Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | JUNE 26 2020

Identifying highly influential nodes in multilayer networks based on global propagation

Xin Li ; Xue Zhang; Chengli Zhao ; Dongyun Yi; Guochen Li

Chaos 30, 061107 (2020)
https://doi.org/10.1063/5.0005602

Identifying highly influential nodes in multilayer networks based on global propagation

Cite as: Chaos **30**, 061107 (2020); doi: 10.1063/5.0005602
Submitted: 25 February 2020 · Accepted: 2 June 2020 ·
Published Online: 26 June 2020

Xin Li,[1,a)] Xue Zhang,[2] Chengli Zhao,[1,b)] Dongyun Yi,[1] and Guochen Li[3]

AFFILIATIONS

[1]Department of Systems Science, College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, Hunan 410073, China
[2]Department of Mathematics, College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, Hunan 410073, China
[3]Department of Biology and Chemistry, College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, Hunan 410073, China

[b)]**Author to whom correspondence should be addressed:** chenglizhao@nudt.edu.cn
[a)]**Electronic mail:** 287998116@qq.com

ABSTRACT

Based on percolation theory and the independent cascade model, this paper considers the selection of the optimal propagation source when the propagation probability is greater than the percolation threshold. First, based on the percolation characteristics of real networks, this paper presents an iterative algorithm of linear complexity to solve the probability of the propagation source transmitting information to the network's giant component, that is, the global propagation probability. Compared with the previous multiple local simulation algorithm, this algorithm eliminates random errors and significantly reduces the operation time. A sufficient and necessary condition is provided, and it is proved that the final propagation range of the propagation source obeys the bimodal distribution. Based on this sufficient and necessary condition, we extend the efficient iterative algorithm proposed in this article to multi-layer networks and find that for two-layer networks, the final propagation range of the propagation source follows a four-peak distribution. Through iterations and calculations, the probability of each peak and the number of nodes included can be directly obtained, and the propagation expectations of the nodes in the multi-layer network can then be calculated, which can result in a better ranking of the propagation influence of the nodes. In addition, to maximize the influence of multi-propagation sources, this paper also presents a de-overlapping method, which has evident advantages over traditional methods.

In many practical works, the identification and measurement of the influence of nodes in complex networks are of great significance. At present, there are many indicators that describe the importance of nodes, but they all have certain limitations in expressing the influence of node propagation. In particular, in the multi-layer network, there is little research on the selection of good propagation sources. In this paper, we use the global propagation probability to describe the propagation influence of the nodes and give an efficient iterative calculation method to calculate it. Furthermore, we extend this method to multi-layer networks and find that the final propagation range of the propagation source follows a multi-peak distribution. By studying this distribution, we can finally estimate the expected value of the final propagation range of the propagation source and select excellent propagation sources, and these source nodes will affect more nodes with greater probability.

I. INTRODUCTION

Before introducing the work of this paper, we first review some classic methods used for maximizing the influence of nodes in the network. In terms of the single-source influence-ranking problem, the simplest and most effective method is the use of a locality index, such as degree centrality,[1] which reflects the number of neighbors of a node—for example, an influencer with a large number of followers in Weibo has a high spread influence. Chen *et al.*[2] presented a local

centrality index that took into consideration the number of nodes' second-order neighbors. These two indicators can only reflect the local topology of the nodes. Furthermore, some global indicators with low algorithmic complexity have been presented, such as K-shell[3] and eigenvector centrality,[4] and the algorithmic complexity of calculating the index values of all the nodes in the network is O(n). Finally, there exist some indicators with high algorithmic complexity, such as the betweenness centrality[5] and closeness centrality.[6] These indicators are not applicable in large networks. Hu *et al.*[7] proposed a method for describing the spread influence based on the global propagation probability of the nodes and used multiple local simulation methods to solve the index value, which has high practical value. The multi-source influence-maximization problem was first proposed by Domingos and Richardson,[8] and it was later proved by Kempe *et al.*[9] that the problem is an NP-hard problem. One of the simplest selection strategies comprises selecting the nodes with the largest index value as the propagation source, but the various influencers selected using this algorithm may have a large influence overlap. Therefore, Leskovec *et al.*[10] used a greedy algorithm to seek the initial propagation set, but the complexity of this algorithm makes it unsuitable for large networks. Chen *et al.*[11] proposed two improved greedy algorithms, the new greedy algorithm and maximum greedy algorithm, both of which have a lower algorithmic complexity than the greedy algorithm, but still have high algorithmic complexity in large networks. Some special heuristic algorithms were also proposed subsequently. The degree–discount algorithm[12] is an efficient heuristic algorithm. It is an improvement on the optimization strategy of the heuristic algorithm, which has a high practical value. Its basic working principle is that when a node's neighbor has been selected as the propagation source, it will have a discount effect on the degree index of the node, thereby achieving the purpose of de-overlapping. Kimura and Saito[13] used the decomposition of maximally connected subgraphs to find the optimal propagation source. Based on this, they also proposed a method using the maximum influence path.[14] Although this method significantly reduced the algorithmic complexity, it is restricted to only spreading along the shortest path. Wang *et al.*[15] presented an algorithm that combines dynamic programming to select the optimal propagation source, which improved the efficiency of the algorithm. Li *et al.*[16] proposed a new model while considering positive and negative influences; although these algorithms were efficient, it was easy for them to fall into a local optimal situation, which reduced their effectiveness as compared to the greedy algorithm.

Online social networks have the characteristics of a large scale and complex structure. Efficiently identifying the influence of nodes in the network is of practical significance. However, there are still two shortcomings in the current research. We generally use the expected value of the number of nodes to be transmitted by the propagation source to characterize the propagation influence of the propagation source; however, the majority of the existing indicators describe the importance of the node from different angles and can only describe the spread influence in an indirect manner. In contrast, the majority of the existing works present analyses and discussions of the situation of single-layer networks. The identification of nodes with a high propagation influence in multi-layer networks has not been studied.

This paper thus makes the following contributions:

(1) Based on the percolation characteristics of real networks, an iterative algorithm is proposed to solve the probability of the propagation source to transmit information to the network's giant component (global propagation probability). Compared with the previous local simulation algorithm,[7] this algorithm eliminates random errors and significantly reduces the computation time, and the propagation expectations of all nodes in the network can be calculated efficiently.
(2) We analyze why the final propagation range of the nodes in the network obeys the bimodal distribution and provide a necessary and sufficient condition and its proof for obeying the bimodal distribution, which explains the existence and uniqueness of the network's giant component.
(3) Based on the efficient iterative algorithm and the characteristics of the network giant component, we extend the model in this paper to multi-layer networks and find that the propagation range of nodes obeys multimodal distribution. Based on this rule, the expected value of the propagation range of the propagation source in the multilayer network is calculated.
(4) We rank the node propagation influence based on the expected value of the node propagation range and design an efficient greedy algorithm for solving the problem of maximizing the influence of multiple propagation sources. This method is applicable to both single-layer and multi-layer networks and provides a better selection effect than traditional methods.

II. PRELIMINARY

Based on the independent cascading propagation model and percolation theory (this paper refers specifically to bond percolation), this paper presents a comprehensive discussion of the relationship between nodes and the giant component in different networks. Based on this, we present a new index for evaluating the spread influence of a node (or nodes), which has high practical value. Before introducing the work of this article, we first briefly introduce the existing works and conclusions that this article is based on.

