Research Article

A Comprehensive Algorithm for Evaluating Node Influences in Social Networks Based on Preference Analysis and Random Walk

Chengying Mao and Weisong Xiao

School of Software and Communication Engineering, Jiangxi University of Finance and Economics, 330013 Nanchang, China

Correspondence should be addressed to Chengying Mao; maochy@yeah.net

Received 18 March 2018; Revised 3 August 2018; Accepted 14 August 2018; Published 8 October 2018

Academic Editor: Ana Meštrović

Copyright © 2018 Chengying Mao and Weisong Xiao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the era of big data, social network has become an important reflection of human communications and interactions on the Internet. Identifying the influential spreaders in networks plays a crucial role in various areas, such as disease outbreak, virus propagation, and public opinion controlling. Based on the three basic centrality measures, a comprehensive algorithm named PARW-Rank for evaluating node influences has been proposed by applying preference relation analysis and random walk technique. For each basic measure, the preference relation between every node pair in a network is analyzed to construct the partial preference graph (PPG). Then, the comprehensive preference graph (CPG) is generated by combining the preference relations with respect to three basic measures. Finally, the ranking of nodes is determined by conducting random walk on the CPG. Furthermore, five public social networks are used for comparative analysis. The experimental results show that our PARW-Rank algorithm can achieve the higher precision and better stability than the existing methods with a single centrality measure.

1. Introduction

With the rapid development of network and information technology, the applications in form of rich media have involved in all aspects of our lives. Accordingly, the interaction and communication between individuals have become more and more convenient and frequent. For example, the platforms such as Facebook, WeChat, QQ, and WhatsApp are very helpful for users to deliver their messages, options, or pictures. As a result, the individuals in the society have been tightened together in an invisible way, that is, the so-called *social network* [1, 2]. Under the impetus of intelligent mobile terminals like iPhone, the scale of social network has a sharp increase in recent years. As reported by TechCrunch, the monthly active users of Facebook have climbed to 2 billion in the middle of 2017 [3]. Similarly, WeChat, as one of the most impactful mobile products, has monthly users which are over 980 million now [4]. Faced with such a huge and complex social network, we usually feel hard and tricky to analyze its overall features and understand the behaviors of the individuals in it. Consequently, the analysis

and modeling of social networks have caused much attention in the recent two decades [5].

At the early stage, the studies mainly focus on the static statistical properties that characterize the structure of social networks [6]. Some concepts, such as degree distribution, clustering coefficient, and the average path length, have been proposed and widely applied to the measurement of social networks. Although the above metrics can reflect the feature of overall network or a single node well, they usually ignore the dynamic interaction behaviors of the individuals in a network [7]. Nowadays, the structure of social networks is not just a mathematical toy; it has been employed extensively as a model of real-world networks in various types, such as network of friendships, network of telephone calls, and network in epidemiology. In most of these application scenarios, the dynamic features of nodes or communities need to be deeply investigated so as to make scientific decisions. For example, in public opinion emergencies, the recognition of special individuals like opinion leaders and the evaluation of the spreading of their influences can contribute to the understanding and controlling of the

opinion (or rumor) transmission [8, 9]. Similarly, the identification of influential individuals is also very helpful in controlling the disease spreading.

As reported in [10–12], many mechanisms such as cascading, spreading, and synchronizing in social networks are highly affected by a tiny fraction of influential nodes. In other words, *identifying influential nodes* is an effective way to reveal the potential disciplines behind the information, rumor, or disease spreading over social networks. Due to the theoretical and practical significance, how to identify the influential nodes in a social network has been widely investigated in recent years [13–15]. At present, quite a few centrality indicators have been presented to address this problem. Typically, *degree centrality* [16], *betweenness centrality* [16, 17], and *closeness centrality* [16] are three well-known measures. However, most of them quantify the influence of nodes in a network from the perspective of a single indicator. Although the single measure is reasonable from its own point of view, it is usually lack of the ability of comprehensive evaluation. At the same time, each measure has its own advantages and limitations. Thus, it is more appropriate to consider multiple different measures simultaneously. In this paper, we attempt to integrate the above three representative measures together by preference analysis and then adopt the random walk algorithm to rank the nodes in a network according to their spreading influences. In order to validate the effectiveness of our proposed algorithm, we use the Susceptible–Infected–Recovered (SIR) [18] model to evaluate the rationality of node ranking results.

