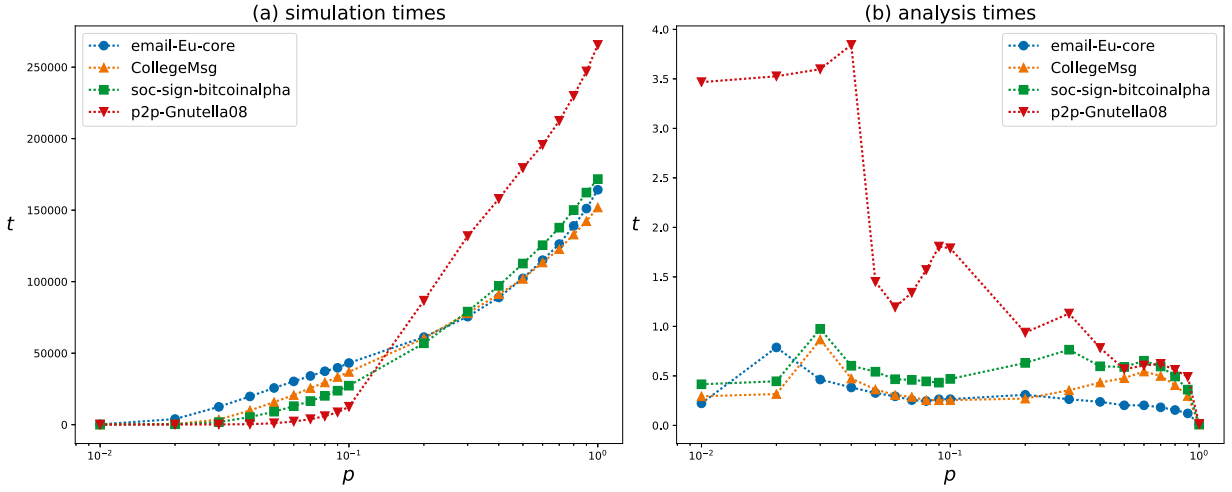

Fig. 2. Simulation times for $\tilde{u}_i$ and analysis times for u_i (or q_i) in different networks with varied p.

and get

$$1 - q_i = \prod_{j \in \mathcal{N}_i} \left(1 - p + p(1 - q_j)\right) \tag{12}$$

from Eq. (3). Then we rearrange Eq. (12), and have

$$q_i = 1 - \prod_{j \in \mathcal{N}_i} \left(1 - pq_j\right). \tag{13}$$

Note that we have an equation for each user from Eq. (13), and then get an equation system with N equations and N variables. Therefore, we can compute q_i by solving this equation system through iterative calculation, and then get the user influence ranking in the duplicate forwarding model accordingly.

4. Verification

In the independent cascade model, we define the user influence $\tilde{u}_i$ as the average number of users which are activated (i.e., receive messages) during a diffusion process, which is caused by user i generating a message.

Noting that there is no efficient approach to calculate the exact value of $\tilde{u}_i$ at this moment, we first conduct simulations to estimate it in this section. Then we introduce a Spearman-like correlation coefficient to measure the correlation between two user influence rankings. Based on this correlation coefficient, we study the correlation between the user influence rankings of $\tilde{u}_i$ and u_i (or q_i), as well as the correlations between the user influence rankings of $\tilde{u}_i$ and other measurements such as degree, betweenness, k-core and PageRank.

4.1. Simulations for $\tilde{u}_i$ in the independent cascade model

We select 4 networks from http://snap.stanford.edu/data/ to conduct simulations. Note that if there are multiple connected components in a network, we can take into account these connected components separately. To simplify the simulations, we choose the largest connected component for each network. Besides, we consider undirected edges and delete edges connecting a node to itself, and then get 4 networks with undirected edges and no self-loops. Some properties of these networks are listed in Table 1.

In each simulation, all users are inactive at the beginning, and a user is selected to be the first active one and start to influence its inactive neighbors with probability p independently, which may activate a chain reaction. Note that if a