A. Equivalence of independent cascade model and bond percolation

The independent cascade (IC) model[17] is one of the most commonly used models for network information dissemination. When a node v is activated, it activates its neighbor nodes with the probability φ. This attempt is made only once, and subsequently, node v is still active but does not have the ability to propagate information. In other words, the edges in the network can propagate information with a probability of φ and fail at information propagation with a probability of $1-\varphi$.

The percolation theory[18–21] has a universal application in the network. It can be used to study classic infectious disease models. It also plays an important role in studying the spread of disease and the spread of public opinion in the network. Our research is based on the theory of bond percolation theory.[22–24] Based on bond percolation theory, there exists a critical threshold φ_c. When the propagation probability $\varphi > \varphi_c$, there exists a giant component with the same order as the network range that is covered by information. If the

initial propagation source we select belongs to a giant component, then the node can propagate information to the network as a whole. Furthermore, as each edge in the network is removed or retained with a probability, the nodes in the giant component are different each time, but the number of nodes in the giant component s^{∞} is approximately a fixed value.

B. Node propagation influence index based on the percolation theory

Based on the percolation theory, the extant literature[7] uses the global propagation probability to reflect the node's propagation capability and calculates it with a multiple local simulation method. This literature presents an important propagation law. When the propagation probability is greater than the percolation threshold, the final propagation range of the selected initial propagation source obeys the bimodal distribution. Therefore, from the selected propagation source, the information is either propagated locally (stops spreading within a small range) or it is propagated to a large component of the network. Therefore, the propagation range expectation can be approximated using Eq. (1) as[7]

$$S(seed) = \sum_{s} s \times p(seed, s) \approx s^{\infty} \times p(seed, s^{\infty}), \tag{1}$$

where $p(seed, s)$ represents the probability that the final propagation range of the propagation source seed is s and s^{∞} represents the number of nodes of the giant component. When the propagation probability is determined, s^{∞} is a fixed value, and thus, the propagation capability of the propagation source can be characterized by the global propagation probability p. Owing to the bimodal distribution of the propagation range, each time a simulation propagation is performed, and Hu *et al.*[7] found through a large number of experiments that when the propagation range is greater than a small threshold m, it can be considered that it will be globally propagated this time. Therefore, in the extant literature,[7] multiple local simulations are used to obtain the global propagation probability.

This indicator directly describes the propagation influence of the transmission source from the expectation of the final transmission range, which is a direct and reliable method. However, no detailed discussions have been presented on the applicable network of this method, the calculation method of the global propagation probability is required to be improved, the results of the local simulation method have a certain random error, and the critical threshold m is difficult to determine. Below, we analyze the applicable network of this method, present the necessary and sufficient conditions for the existence of a giant connected piece, and, based on the meaning of global propagation probability, design an iterative method to solve the system of equations to calculate the global propagation probability of the network nodes. Compared with the local multiple simulation method, the calculation time required is significantly reduced. Based on the efficiency of the algorithm, a method of maximizing the influence of multiple propagation sources is presented using the idea of the greedy algorithm, which has a better effect than the classic method. We also extended this method to multi-layer networks and derived many meaningful conclusions.

III. MODEL ESTABLISHMENT AND ANALYSIS

A. Iterative algorithm for global propagation probability of a single-layer network

In previous works, multiple local simulations were used to calculate the global propagation probability of the propagation source, but there were two shortcomings. First, there are random errors in the simulation values, and second, when there are more nodes to be evaluated, the simulation efficiency is low. In the literature,[25] the percolation characteristics of the network were researched, and it was shown that, for a local tree network, each node in the network has a unique probability that belongs to a giant component and satisfies a certain local equation. As this study is based on an independent cascade propagation mechanism, on considering the bond percolation model, the equations can be solved using an iterative algorithm to efficiently find the global propagation probability of the nodes.

1. First-order iterative form

Let φ be the propagation probability. Let the global propagation probability of node i in network G be p_i, $i \in \{1, 2, \ldots, N\}$. According to the bond percolation theory, we delete the edges in network G with probability $1 - \varphi$; then, the probability that node i belongs to the largest connected component of the remaining network is equal to the probability that node i can propagate the information to the giant component of network G, which is equal to p_i. As shown in Fig. 1(a), we know that the probability that node v belongs to the giant component is equivalent to that of node v having at least one edge connected to the giant component; that is, node v has at least one neighbor belonging to the giant component, and u and v are connected. Thus, the following equation is presented:

$$p_v \approx 1 - \prod_{u \in G.nei(v)} (1 - \varphi p_u), \tag{2}$$

where $G.nei(v)$ represents the set of neighbor nodes of node v. The equal sign cannot be strictly used here because, when the neighbor node u belongs to the giant component, the probability of the existence of the uv edge should be a conditional probability. Thus, the probability is greater than φ, and we still use φ as an approximation. Moreover, when the average degree is large, the error in this approximation is small. This is a high-order nonlinear equation. There are N unknowns, and thus, a direct solution would be complicated. We use an iterative solution method,

$$p_v(t+1) = 1 - \prod_{u \in G.nei(v)} (1 - \varphi p_u(t)), \tag{3}$$

Further improvements can be made as follows:

$$p_v(t+1) = 1 - \prod_{u \in G.nei1(v)} (1 - \varphi p_u(t)) \prod_{u \in G.nei2(v)} (1 - \varphi p_u(t+1)), \tag{4}$$

where $G.nei1(v)$ represents a node set for which the neighbors of node v have not been updated to time $t+1$, and $G.nei2(v)$ indicates a node set for which the neighbors of node v have been updated to time $t+1$.

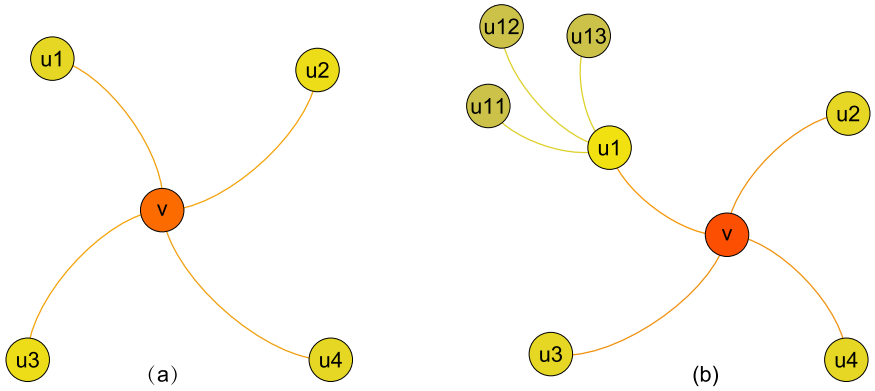


FIG. 1. Local neighbor graph of node *v*. (a) shows the first-order neighbor case of node *v*; (b) shows the second-order neighbor case of node *v*.

2. Second-order iterative form

When the average degree of the network is small, there are some errors in the above-mentioned first-order form model. In order to reduce system errors, the algorithm can be generalized to the second-order form. As shown in Fig. 1(b), when considering the probability that node v leads to a giant component through node u_1, node u_2 must reach the giant component through neighbors other than node v, such that the iterative formula can be improved as follows:

$$\begin{aligned} p_v t+1 &= 1-\prod_{u\in G.nei(v)}(1-\varphi \tilde{p}_u(t)) \\ &= 1-\prod_{u\in G.nei(v)}\left[1-\varphi\left(1-\prod_{s\in G.nei(u),s\neq v}(1-\varphi p_s(t))\right)\right]. \end{aligned} \tag{5}$$

This iterative form significantly reduces the system error, and it can be considered that this value represents the node's global propagation probability. It should be noted that, when selecting the initial value, we can set the initial value of the isolated node as 0 and the initial values of the remaining nodes as non-zero constant values (such as 0.5).