Recently, some hybrid approaches have been proposed for the influence maximization problem. In their solutions, several different measures like degree centrality are usually taken into account to design a comprehensive model for evaluating the influence spread of node. Typically, Jalayer et al. proposed a “greedy TOPSIS and community-based” (GTaCB) algorithm [19] for this problem. It could be seen that the TOPSIS in [19, 20] belongs to a greedy technique. Thus, it only generates a local optimal solution for identifying the influential nodes in a social network. By contrast, the random walk technology used in our solution is a global optimization algorithm. In theory, the rank generated by random walk is more reasonable than that of the TOPSIS-based method. In literature [21], Ko et al. proposed the Hybrid-IM algorithm to maximize the influence spread over a social network by combining PB-IM (path-based influence maximization) and CB-IM (community-based influence maximization). In general, it is not easy to collect the information about path and community from a social network. By contrast, in our algorithm, three basic and typical measures are used for the preference analysis, so it can be easily implemented and has a certain advantage in efficiency. The main contribution of our work is the comprehensive evaluation framework for node influences by combining *preference analysis* and *random walk*. In this paper, although we only adopt three basic centrality measures, the measures used in the framework can be replaced and extended according to some specific requirements. That is, our approach has good scalability in the integration of multiple measures.

The remainder of this paper is organized as follows. In Section 2, we describe the problem to be solved and review some background knowledge. Section 3 presents the overall framework of a comprehensive algorithm for evaluating node influences firstly, and then addresses the technical details. Subsequently, the experimental comparison and analysis are conducted in Section 4. Section 5 discusses the threats to validity and the potential extension. The related studies about the evaluation of node influences are addressed in Section 6. Finally, the conclusions and further research directions are stated in Section 7.

2. Preliminaries

2.1. Problem Description. During the spreading process of diseases or rumors, their influences are usually sparked by one or several initial nodes in a social network. Due to the difference in the location of node in the entire network structure, different nodes will have different transmission abilities for disease or rumors and thus will bring different influences on the network. Therefore, it is very necessary to evaluate nodes’ influences and then rank them. This measurement is helpful in scientific decision-making on social networks, such as the monitoring of public opinion transmission and the controlling of disease propagation.

In this paper, we assume that the initial source of spreading is only due to one node in a network. Then, the *node influence analysis* in a social network can be formally described as below: given a social network represented by a directed graph $G = (N, E)$, for each node $i \in N$, its influence is firstly measured by considering its location and the connections to other nodes in G , and then all nodes are ranked according to their influence metrics. Here, N and E are the node set and edge set of such network, respectively.

It should be noted that the information or disease propagation may be caused by several original source nodes in a social network, and hence to identify multiple influential nodes is also an interesting problem [22]. In this paper, however, we mainly focus on the influence evaluation for a single source node.

2.2. Three Typical Measures for Node Influences. In this study, our objective is to design a framework for evaluating node influences in a social network through comprehensively considering some basic measures. In the past, quite a few measures have been presented to capture the importance of each node in a network. *Degree centrality* [16], *betweenness centrality* [16, 17], and *closeness centrality* [16] are three basic, representative, and widely used measures to reflect the influence of node. As a result, the above three measures are very suitable for use in our comprehensive model of influential node identification. Similarly, in the literatures [23–26], the three measures are also used in their MADM (multiple-attribute decision-making) models to identifying influential nodes. Here, we firstly give a brief review on them.

2.2.1. Degree Centrality. The degree centrality is the earliest and most simple method to depict the influence of node in a network. For node i , the influence is directly reflected by

its degree, that is, so-called degree centrality. Here, it is denoted as $C_D(i)$ and formally defined as

$$C_D(i) = \frac{\text{degree}(i)}{n-1}, \quad (1)$$

where $\text{degree}(i)$ is the degree of node i , and n is the number of nodes in the given network.

Degree centrality measures the node's importance from the perspective of degree. Its inherent limitation lies in that it can only reflect the local structure around a given node, i.e., the node and its neighbors, but the reachability from it to the nodes beyond its neighborhood is completely ignored.