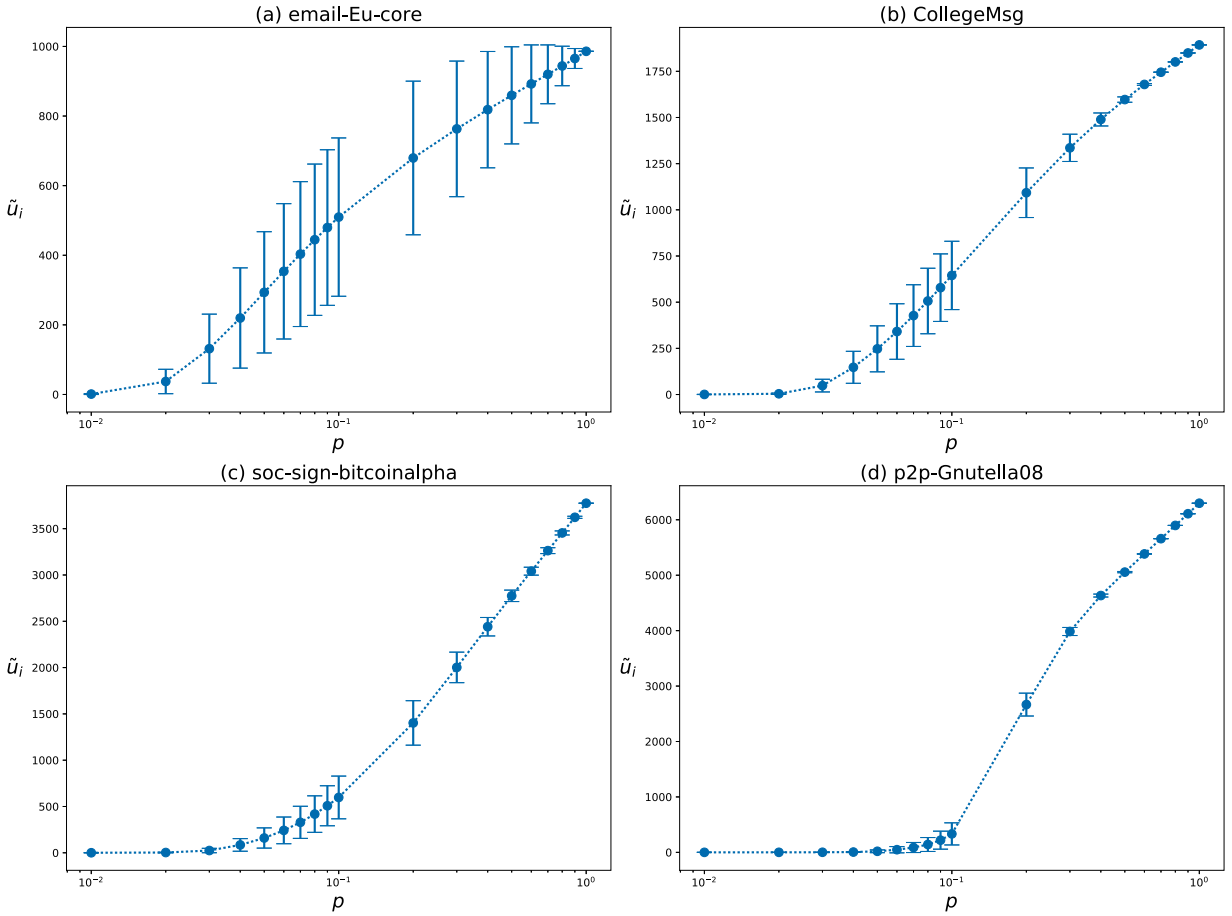

Fig. 3. Simulation results for the mean and standard deviation of $\tilde{u}_i$'s for different networks with varied p in the independent cascade model.

Table 2
Diffusion threshold ρ in the duplicate forwarding model.

Network	$\rho = 1/\lambda_{\mathbf{XA}}$
email-Eu-core	0.013
CollegeMsg	0.021
soc-sign-bitcoin-alpha	0.021
p2p-Gnutella08	0.035

user fails to activate a neighbor, it cannot influence this neighbor again. Furthermore, we repeat each simulation 100,000 times, and calculate the average number of activated users $\tilde{u}_i$ to estimate the user influence for user i.

Since the simulations are quite time consuming, we choose all users for the network email-Eu-core, and 1000 users with the largest degrees for the networks CollegeMsg, soc-sign-bitcoin-alpha and p2p-Gnutella08 to estimate $\tilde{u}_i$. All simulations are carried out on a machine with an Intel(R) Core(TM) I7-8700 CPU (3.20 GHz, 12 threads) and 32 GB main memory. We also adopt the multi-threading technology to accelerate the simulations. The times spent in estimating $\tilde{u}_i$ through simulations for different networks with varied p are plotted in Fig. 2(a), from which we know the simulation times increase with p for all networks. Besides, we also depict the times spent in calculating u_i (or q_i) through theoretical analysis for different networks with varied p in Fig. 2(b), and note that the analysis times are much less than the simulation ones, especially when p is large. For example, the simulation time for $\tilde{u}_i$ in the network p2p-Gnutella08 with $p = 0.5$ is 179, 497.45 s, but the corresponding analysis time for q_i is only 0.57 s, which is about 300,000 times less than the simulation one. So we can conclude that the user influence in the duplicate forwarding model (i.e., u_i and q_i) can be analyzed efficiently.

To study the gaps between different $\tilde{u}_i$'s, we calculate the mean and standard deviation for $\tilde{u}_i$'s in each network. We also vary the value of p to show the impact of p on the mean and standard deviation of $\tilde{u}_i$'s. The simulation results for the mean and standard deviation of $\tilde{u}_i$'s with varied p are plotted in Fig. 3, from which we know the mean of $\tilde{u}_i$'s increases with p, and the standard deviation of $\tilde{u}_i$'s approaches 0 when p is large for all networks. The reason is when p is large,