3. Algorithmic complexity analysis

In order to determine the complexity of the algorithm, we first present the pseudo-code form of the above two algorithms. The algorithm flow of the first-order iterative form is presented in Table I, and the algorithm flow of the second-order iterative form is presented in Table II.

Let the average degree of the network be $\bar{d}$; then, in order to determine the global propagation probability of all the nodes in the network, the time complexity of the first-order iterative form is $m\bar{d}N$, and the second-order form is $m\bar{d}(\bar{d}-1)N$. As is apparent, the algorithmic complexity of the iterative method proposed in this paper is linear. Therefore, using Eq. (1), we can efficiently rank the propagation influence of nodes in the network.

Compared with the local multiple simulation method in the literature,[7] this method is deterministic, that is, the result of iterative convergence is uniquely determined. Therefore, this method eliminates random errors caused by a small number of simulations

TABLE I. First-order iterative form for determining the global propagation probability of the nodes in the network.

Iterative algorithm: Approximately solve the global propagation probability of all nodes in the network

Input: **input network G, propagation probabilityφ, and number of iterations m**
output: **global propagation probability vector p of each node in the network**

Obtain the number of nodes in network G as N
Let p be an N-dimensional vector, and each element of the vector is 0.5
for $j = 1 \rightarrow m$ **do**
  for $k = 1 \rightarrow N$ **do**
    sum = 1
    for **i in G.neighbors(k) % i traverses the neighbors of node k**
      sum = sum $\times$ $(1 - \varphi \times p[i])$
    end for
    $p[k] = 1 -$ sum
  end for
end for
return p

TABLE II. Second-order iterative form for determining the global propagation probability of nodes in the network.

Iterative algorithm: Approximately solve the global propagation probability of all the nodes in the network

Input: **input network G, propagation probability φ, and number of iterations m**
output: **global propagation probability vector p of each node in the network**

Obtain the number of nodes in network G as N
Let p be an N-dimensional vector, and each element of the vector is 0.5
for $j = 1 \rightarrow m$ **do**
  for $k = 1 \rightarrow N$ **do**
    sum = 1
    for **i in G.neighbors(k) % i traverses the neighbors of node k**
      $su = 1$
      for **nei in G.neighbors(i) % nei traverses the neighbors of node i**
        if $nei \neq k$
          $su = su \times (1 - \varphi p[nei])$
        end if
      end for
      sum = sum $\times (1 - \varphi \times (1 - su))$
    end for
    $p[k] = 1 -$ sum
  end for
end for
return p

or the use of different experiments. However, through simulation experiments, we found that, although the method of multiple local simulations has a linear order complexity, in order to obtain more accurate results, a large number of simulations is often required. For a network comprising tens of thousands of nodes, the calculation time required for the multiple simulation method is tens to hundreds of times that of the iterative method. Therefore, the method presented in this paper also significantly reduces the calculation time.

B. Model suitability analysis

In this model, we evaluate the propagation influence of a node by its global propagation probability. The first major premise is the propagation probability $\varphi > \varphi_c$, and it is necessary to satisfy that the final propagation range obeys a bimodal distribution. This is equivalent to eliminating the edges in the network that have a probability of $1 - \varphi$; the network has a unique giant component, which indicates that the range of the giant component of the remaining network is approximately a fixed value s^{∞}. It has certain requirements for the network and propagation probability, and a sufficient and necessary condition for the above proposition is presented below:

For a network, the number of nodes is N (N is sufficiently large), the propagation probability is φ, and $\varphi > \varphi_c$; then, the necessary and sufficient condition for the final propagation range of the propagation source to follow the bimodal distribution is as follows. We arbitrarily divide the network into two (or more) subnetworks of the same order as N, as shown in Fig. 2. If both (or more) giant components of these subnetworks exist, the probability that these giant components are connected by edges between the subnetworks will approach 1,

$$1 - \prod_{(u,v)\in E_{\alpha\beta}} (1 - p_{\alpha u} p_{\beta v} \varphi) \rightarrow 1^{-}, \tag{6}$$

then,

$$\prod_{(u,v)\in E_{\alpha\beta}} (1 - p_{\alpha u} p_{\beta v} \varphi) \rightarrow 0^{+}, \tag{7}$$

where $E_{\alpha\beta}$ is the set of connected edges between two subnetworks, $p_{\alpha u}$ is the probability that node u belongs to a giant component of the α subnetwork, and $p_{\beta v}$ is the probability that v belongs to the giant component of the β subnetwork. The proof of the condition is as follows:

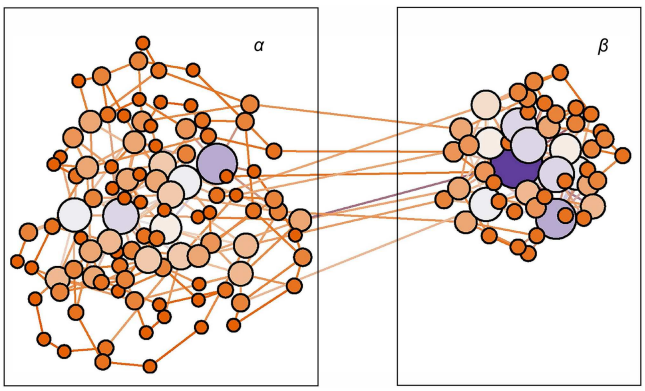


FIG. 2. Randomly divide the network into two subnets: α and β.

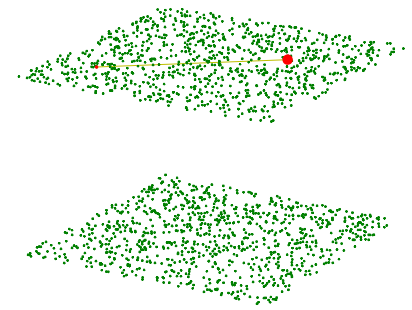
Necessity: We use contradiction if the network has a segmentation of α and β subnetworks, which makes the probability $\sigma < 1$ of the left-hand side of Eq. (6). Then, for a node u in the α subnetwork, the final propagation range is s_α^∞ with a probability $p_{\alpha u} \times (1-\sigma)$, and the final propagation range is $s_\alpha^\infty + s_\beta^\infty$ with a probability $p_{\alpha u}\sigma$. Therefore, the final propagation range is at least a three-peak distribution, which is contradictory to the bimodal distribution, and the null hypothesis does not hold.

Adequacy: We let the propagation range of any two global propagations be S^∞ and s^∞ and set $S^\infty > s^\infty$ and $S^\infty - s^\infty = M$. Conversely, if M and the network range N have same order of magnitude. We consider all the nodes in s^∞ and all edges among this nodes as alpha subnetwork, and all other nodes and their internal edges as the β subnetwork; the following introduces contradictions. We consider two cases. If the two sub-networks have giant components, according to Eq. (7), the probability of the two giant components being disconnected tends to 0, and the second case is unlikely to occur; this is a contradiction. However, if there is no giant component in the β subnetwork, for the first case of S^∞, at least M nodes in the β subnetwork are connected to the α subnetwork. As there is no giant component in the β subnetwork, these M nodes are connected to the α subnetwork through at least m edges, which are of the same order as M, and m and N are thus also of the same order. At this time, the probability that the second case occurs should be less than $(1-\varphi)^m \to 0$, which is contradictory to the fact. Therefore, it is untrue that there is no giant component in the α subnetwork. Therefore, the null hypothesis is untrue, and $M/N \to 0$ while $S^\infty \approx s^\infty$. The final propagation range of the transmission source obeys the bimodal distribution, which is proven.