2.2.2. Betweenness Centrality. The betweenness centrality is used to capture how well situated a node is in terms of paths that it lies on. Specifically, for a node i in network G , its betweenness centrality (denoted as $C_B(i)$) is the fraction of the shortest paths passing through node i to all shortest path pairs in network G .

$$C_B(i) = \sum_{s \neq i \neq t \in N} \frac{\sigma_{st}(i)}{\sigma_{st}}, \quad (2)$$

where σ_{st} is the number of the shortest paths between nodes s and t , and $\sigma_{st}(i)$ denotes the number of the shortest paths between s and t which pass through node i .

It is easy to see that betweenness centrality is a measure to reflect the gateway feature of a node. But it has poor capability to express the strength of connections from the node of interest to its neighbors.

2.2.3. Closeness Centrality. The closeness centrality is a measure of tracking how close a given node is to any other nodes in a network. For node i , its closeness centrality, denoted as $C_C(i)$, can be defined as

$$C_C(i) = \frac{n-1}{\sum_{j=1}^n d_{ij}}, \quad (3)$$

where d_{ij} represents the distance between node i and node j , and n is the number of nodes in the network.

According to the definition in (3), closeness centrality can characterize the speed of information propagation for a given node, but it cannot distinguish the difference in node location information like the gateway.

Based on the analysis on the above three measures, we can find that each measure has its own specialty for reflecting information (or disease) propagation, but it also has shortcomings. Therefore, combining these representative issues into a comprehensive measure is probably a rational way for identifying influential nodes. As mentioned earlier, the basic measures in our framework can be extended or replaced according to the specific requirements. Besides the above three centrality measures, quite a few other measures have been presented in recent years, such as diffusion centrality [27], sociability centrality [28], and BridgeRank [29]. In fact, all these basic measures can be applied into our comprehensive framework for evaluating the influences of nodes in a

social network. For the sake of simplicity, we only take the three basic and representative centrality measures into consideration in this study.

2.3. Random Walk Model. The random walk model is a special case of Markov chain, that is, a finite and time-reversible Markov chain. It arises in many models in mathematics and physics [30]. In the field of computer science, the rank walk is usually modeled in the following way: suppose there is a system with m states, and the initial probability distribution of these states is represented as $\pi_0 = [p_0(1), p_0(2), \dots, p_0(m)]$. In this system, the states can be transited to each other. Specifically, if state i has different transition probabilities to other k states, the sum of these probabilities should be 1.0, that is, $\sum_{j=1}^k p_{ij} = 1$, where p_{ij} is the transition probability from state i to state j . The transition probabilities of all state pairs can be represented as a probability matrix of state transition, i.e., $\mathbf{P}$. Thus, the random walk model can be clearly described through using matrix notations [31]. Let $\pi_t = [p_t(1), p_t(2), \dots, p_t(m)]$ be the probability distribution of m states after walking t steps, then it can be iteratively calculated according to the initial distribution π_0 as below.

$$\pi_t = \pi_{t-1} \mathbf{P}. \quad (4)$$

In fact, the random walk model can be easily applied to the directed graph [30]. For nodes in a directed graph, we can consider them as states. At the same time, the connection strength of a directed edge between two nodes is treated as transition probability. Once an initial distribution for all nodes is determined, the stationary distribution can be yielded finally through the finite-step transitions shown in (4).

3. Comprehensive Algorithm for Evaluating Node Influences

3.1. The Overall Framework. In the paper, we attempt to design a comprehensive algorithm for evaluating node influences by synthetically considering three basic and independent measures about influence. Thus, the three basic measures are the input data for further processing in our algorithm. Here, assume that the basic measures, such as C_D , C_B , and C_C , are obtained by degree counting and path analysis on the given network according to (1), (2), and (3).