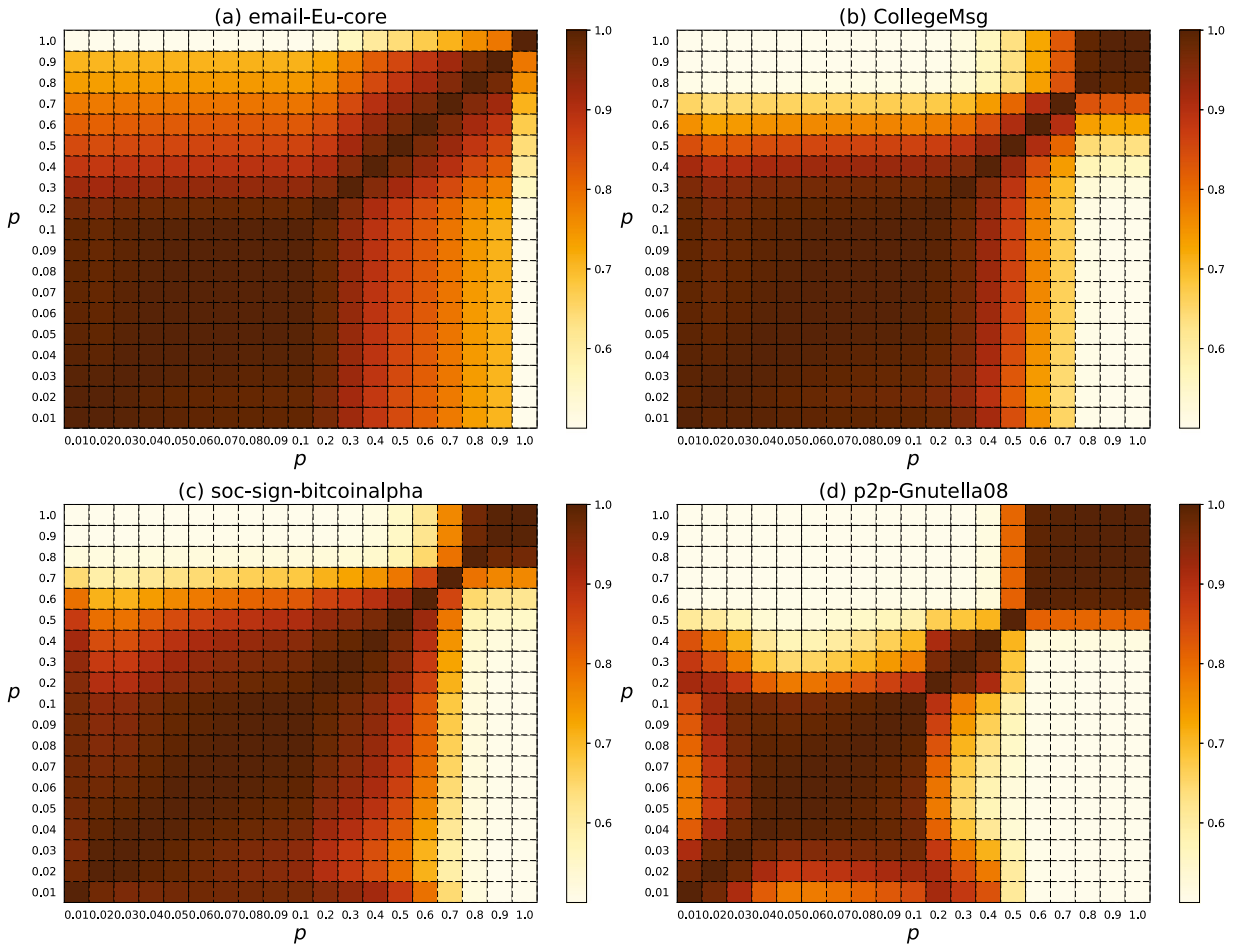

Fig. 4. The Spearman-like correlation coefficients between user influence rankings of $\hat{u}_i$ with varied p in the independent cascade model, where $0.01 \leq p \leq 1$.

any user may be able to influence a large fraction of users, and the gaps between different $\tilde{u}_i$'s are small. Therefore, we can claim that if the influence probability in a network is high enough, the user influences for different users will be close in the independent cascade model, and we can choose any user to incur a wide spread. Besides, since the values of $\tilde{u}_i$'s are small at $p = 0.01$, the standard deviation of $\tilde{u}_i$'s is almost invisible. Actually, the value of the standard deviation at $p = 0.01$ is much larger than that of the mean.

4.2. Correlation definition

In this paper, we introduce a Spearman-like correlation coefficient to measure the correlation between two user influence rankings. Let X_i and Y_i be the rank values for user i in two user influence rankings, and $d_i = X_i - Y_i$. Here we assign users of an identical user influence with the same rank value, which is equal to the mean of their positions in the ascending order. This is actually equivalent to averaging out the rank value over all possible permutations.

The formula for the Spearman-like correlation coefficient is given by

$$r = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}, \tag{14}$$

where N is the user number. Note that the formula for the Spearman correlation coefficient is

$$r_s = \frac{\sum_i (X_i - \sum_i X_i/N)(Y_i - \sum_i Y_i/N)}{\sqrt{\sum_i (X_i - \sum_i X_i/N)^2 \sum_i (Y_i - \sum_i Y_i/N)^2}}, \tag{15}$$

and Eq. (14) is actually the same as Eq. (15) if all rank values are distinct integers.

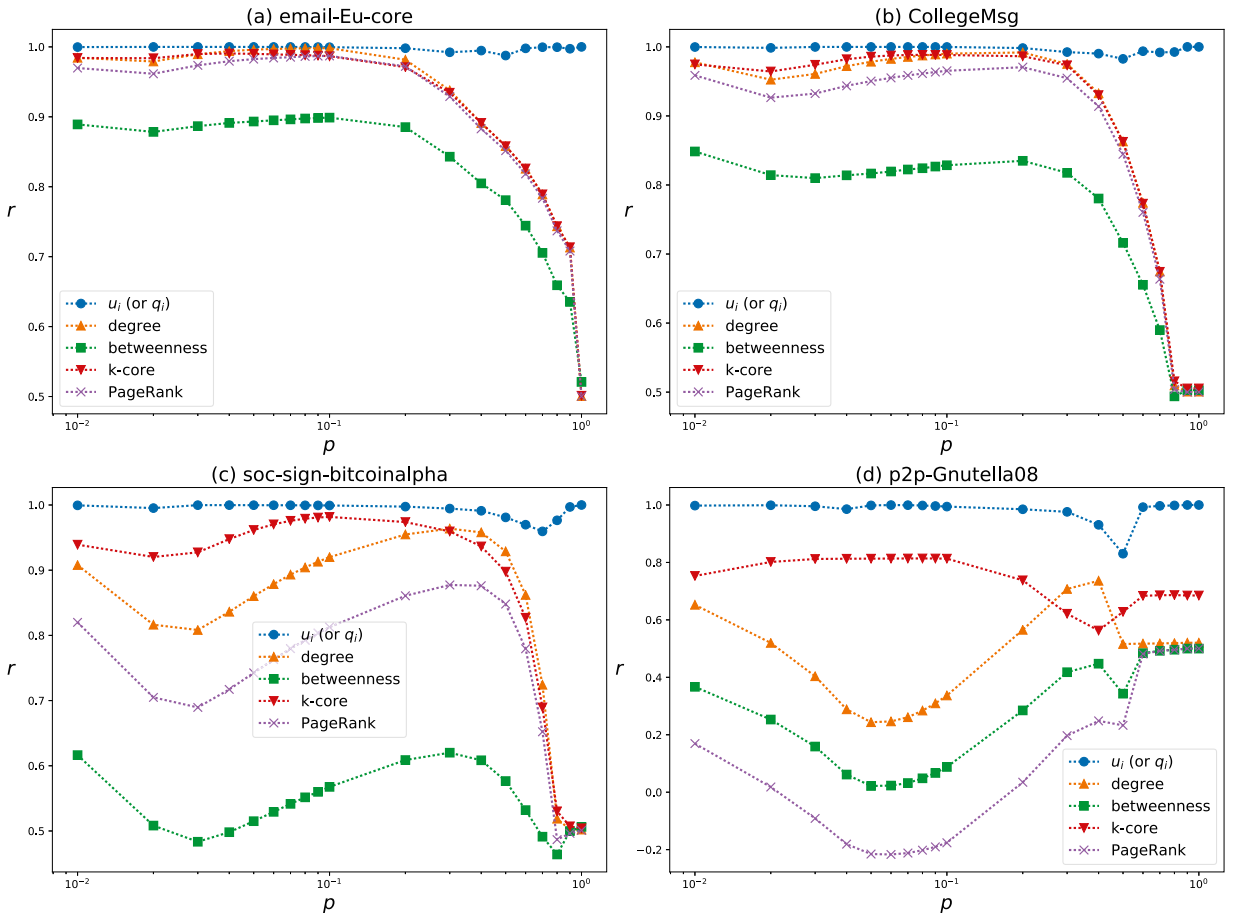

Fig. 5. The Spearman-like correlation coefficient between the user influence rankings of $\hat{u}_i$ and u_i (or q_i) with varied p, as well as the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and other measurements such as degree, betweenness, k-core and PageRank.