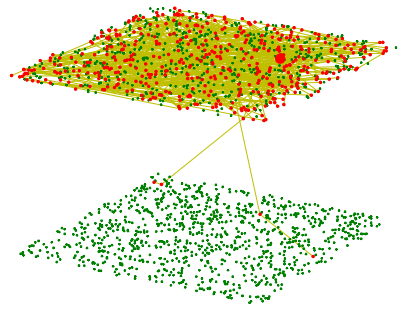
C. Iterative algorithm for global propagation probability of multilayer networks

In real life, many networks can be clearly divided into several layers, such as individuals of different species, different ages, or geographically isolated groups. There are often fewer channels for disease and information transmission among them, and thus, they demonstrate an evident layering phenomenon. Based on the analysis in Sec. III B, this type of network with an evident multilayer structure does not satisfy the condition that the propagation range obeys the bimodal distribution. In the following, we extend the iterative method in Sec. III A to multilayer networks. As the propagation range of the propagation source in the network does not obey the bimodal distribution, but a multimodal phenomenon appears, it is necessary to discuss the probability of propagation to each peak in detail. Below, we analyze the situation of the two-layer network, and the multilayer network can be discussed in the same manner.

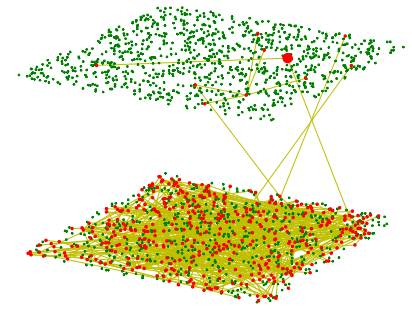
We set the propagation probability of the α subnetwork as φ_1, the propagation probability of the β subnetwork as φ_2, and the propagation probability between the subnetworks as γ. Here, we also consider the case wherein the propagation probability is greater than the seepage threshold, that is, $\varphi_1 > \varphi_{1c}$ and $\varphi_2 > \varphi_{2c}$. We use $p_{\alpha u}$ to represent the probability that node u in the α network belongs to the giant component of the α network, and we use $p_{\beta v}$ to represent



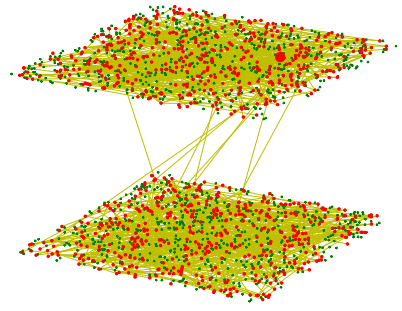
(a)The final propagation of the first peak



(b)The final propagation of the second peak



(c)The final propagation of the third peak



(d)The final propagation of the fourth peak

FIG. 3. Four types of propagation diagram of a two-layer network. We use a computer to generate a two-layer scale-free network. Each layer of the network contains 1000 nodes with an average degree of 3, and 40 edges are randomly added between the two layers of the network. We randomly select a propagation source for multiple simulations and find that only the above four propagation situations occur, where red indicates that the node is propagated and the yellow line is the edges that propagate the information. (a) shows local propagation; (b) shows global propagation on the first layer of the network; (c) shows global propagation on the second layer of the network; (d) shows global propagation on the entire network.

the probability that node v in the β network belongs to the component of the β network. Therefore, using Eqs. (4) and (5), it is possible to iteratively calculate the probability that a node belongs to a giant component of the respective network. We consider the following:

$$\sigma = 1 - \prod_{(u,v)\in E_{\alpha\beta}} (1 - p_{\alpha u} p_{\beta v} \gamma), \tag{8}$$

where $E_{\alpha\beta}$ represents the set of connected edges between two networks and σ represents the probability that the giant components of two networks are connected. The iterative calculation of Eqs. (4) and (5) on the sum network of the α and β networks can provide the iteration probability $p_{\alpha\beta w}$ of any node of the α and β networks. It is easy to analyze the probability that the w node belongs to at least one layer of the giant component in the network, and we thus obtain the following:

$$p_{\alpha u} + \hat{p}_{\beta u} = p_{\alpha\beta u}\,, \quad \hat{p}_{\alpha v} + p_{\beta v} = p_{\alpha\beta v}\,, \tag{9}$$

where $\hat{p}_{\beta u}$ represents the probability that node u in the α subnetwork only propagates to the giant component of the β subnetwork and $\hat{p}_{\alpha v}$ represents the probability that node v in the β subnetwork only propagates to the giant component of the α subnetwork, and thus,

$$\hat{p}_{\beta u} = p_{\alpha\beta u} - p_{\alpha u}\,, \quad \hat{p}_{\alpha v} = p_{\alpha\beta v} - p_{\beta v}. \tag{10}$$

Therefore, the final propagation range of the source node should obey the four-peak distribution, corresponding to the four propagation situations in Fig. 3, respectively. Its propagation probability and range are presented in Table III.

Therefore, the spread of influence indicators can be generalized as follows:

$$S(w) \approx \begin{cases} p_{\alpha w}(1-\sigma)s_\alpha^\infty + (p_{\alpha\beta u} - p_{\alpha u})s_\beta^\infty + p_{\alpha w}\sigma s_{\alpha\beta}^\infty & w \in net\ \alpha, \\ (p_{\alpha\beta v} - p_{\beta v})s_\alpha^\infty + p_{\beta w}(1-\sigma)s_\beta^\infty + p_{\beta w}\sigma s_{\alpha\beta}^\infty & w \in net\ \beta. \end{cases} \tag{11}$$

The number of nodes in a giant component can be approximated according to Eq. (12),

$$s_\alpha^\infty \approx \sum_{u\in net\ \alpha} p_{\alpha u},\ \ s_\beta^\infty \approx \sum_{v\in net\ \beta} p_{\beta v},\ \ s_{\alpha\beta}^\infty \approx s_\alpha^\infty + s_\beta^\infty. \tag{12}$$

In this manner, we can comprehensively discuss the situation of the two-tier network, and the situation of the multi-tier network can be analyzed based on the same analysis.

TABLE III. Correspondence table of peak spread and its probability.

Node u in the α subnetwork				
Peak	Local	s_α^∞	s_β^∞	$s_{\alpha\beta}^\infty$
Probability	$1 - p_{\alpha\beta u}$	$p_{\alpha u}(1-\sigma)$	$p_{\alpha\beta u} - p_{\alpha u}$	$p_{\alpha u}\sigma$
Node v in the β subnetwork				
Peak	Local	s_α^∞	s_β^∞	$s_{\alpha\beta}^\infty$
Probability	$1 - p_{\alpha\beta v}$	$p_{\alpha\beta v} - p_{\beta v}$	$p_{\beta v}(1-\sigma)$	$p_{\beta v}\sigma$

D. Discrimination of communication influence

Based on the influence index of the global propagation probability, we can judge the propagation influence of the nodes. For the problem of ranking the influence of the single propagation source, we can directly use the calculated values of Eqs. (1) and (11) to compare the propagation influence of each node. For the problem of maximizing the influence of multi-propagation sources, in order to consider the overlap of the influence, we design a greedy algorithm based on the efficient iterative method proposed in this paper.

Before introducing the algorithm, let us suppose that the network we are analyzing contains m layers, and the number of nodes of their giant components under a specific probability is $Peak = \{s_1, s_2, \ldots, s_m\}$, then the algorithm execution steps are as follows:

Step 1. Let R $= \phi$. We iteratively calculate the node number set of the peak of the giant components and the corresponding probabilities $p_0(u, s)$, $s \in Peak$, $s \in Peak$ of all the nodes, and we calculate the expected propagation value of all the nodes,

$$Su = \sum_{s\in Peak} p_0(u, s) \times s. \tag{13}$$

The largest node v_1 that maximizes Eq. (13) is selected to join the set R.

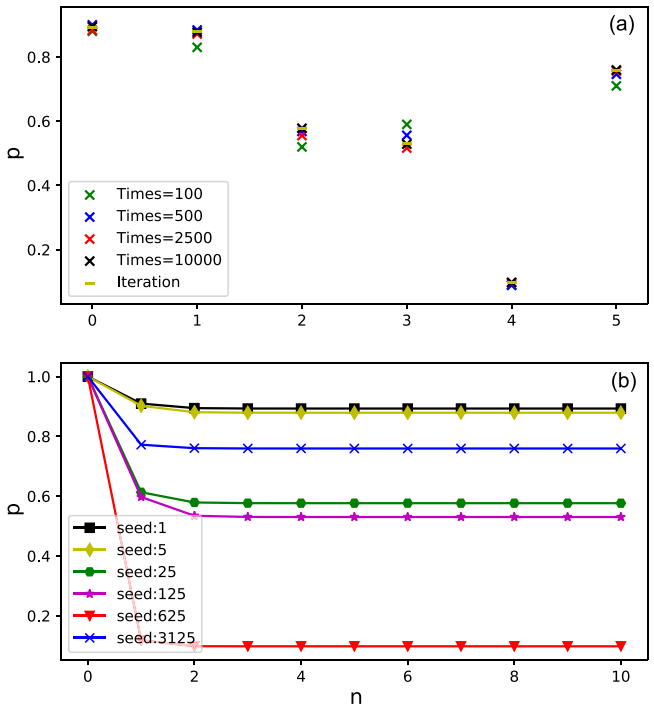


FIG. 4. BlogCatalog friends network is considered in this figure with $\varphi = 0.02$. In (a), we randomly select nodes 1, 5, 25, 125, 625, and 3125; it presents the comparison between the global propagation probability of the simulation calculation results and the results obtained via the iterations. (b) shows the convergence of the global propagation probability values of these nodes with the number of iteration steps. Other nodes and other networks have similar rules.

Step 2. When k nodes have been selected in R ($k \geq 1$), we eliminate the nodes from the set R and their neighboring edges, iteratively calculate the corresponding probability $p_k(u, s)$, $s \in Peak$, and then calculate the expected discount value for the remaining nodes,

$$S(u) = \sum_{s \in Peak} \left\{ p_k(u, s) \times s \times \prod_{v \in R} [1 - \tilde{p}_0(v, s)] \right\}, \quad (14)$$

where $\tilde{p}_0(v, s)$ represents the probability that node v in R can propagate information to peak s (spread range includes peak s), and we select the largest node v_{k+1} that maximizes Eq. (14) to join the set R until the preset requirements are met.

As the algorithm not only provides accurate iteration values but also does not use excessive approximation in the process of de-overlapping, we expect that this multi-source initial node screening effect should provide a better performance than the traditional algorithm. We present the corresponding verification in Sec. IV.

IV. EXPERIMENTS

A. Dataset

In order to verify some of the main conclusions derived using the model in this paper, in addition to performing experiments on the scale-free network generated by the program, we also conduct a simulation analysis using the following real network datasets:

① Deezer: These data are obtained from Deezer, a music-streaming service website, which presents the network of friends of Romanian users. The data are downloaded from http://snap.stanford.edu/data/ (referred to as SLNDC) with a total of 41 737 nodes and an average degree of 6.024.

② Email-Enron: These data are obtained from the Enron e-mail network, downloaded from SLNDC, comprising a total of 36 692 nodes and an average degree of 10.020.