However, if all users have an identical user influence, the numerator and denominator of Eq. (15) will be 0, and then r_s will make no sense. Unfortunately, we find users will indeed have an identical user influence if p is large enough. Therefore, we cannot use the Spearman correlation coefficient, i.e. Eq. (15), to measure the correlation between two user influence rankings. In this paper, we adopt Eq. (14), and call r the Spearman-like correlation coefficient. Besides, we know that $r \leq 1$, and the value of r should be close to 1 if two user influence rankings are highly positively correlated.

4.3. Correlations between user influence rankings

In the independent cascade model, to reduce the impact of simulation fluctuations on the value of $\tilde{u}_i$, we normalize $\tilde{u}_i$ by letting

$$\hat{u}_i = \frac{\tilde{u}_i}{\max_j \tilde{u}_j}, \tag{16}$$

and keep three decimal places. Then we rank users according to $\hat{u}_i$. We compute the Spearman-like correlation coefficient between user influence rankings of $\hat{u}_i$ with varied p, and plot the results in Fig. 4. We find that the user influence ranking varies dramatically with p in the independent cascade model.

To get the user influence ranking in the duplicate forwarding model, we first calculate the diffusion threshold ρ for each network from Eq. (9), which is listed in Table 2. For $p < \rho$, we compute the user influence u_i from Eqs. (2) and (8), and get the user influence ranking for each network. For $p \geq \rho$, we calculate q_i, which is the probability that infinite messages will be received after user i generates a message, from Eq. (13), and get the corresponding user influence ranking. Note that we do the same operations of normalization and rounding for u_i and q_i as those for $\tilde{u}_i$.

For comparison, we adopt different measurements for user influence such as degree, betweenness, k-core and PageRank here. We compute the values of these measurements for each user, and do the same operations of normalization and

Table 3
P-values for the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and u_i (or q_i) for different networks with varied p.

email-Eu-core		CollegeMsg		soc-sign-bitcoin-alpha		p2p-Gnutella08	
p	p-value	p	p-value	p	p-value	p	p-value
0.01	0	0.01	0	0.01	0	0.01	0
0.02	0	0.02	0	0.02	0	0.02	0
0.03	0	0.03	0	0.03	0	0.03	0
0.04	0	0.04	0	0.04	0	0.04	0
0.05	0	0.05	0	0.05	0	0.05	0
0.06	0	0.06	0	0.06	0	0.06	0
0.07	0	0.07	0	0.07	0	0.07	0
0.08	0	0.08	0	0.08	0	0.08	0
0.09	0	0.09	0	0.09	0	0.09	0
0.1	0	0.1	0	0.1	0	0.1	0
0.2	0	0.2	0	0.2	0	0.2	0
0.3	0	0.3	0	0.3	0	0.3	0
0.4	0	0.4	0	0.4	0	0.4	0
0.5	0	0.5	0	0.5	0	0.5	1.08×10^{-255}
0.6	0	0.6	0	0.6	0	0.6	0
0.7	0	0.7	0	0.7	0	0.7	0
0.8	0	0.8	0	0.8	0	0.8	0
0.9	0	0.9	0	0.9	0	0.9	0
1	0	1	0	1	0	1	0

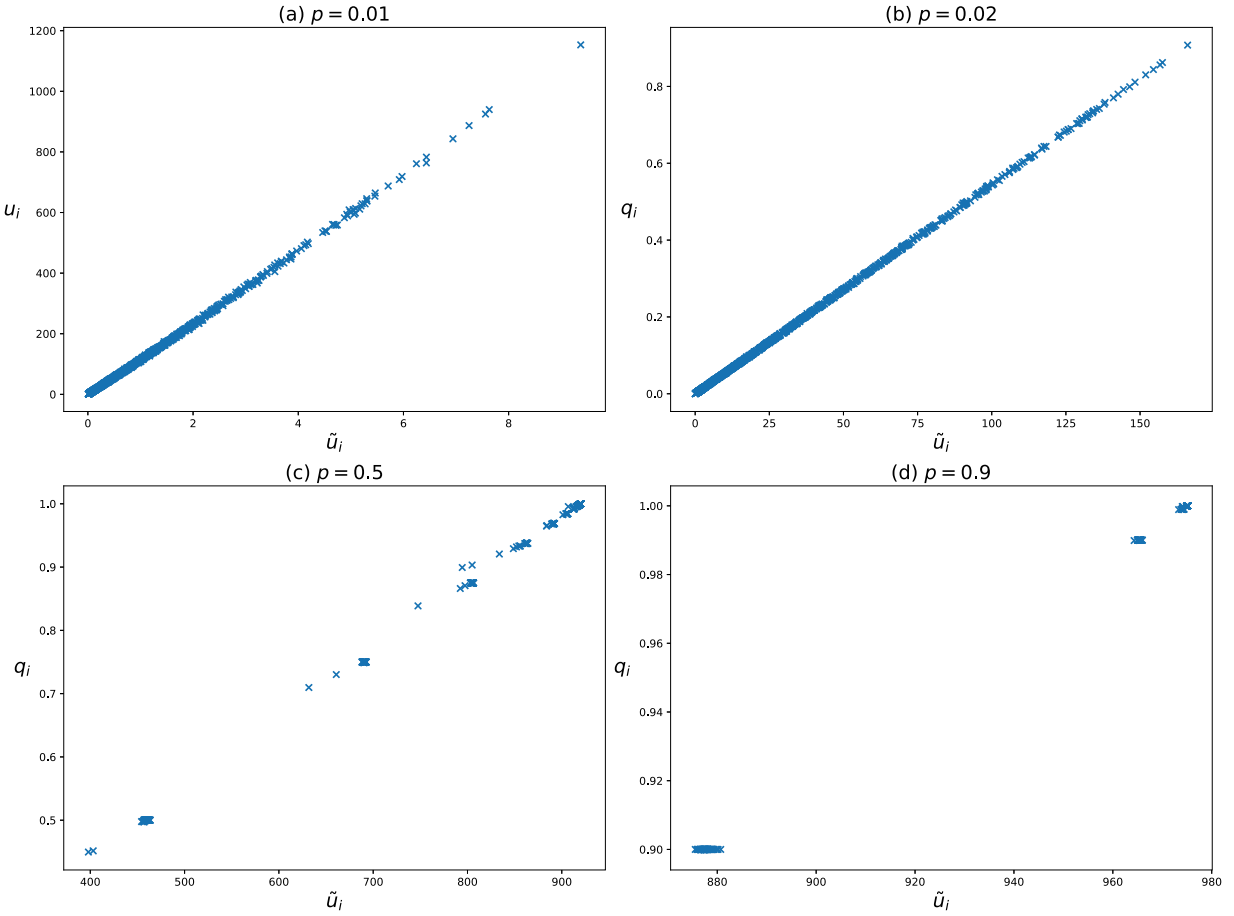

Fig. 6. Comparison of $\tilde{u}_i$ and u_i (or q_i) for users in the network email-Eu-core with different p.

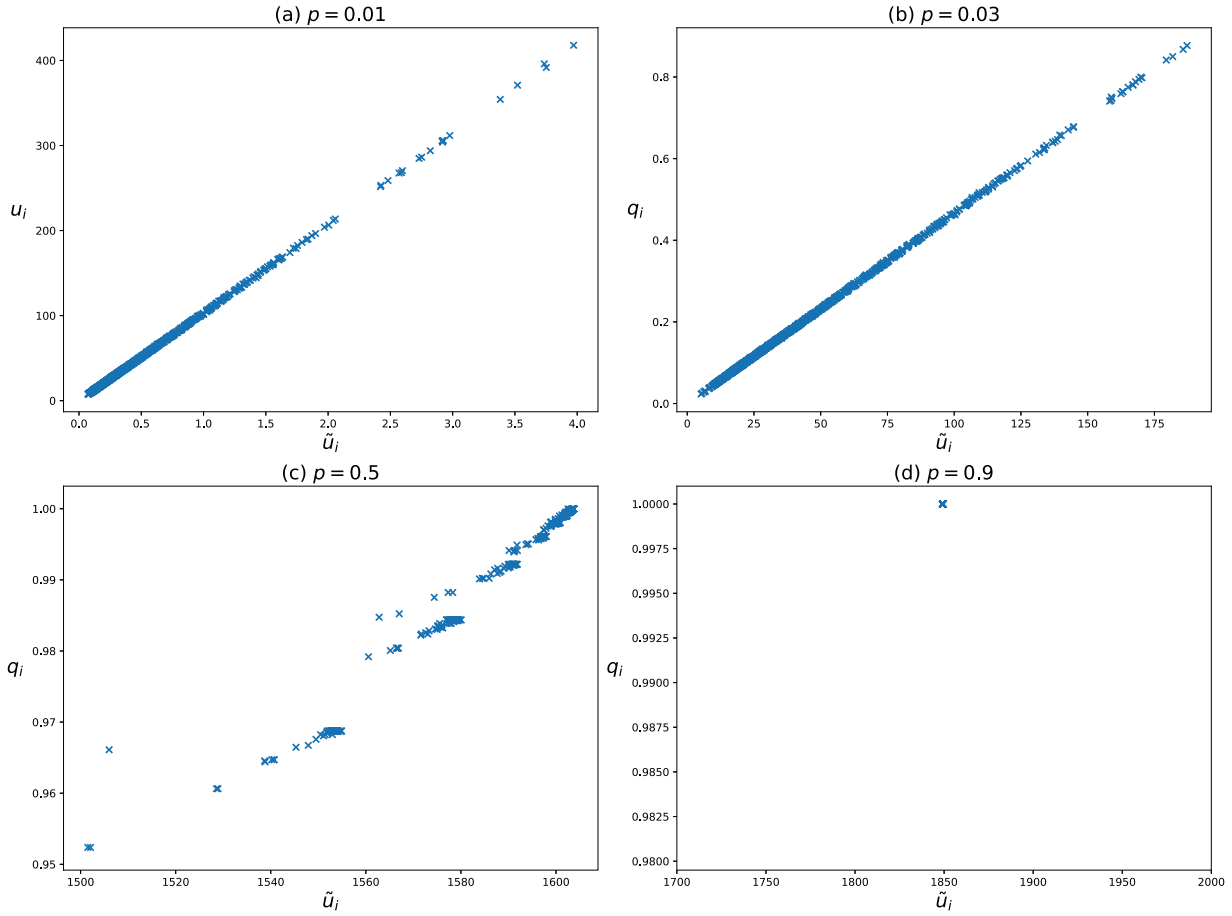

Fig. 7. Comparison of $\tilde{u}_i$ and u_i (or q_i) for users in the network CollegeMsg with different p.

rounding for these values as those for $\tilde{u}_i$. Then we get the user influence rankings accordingly. Note that these user influence rankings do not change with p.

Then we calculate the Spearman-like correlation coefficient between the user influence rankings of $\hat{u}_i$ and u_i (or q_i) from Eq. (14), as well as the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and other measurements such as degree, betweenness, k-core and PageRank. The results are depicted in Fig. 5.

We observe that the Spearman-like correlation coefficient between the user influence rankings of $\hat{u}_i$ and u_i (or q_i), which is labeled will "u_i (or q_i)" in Fig. 5, is very close to 1. That is to say, these two user influence rankings are highly positively correlated, and we can use the user influence ranking of u_i (or q_i) to estimate that of $\hat{u}_i$. We also note that there are some fluctuations around $p = 0.7$ in the network soc-sign-bitcoin-alpha, and around $p = 0.5$ in the network p2p-Gnutella08. There may be some properties of these networks which lead to these fluctuations in the independent cascade model, and we will consider them in our future work. So we can conclude that the user influence in the duplicate forwarding model can be used as a good measurement to estimate the user influence ranking in the independent cascade model.