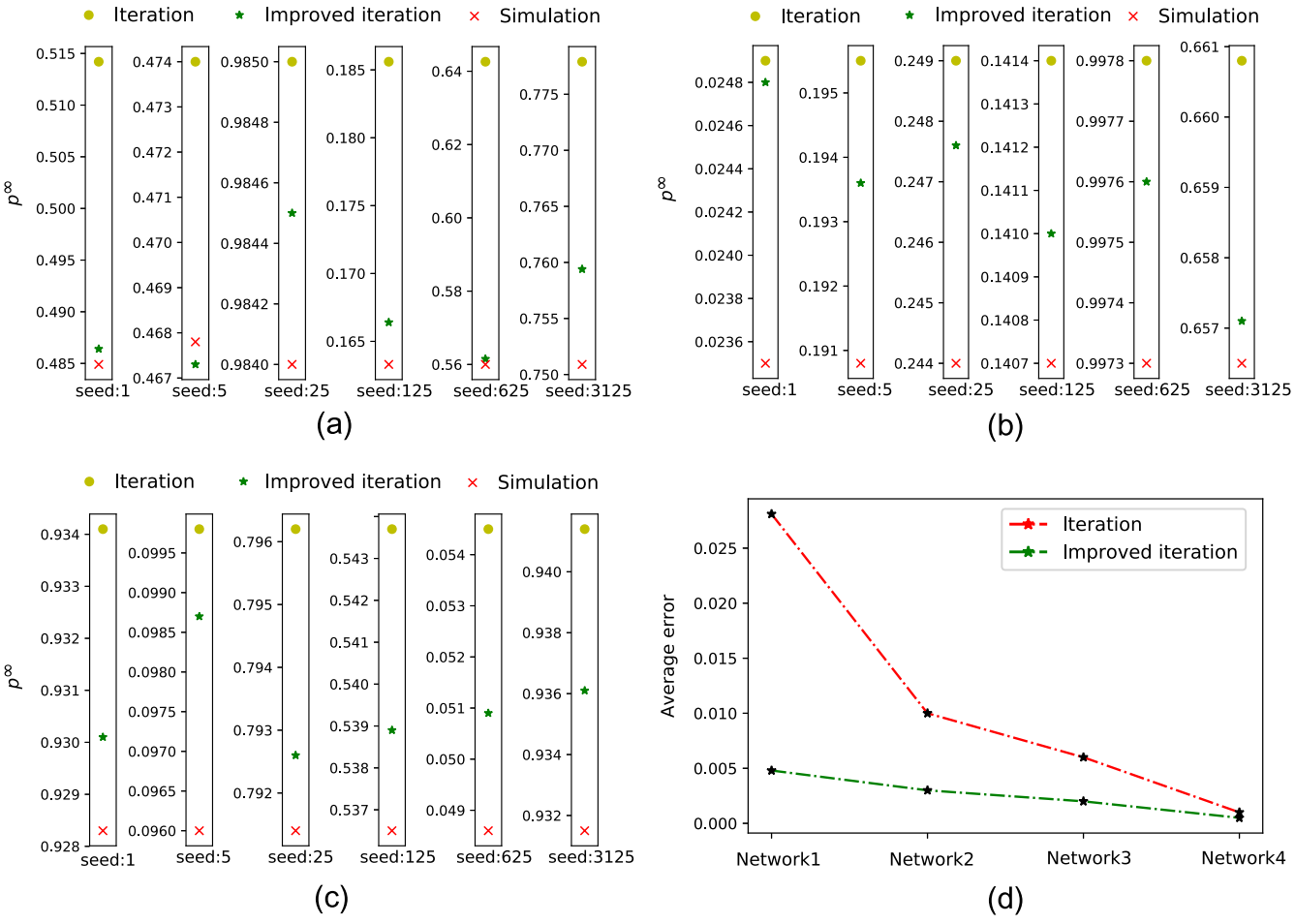


FIG. 5. (a)–(c) reflect the comparison between the global propagation probability of the iterative values and the average of 10 000 simulations in the Deezer network, Email-Enron network, and Facebook network (the propagation probability is 0.3, 0.05, and 0.1, respectively). (d) reflects the simulated errors and true average errors of these six nodes.

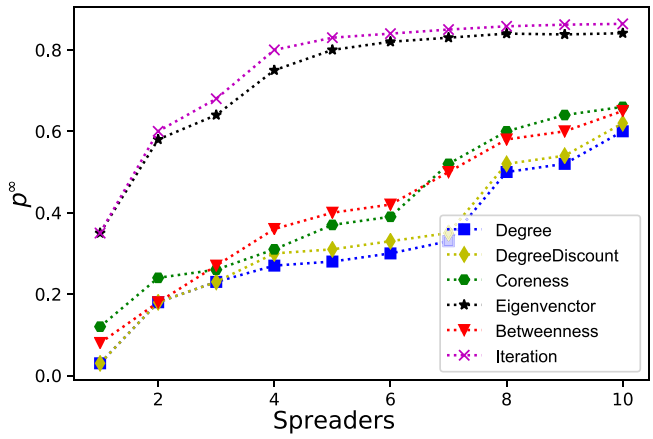


FIG. 6. Comparison of the effect of iterative algorithms and classical methods for screening multiple propagation sources in the Deezer network, where the propagation probability is 0.1, and the pre-selected nodes are nodes with degrees 4–7.

③ Facebook: These data are obtained from the Facebook friends network, are downloaded from SLNDC, and comprise a total of 13 866 nodes with an average degree of 12.528.

④ BlogCatalog: These data are obtained from the BlogCatalog friends network and are downloaded from http://www.blogcatalog.com, with a total of 10 312 nodes and an average degree of 64.776.

B. Experimental analysis in general networks

1. Ranking of single-source influence

We apply the first-order form iterative algorithm to the BlogCatalog friend network with a higher average degree. It can be observed in Fig. 4 that, as the number of simulations increases, the error between the simulation value and the iteration calculation value decreases. Moreover, the algorithm only requires a few steps to converge, which significantly reduces the operation time.

For networks with smaller average degrees, we used the first-order and second-order iterative forms, and we found that the second-order form further reduces the error. The error of some nodes is presented in Fig. 5.

As the index defined in Eq. (1) represents the expected value of the node's final propagation range, it is reasonable to use it to reflect the node's propagation influence. Combined with the accuracy of the iterative algorithm, we believe that this method is effective in ranking the influence issues.

2. Maximizing multi-source influence

For the multi-source problem, we also compare the proposed algorithm with the classic algorithm. Considering the economic benefits, we first select the nodes near the average degree as candidate nodes, and we then use the multiple iterative de-overlapping algorithm described in Sec. III D to screen the initial propagation source. The effect of the application of the proposed algorithm in the Deezer network is presented in Fig. 6, and the corresponding application effect in the other two networks is shown in Fig. 7. It can be observed from Figs. 6 and 7 that the proposed method provides better results than the classical method.

C. Experimental analysis in two-layer network

1. Ranking of single-source influence

In order to verify the distribution characteristics derived in Sec. III C, we considered a constructed two-layer scale-free network. Each layer of the network contains 20 000 nodes with average degrees of 10 and 20, and 150 edges are randomly added between

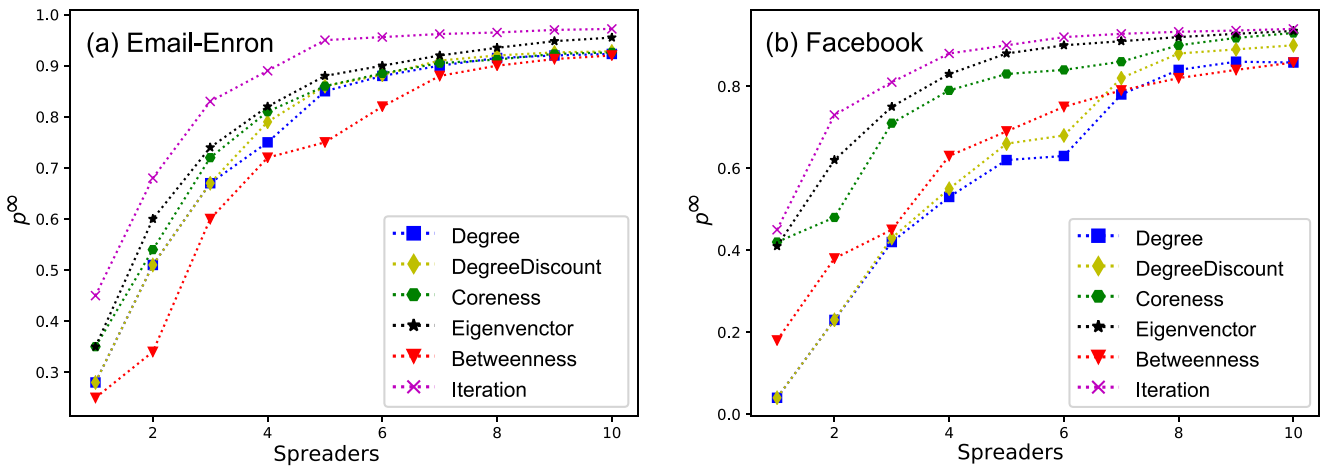


FIG. 7. Screening effect of multiple propagation sources in (a) Email-Enron and (b) Facebook networks.

the two layers of the network. We consider the probability $\varphi_1 = \varphi_2 = \gamma = 0.06$. We randomly selected a node from the first layer of the network as the propagation source and performed 1000 simulations. The distribution of the propagation range is shown in Fig. 8.

The theoretical values of the propagation range and appearance probability of each peak are indicated in the figure. It can be observed that the final propagation range of the propagation source obeys the four-peak distribution, and the propagation range and appearance probability of each peak can be quickly calculated using the method presented in Sec. III C. When the propagation probability and number of edges between the layers of the network are increased, it is easier for the source to propagate into the giant components of the two subnetworks at the same time, and when it increases to a certain threshold, it degenerates into a bimodal distribution.

Using Eq. (11), we can calculate the final propagation expectation of the node. In order to verify the accuracy of the calculated value, we randomly selected 100 nodes from the network, each node carried out 200 simulation propagations, and compared the average value of the final propagation range of these 200 simulations with our calculated values, see Fig. 9.