Besides, we observe that the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and other measurements, which are labeled will "degree", "betweenness", "k-core" and "PageRank" in Fig. 5, will approach 0.5 if p is close to 1 in these networks. That is because if p is close to 1, all users may have identical $\hat{u}_i$ in the independent cascade model, and then have identical rank value in the user influence ranking. However, the rank values for the measurements degree, betweenness, k-core and PageRank may be distinct integers. For example, we take any measurement (say betweenness) into account, and let the user number $N = 2n + 1$. We know the rank values for betweenness may be distinct integers, and the range is $[1, 2n + 1]$. Since all users may have an identical rank value of n for $\hat{u}_i$ if p is close to 1, the Spearman-like correlation coefficient should be

$$r = 1 - \frac{6\sum_i d_i^2}{N(N^2 - 1)}$$

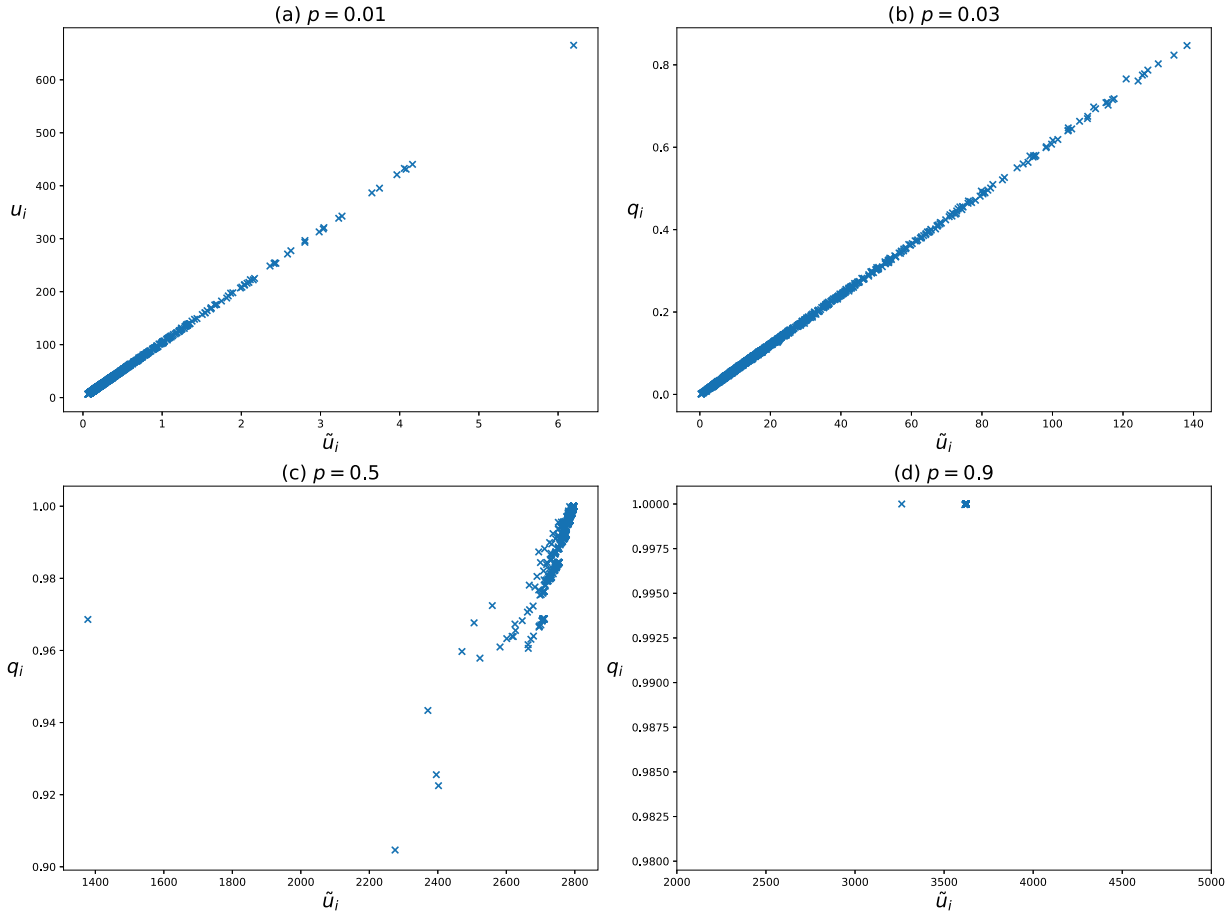

Fig. 8. Comparison of $\tilde{u}_i$ and u_i (or q_i) for users in the network soc-sign-bitcoin-alpha with different p.

$$\begin{aligned} &\approx 1 - \frac{6 \times 2 \sum_{1 \le k \le n} k^2}{(2n+1)\left((2n+1)^2 - 1\right)} \\ &= 1 - \frac{6 \times 2 \times n(n+1)(2n+1)/6}{4n(n+1)(2n+1)} \\ &= \frac{1}{2}. \end{aligned} \tag{17}$$

We can get the same result if we let $N = 2n$. This is the reason why the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and other measurements will approach 0.5 if p is close to 1.

Actually, PageRank has been widely used to measure the importance of website pages, which are connected by directed hyperlinks. However, from Fig. 5 we find that PageRank achieves poor performance to estimate the user influence ranking in the independent cascade model. This phenomenon may be caused by the reason that the diffusion dynamics in the independent cascade model are different from the browsing behavior dynamics in the Internet, and another potential reason may be the networks we consider here are symmetric. It is interesting to study this phenomenon, and will be included in our future work.

Then we calculate the corresponding p-values for the Spearman-like correlation coefficients between the user influence rankings of $\hat{u}_i$ and u_i (or q_i) for different networks with varied p, and list the results in Table 3, from which we know the correlations between these two user influence rankings are strongly statistically significant.

Finally, we compare the values of $\tilde{u}_i$ and u_i (or q_i) for users with different p. For each network in Table 1, we let $p = 0.01, 0.5, 0.9$ respectively, and also choose a value of p which is slightly larger than the diffusion threshold in Table 2. The results are depicted in Figs. 6–9. We observe that $\tilde{u}_i$ seems to be linearly correlated with u_i (or q_i) when p is small. That is to say, after getting the values of $\tilde{u}_i$ for some users by simulations, we can use linear fitting to estimate the influences for the rest users. We believe this operation can significantly reduce the amount of samples in estimating $\sigma(S)$, and then provide help in solving the influence maximization problem. Besides, although this linear correlation does not hold for

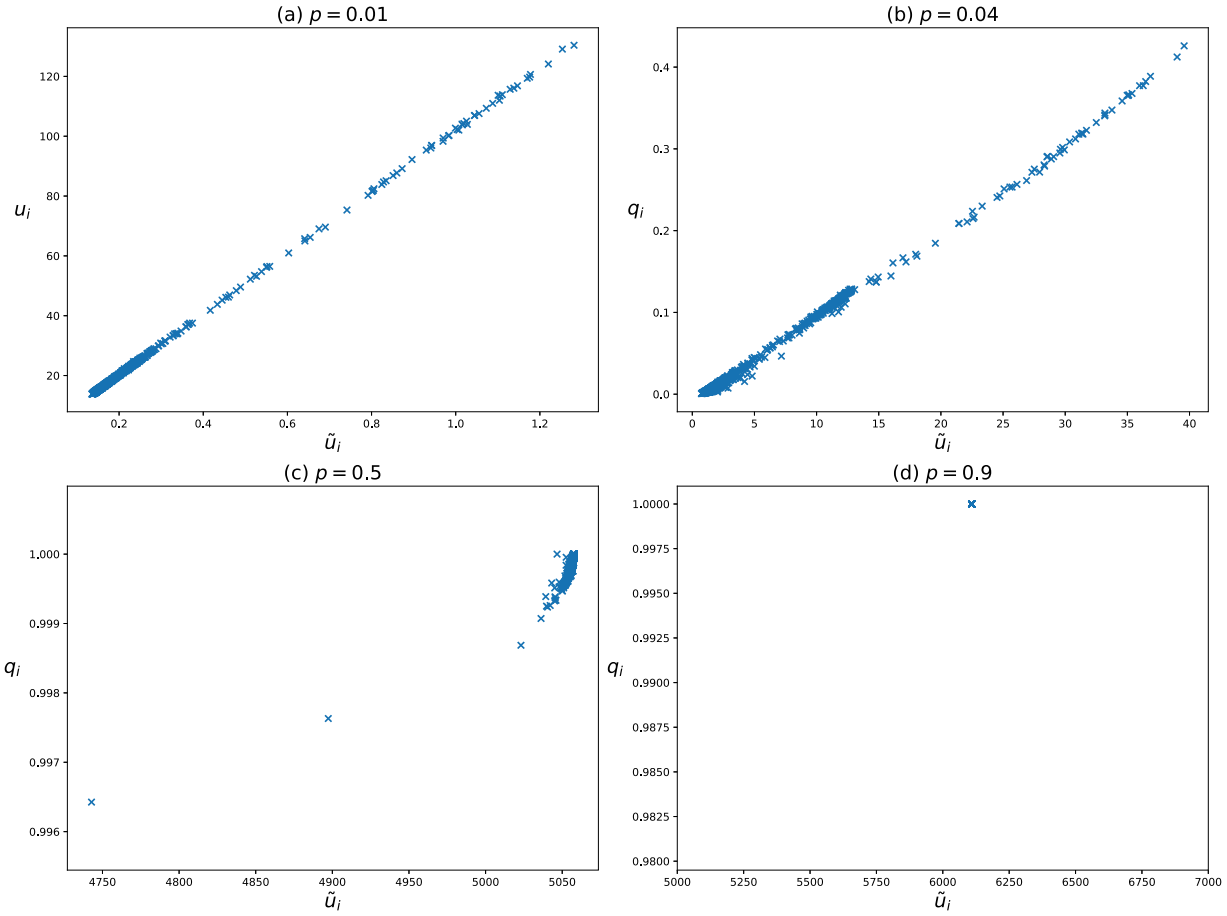

Fig. 9. Comparison of $\tilde{u}_i$ and u_i (or q_i) for users in the network p2p-Gnutella08 with different p.

$p = 0.5$ (especially for the networks soc-sign-bitcoin-alpha and p2p-Gnutella08), the values of $\tilde{u}_i$ and q_i are still highly positively correlated for most users. When $p = 0.9$, most users get identical $\tilde{u}_i$ (especially for the networks CollegeMsg, soc-sign-bitcoin-alpha and p2p-Gnutella08), which can be predicted through the analysis results for q_i from our approach.

5. Conclusion and discussion

In this paper, we propose a duplicate forwarding model to characterize the diffusion process in social networks, and analyze the user influences below and above the diffusion threshold theoretically. Through extensive simulations, we find that the analysis results from the duplicate forwarding model achieve much better accuracy than the measurements degree, betweenness, k-core and PageRank in estimating the user influence ranking in the independent cascade model. However, these inaccuracies of degree, betweenness, k-core and PageRank in estimating the user influence ranking may be caused by the reason that these measurements do not consider the diffusion dynamics in the independent cascade model. Actually, they have been introduced to characterize different aspects of node importance in a network, and have been successfully used in different areas.

In this paper, we assume symmetric relationships and homogeneous user behaviors to simplify the models, which are unrealistic in reality. In our future work, we will consider asymmetric relationships (e.g., relationships in Twitter and Weibo) and the situation that different users may forward messages with different probabilities. That is to say, a weighted directed network will be adopted to extend the network model. Then we can study the relationships between the user influence rankings of the independent cascade model and the duplicate forwarding model in a more realistic environment. Note that the diffusion dynamics in other diffusion models such as SIS model, SIR model and linear threshold model also need to be studied. What are the differences between the diffusion dynamics of the independent cascade model and these models? Can the duplicate forwarding model be modified to analyze the diffusion processes in these models? We will consider these problems in the future. Besides, in this paper we choose 4 moderate-size real-world networks in the simulations, since it takes too much time to repeat each simulation 100,000 times, especially when the parameter p is

large. Actually, the user influence in the duplicate forwarding model can be analyzed efficiently, even for a network with millions of users. We will try to accelerate the simulations, and verify the analysis results in larger networks. Finally, we plan to adopt the proposed approach in the influence maximization problem, and verify its efficiency and accuracy.