It is not difficult to see from Fig. 9 that the simulated mean and the calculated value from formula (11) are approximately in a proportional relationship. The error comes from two aspects. On the one hand, the number of simulations is not large enough, it will bring random errors. On the other hand, Eq. (2) is an approximate equation, which will make the model calculation value slightly larger. Nevertheless, their regression equation is $y = 0.998x + 69.43$, and their linear correlation coefficient reaches 0.9923. Therefore, the use of this index to reflect the propagation influence of the node has an advantageous effect.

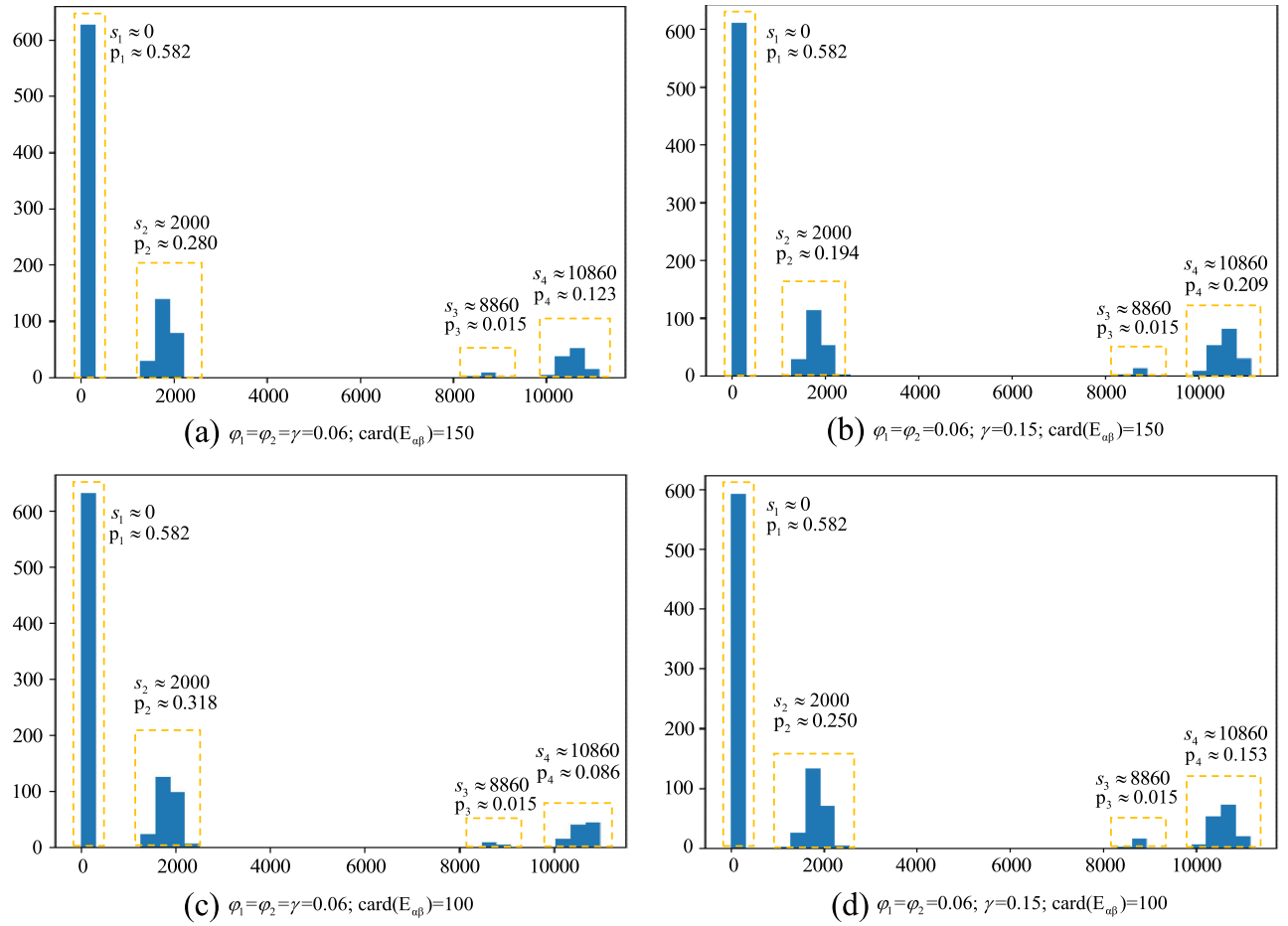


FIG. 8. Distribution of the final propagation range of a propagation source in a two-layer scale-free network. The simulation distribution and theoretical calculation of the final propagation range are presented in (a). Let $\gamma = 0.15$ and the other conditions remain unchanged to obtain (b). The number of edges between the two layers of the network becomes 100, and other conditions remain unchanged to obtain (c). Let $\gamma = 0.15$, and the number of edges between the two layers of the network becomes 100, and other cases remain unchanged to obtain (d).

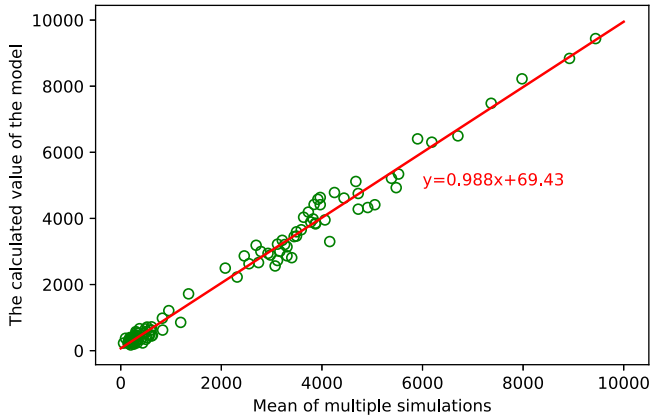


FIG. 9. Comparison of the calculated and simulated values of the final propagation expectations.

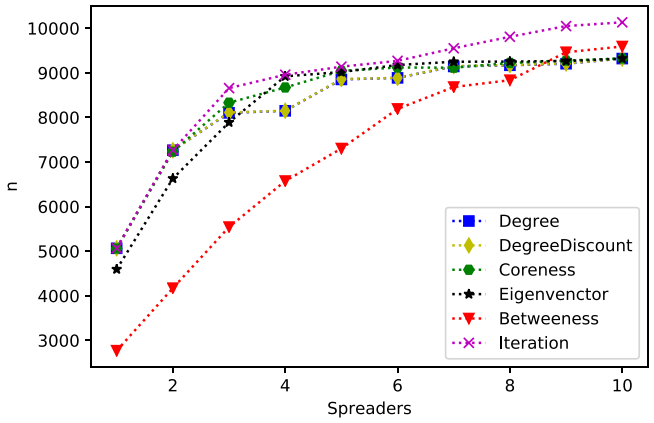


FIG. 10. Comparison of screening effects of multiple propagation sources with different algorithms in a two-layer network.

2. Maximizing multi-source influence

Using the method presented in Sec. III D, we can select the optimal set of propagation sources. We also consider the scale-free networks of the two communities as an example and compare the proposed method with the classical method as follows. In order to compare the pros and cons of each index more clearly, we first randomly select some candidate sets of the propagation sources from the network, and we then use each index to select the optimal node set as the propagation source. The optimal top 10 nodes screened by different algorithms are shown in Table IV. the screening effect of the transmission sources is presented in Fig. 10.