CRediT authorship contribution statement

Pei Li: Conceptualization, Methodology, Writing - original draft. **Ke Liu:** Software, Data curation. **Keqin Li:** Supervision, Validation. **Jianxun Liu:** Visualization, Investigation. **Dong Zhou:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Scientific Research Fund of Hunan Provincial Education Department, China (No. 18B199), and the National Natural Science Foundation of China (Nos. 61872139, 61876062).

References

[1] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? In: International Conference on World Wide Web, 2010, pp. 591–600.
[2] W. Yao, P. Jiao, W. Wang, Y. Sun, Understanding human reposting patterns on Sina Weibo from a global perspective, Physica A 518 (2019) 374–383.
[3] S. Talukder, B. Carbunar, A study of friend abuse perception in Facebook, ACM Trans. Soc. Comput. 3 (4) (2020) 17.
[4] J. Arndt, Role of product-related conversations in the diffusion of a new product, J. Mark. Res. 4 (3) (1967) 291–295.
[5] J. Leskovec, L.A. Adamic, B.A. Huberman, The dynamics of viral marketing, ACM Trans. Web 1 (1) (2007) 5.
[6] Q. Wang, F. Miao, G.K. Tayi, E. Xie, What makes online content viral? The contingent effects of hub users versus non-hub users on social media platforms, J. Acad. Mark. Sci. 47 (6) (2019) 1005–1026.
[7] Y. Li, B.Q. Zhao, J.C.S. Lui, On modeling product advertisement in large-scale online social networks, IEEE/ACM Trans. Netw. 20 (5) (2012) 1412–1425.
[8] P. Li, Y. Sun, Y. Chen, Z. Tian, Estimating user influence in online social networks subject to information overload, Internat. J. Modern Phys. B 28 (3) (2014) 1450004.
[9] P. Li, H. Nie, F. Yin, et al., Modeling and estimating user influence in social networks, IEEE Access 8 (2020) 21943–21952.
[10] N.T.J. Bailey, The Mathematical Theory of Infectious Diseases and Its Applications, Griffin, London, 1975.
[11] H.W. Hethcote, The mathematics of infectious diseases, SIAM Rev. 42 (4) (2000) 599–653.
[12] C. Moore, M.E.J. Newman, Epidemics and percolation in small-world networks, Phys. Rev. E 61 (5) (2000) 5678–5682.
[13] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (14) (2001) 3200–3203.
[14] R. Parshani, S. Carmi, S. Havlin, Epidemic threshold for the susceptible-infectious-susceptible model on random networks, Phys. Rev. Lett. 104 (25) (2010) 258701.
[15] G.O. Agaba, Y.N. Kyrychko, K.B. Blyuss, Time-delayed SIS epidemic model with population awareness, Ecol. Complex. 31 (2017) 50–56.
[16] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (3) (2001) 211–223.
[17] W. Yang, L. Brenner, A. Giua, Influence maximization in independent cascade networks based on activation probability computation, IEEE Access 7 (2019) 13745–13757.
[18] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. 83 (6) (1978) 1420–1443.
[19] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 137–146.
[20] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 199–208.
[21] B. Nettasinghe, V. Krishnamurthy, Influence maximization over Markovian graphs: A stochastic optimization approach, IEEE Trans. Signal Inf. Process. Netw. 5 (1) (2019) 1–14.
[22] J. Tang, R. Zhang, Y. Yao, et al., An adaptive discrete particle swarm optimization for influence maximization based on network community structure, Internat. J. Modern Phys. C 30 (6) (2019) 1950050.
[23] X. Li, J.D. Smith, T.N. Dinh, M.T. Thai, Tiptop: (almost) exact solutions for influence maximization in billion-scale networks, IEEE/ACM Trans. Netw. 27 (2) (2019) 649–661.
[24] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 1029–1038.
[25] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: IEEE International Conference on Data Mining, 2010, pp. 14–17.
[26] J. Leskovec, A. Krause, C. Guestrin, et al., Cost-effective outbreak detection in networks, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 420–429.
[27] A. Goyal, W. Lu, L.V. Lakshmanan, Celf++: Optimizing the greedy algorithm for influence maximization in social networks, in: International Conference Companion on World Wide Web, 2011, pp. 47–48.
[28] A. Arora, S. Galhotra, S. Ranu, Debunking the myths of influence maximization: An in-depth benchmarking study, in: ACM International Conference on Management of Data, 2017, pp. 651–666.
[29] C. Zhou, P. Zhang, W. Zang, L. Guo, On the upper bounds of spread for greedy algorithms in social network influence maximization, IEEE Trans. Knowl. Data Eng. 27 (10) (2015) 2770–2783.

[30] Y. Tang, X. Xiao, Y. Shi, Influence maximization: Nearoptimal time complexity meets practical efficiency, in: ACM International Conference on Management of Data, 2014, pp. 75–86.
[31] H.S. Wilf, Generatingfunctionology, second ed., Academic Press, London, 1994.
[32] S. Dolev, Y. Elovici, R. Puzis, Routing betweenness centrality, J. ACM 57 (4) (2010) 25.
[33] M. Kitsak, L.K. Gallos, S. Havlin, et al., Identification of influential spreaders in complex networks, Nature Phys. 6 (11) (2010) 888–893.
[34] N. Perra, S. Fortunato, Spectral centrality measures in complex networks, Phys. Rev. E 78 (3) (2008) 036107.
[35] Y. Murase, J. Török, H.H. Jo, et al., Multilayer weighted social network model, Phys. Rev. E 90 (5) (2014) 052810.
[36] P. Li, J. Yu, J. Liu, et al., Generating weighted social networks using multigraph, Physica A 539 (2020) 122894.
[37] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E 64 (2) (2001) 026118.