It can be observed that the multiple propagation sources selected using the iterative method have a higher propagation influence. In the final saturation situation, the propagation source selected by the iterative method is significantly better than the classic method. This is because, when other algorithms select multiple propagation sources, they tend to be focused on the layer having a higher average degree. Therefore, the probability of propagation to the giant component of the network with a lower average degree is smaller, and the final total propagation range is also smaller.

V. CONCLUSION

This article is focused on the problem of maximizing the influence of the propagation source in a real network when the propagation probability is greater than the percolation threshold. First, based on the local topology of the network, an efficient iterative algorithm is presented to solve the global propagation probability of all the nodes in the network. Through multiple simulation experiments, we verified that this iterative algorithm can converge quickly and has good accuracy. Second, through an analysis, a sufficient and necessary condition for the final propagation range of the propagation source to follow the bimodal distribution is presented in addition to its proof. Third, based on the necessary and sufficient conditions, the iterative algorithm is extended to multi-layer networks. This paper mainly discusses the situation of the two-layer network and finds that the final propagation range follows the four-peak distribution. Combined with the global propagation probability obtained by the iterative algorithm, the size and probability of propagation to each peak can be calculated approximately through theoretical derivation. Therefore, we can calculate the expected value of the final propagation range, which can well reflect the propagation influence of the node. Finally, when there are multiple propagation sources to be selected, we also propose an algorithm to eliminate influence

TABLE IV. The optimal top 10 nodes screened by different algorithms.

	1	2	3	4	5	6	7	8	9	10
Degree	20 644	21 142	21 275	21 335	21 366	21 483	21 540	21 698	21 756	21 765
Degree–Discount	20 644	21 142	21 275	21 335	21 366	21 483	21 540	21 698	21 756	21 765
Coreness	20 644	20 915	21 036	21 048	21 062	21 142	21 275	21 283	21 295	21 335
Eigenvenctor	22 423	25 163	25 571	24 474	38 025	29 854	30 008	24 048	27 617	22 895
Betweenness	7150	2672	30 621	25 000	38 177	6768	29 539	30 641	4293	824
Iteration	20 644	21 295	21 335	22 395	106	21 142	334	258	21 366	274

overlap. Through experimental data verification, it can be determined that the model proposed in this paper is effective in ranking the influence of a single propagation source and maximizing the influence of multiple propagation sources.

This paper uses the global propagation probability to derive the expected value of the node's propagation range, which is an extension of the node's many indicators and has a better recognition effect in specific problems. However, there are still some issues to be further studied. First, the convergence and efficiency of the iterative equations proposed in this paper also require strict proof. Then, the method presented in this paper requires that the network structure be known. The applicability of the method is required to be further analyzed for cases wherein the network structure information is incomplete. Finally, this article only presents a detailed discussion of a situation of two-layer networks. For networks comprising more layers, we can use the same idea for solving the problem of estimating the influence of nodes in multi-layer networks from the perspective of information explosion.

DATA AVAILABILITY

All data used during the study are available online in dezeer at http://snap.stanford.edu/data/gemsec-Deezer.html, Ref. 26; Email-Enron at http://snap.stanford.edu/data/email-Enron.html, Ref. 27; Facebook at http://snap.stanford.edu/data/gemsec-Facebook.html, Ref. 26; BlogCatalog at http://www.blogcatalog.com, Ref. 28.

REFERENCES

[1] M. Fershtman, "Cohesive group detection in a social network by the segregation matrix index," Soc. Networks **19**, 193–207 (1997).
[2] D. B. Chen, L. Y. Lu, M. S. Shang, Y. C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," Physica A **391**, 1777–1787 (2012).
[3] B. O. Dabbousi, J. Rodriguez-Viejo, F. V. Mikulec, J. R. Heine, H. Mattoussi, R. Ober, K. F. Jensen, and M. G. Bawendi, "(CdSe) ZnS core–shell quantum dots: Synthesis and characterization of a size series of highly luminescent nanocrystallites," J. Phys. Chem. B **101**, 9463–9475 (1997).
[4] M. Lei, L. Zhi, and T. Xiang-Yang, C. Mao, and L. San-Ya, "Evaluating influential spreaders in complex networks by extension of degree," Acta Phys. Sin. **8**, 387–397 (2015).
[5] L. C. Freeman, "A set of measures of centrality based on betweenness," Sociometry **40**, 35–41 (1977).
[6] A. Bavelas, "Communication patterns in task-oriented groups," J. Acoust. Soc. Am. **22**, 725730 (1950).
[7] Y. Q. Hu, S. G. Ji, Y. L. Jin, H. E. Stanley, and S. Havlin, "Local structure can identify and quantify influential global spreaders in large scale social networks," Proc. Natl. Acad. Sci. U. S. A. **115**, 7468–7472 (2018).
[8] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2001), pp. 57–66.
[9] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2003), pp. 137–146.
[10] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2007), pp. 420–429.
[11] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2009), pp. 199–218.
[12] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Shokrollahi, "Improving detection of influential nodes in complex networks," Physica A **436**, 833–845 (2015).
[13] M. Kimura and K. Saito, "Approximate solutions for the influence maximization problem in a social network," in Knowledge-Based Intelligent Information and Engineering Systems: 10th *International Conference,* Bournemouth (KES, 2006), p. 259.
[14] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in Proc*eedings of the* 16th ACM SIGKDD Int*ernational* Conf*erence* on Knowledge *D*iscovery and *Data Mining* (ACM, 2010), pp. 1029–1038.
[15] Y. Wang, G. Cong, and G. Song, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks," in Proc*eedings of the 16th ACM SIGKDD International* Conf*erence* on Knowledge *D*iscovery and *M*ining (ACM, 2010), pp. 1039–1048.
[16] D. Li, Z. M. Xu, and N. Chakraborty, "Polarity related influence maximization in signed social networks," PLoS One **9**, e102199 (2014).
[17] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," Nature **464**, 1025–1028 (2010).
[18] S. Havlin, H. E. Stanley, A. Bashan, J. Gao, and D. Y. Kenett, "Percolation of interdependent network of networks," Chaos Solitons Fractals **72**, 4–19 (2015).
[19] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," Nature **524**, 65–68 (2015).
[20] R. Pastor-Satorras, C. Castellano, P. V. Mieghem, and A. Vespignani, "Epidemic processes in complex networks," Rev. Mod. Phys. **87**, 925–979 (2015).
[21] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Critical phenomena in complex networks," Rev. Mod. Phys. **80**, 1275–1335 (2008).
[22] H. Kesten, "The critical probability of bond percolation on the square lattice equals 1/2," Commun. Math. Phys. **74**, 41–59 (1980).
[23] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin, "Networks formed from interdependent networks," Nat. Phys. **8**, 40–48 (2012).
[24] X. Liu, H. E. Stanley, and J. Gao, "Breakdown of interdependent directed networks," Proc. Natl. Acad. Sci. U. S. A. **113**, 1138–1143 (2016).
[25] F. Radicchi, "Percolation in real interdependent networks," Nat. Phys. **11**, 597–602 (2015).
[26] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "GEMSEC: Graph embedding with self clustering," in ACM International Conference on Advances in Social Networks Analysis and Mining (ACM 2019), pp. 65–72.
[27] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," Internet Mathematics **6**, 29–123 (2009).
[28] R. Zafarani, and H. Liu, "Users joining multiple sites: Distributions and patterns," in International Conference on Weblogs and Social Media (AAAI, 2014), pp. 635–638 (2018).