As shown in Figure 1, the procedure of comprehensively ranking nodes according to their influences can be divided into the following two steps: At the first step, for each basic measure, the metrics of all nodes are firstly regulated into the interval from 0 to 1. Then, the preference relation of each node pair is analyzed by comparing the metrics of two nodes in the pair. Based on the preference relations, a sub-graph of the preference relation (also known as partial preference relation graph) can be built. In this graph, nodes are still the nodes in the original network, but each edge represents the preference relation of two nodes with respect to the given basic measure. Secondly, a complete preference relation graph is formed by adding l subgraphs together. In this paper, l is set to 3 because we mainly

FIGURE 1: The overall framework for ranking nodes according to their influences.

combine three centrality measures (i.e., C_D , C_B , and C_C) together in our algorithm. Then, the complete graph is converted to a matrix, and the regulation is performed on it for further computation. Finally, a ranked list of nodes can be generated by applying a random walk on the complete model of preference relations.

It should be noted that, in this paper, only three basic centrality measures are taken into account in the algorithm. However, the above framework is a scalable model for evaluating node influences. That is, besides the basic measures, some other advanced measures can also be adopted in it. Each measure has its own advantages and limitations in representing the influences of nodes, so the complementarity should be considered when choosing the measures for use in our framework.

3.2. The Technical Details

3.2.1. Analysis on Partial Preference Relations. In order to address the technical details of our proposed algorithm, a small network (graph) is used as a running example. As shown in Figure 2, the whole network consists of seven nodes and eight undirected edges. Intuitively, node 3 is the most influential node in this network, and node 1 is the second one. Node 7 should be the weakest one with respect to the influence or the ability of information spreading. Furthermore, node 5 has higher influence than node 7, but is lower than other remainders. However, the influences of nodes 2, 4, and 6 are difficult to distinguish only by subjective assessment.

According to the definitions in Section 2.2, three measures of each node in Figure 2 can be calculated and illustrated in Table 1. For the measure of degree centrality (C_D), node 3 has the max degree (i.e., 4) and node 7's degree is just only one. If we rank nodes in accordance with this indicator, the order can be expressed as below: $3 \rightarrow 1 \rightarrow 2 \sim 4 \sim 5 \sim 6 \rightarrow 7$. Here, $i \rightarrow j$ means node i has higher influence than node j , and $i \sim j$ means nodes i and j have no significant difference in influence. Although the above ranking result about C_D is consistent with the subjective and intuitive judgment, quite a few nodes have

FIGURE 2: A running example for evaluating node influences.

the same degree so that it is hard to provide an accurate order. For the running example in Figure 2, it is difficult to distinguish nodes 2, 4, 5, and 6 in a strictly ordered relation.

With regard to the basic measure of betweenness centrality (C_B), node 3 also has the highest value, but the betweenness values of nodes 2 and 7 are both zero. The rank about C_B is as follows: $3 \rightarrow 6 \rightarrow 1 \rightarrow 4 \rightarrow 5 \rightarrow 2 \sim 7$. Since the betweenness centrality mainly focuses on the connection feature of nodes and ignores the spreading breadth of node influences, the above rank has little conflict to the subjective result.

For the third measure, i.e., closeness centrality (C_C), it is defined as the inverse of farness, which in turn is the sum of distances from the current node to all other nodes. In this example, the rank of all seven nodes with respect to C_C is $3 \rightarrow 1 \rightarrow 2 \sim 4 \sim 6 \rightarrow 5 \rightarrow 7$. We can find that there are still quite a few nodes in the same order, such as nodes 2, 4, and 6 here.

As mentioned above, the basic measures of node influences merely reflect one aspect of information (or disease) spreading features and behaviors. On the other hand, it may be difficult to distinguish the order of some nodes due to the same metric values. In the paper, we present a new ranking algorithm by comprehensively considering the above

TABLE 1: Three typical measures of node influences.

Measures	1	2	3	Nodes 4	5	6	7
Degree (C_D)	3	2	4	2	2	2	1
Betweenness (C_B)	5	0	19	3	1	10	0
Closeness (C_C)	0.06	0.05454	0.075	0.05454	0.04615	0.05454	0.0375

TABLE 2: The normalized data of three measures for the running example.

Measures	1	2	3	Nodes 4	5	6	7
Degree (C_D)	0.6667	0.3333	1.0	0.3333	0.3333	0.3333	0.0
Betweenness (C_B)	0.2632	0.0	1.0	0.1579	0.0526	0.5263	0.0
Closeness (C_C)	0.6	0.4544	1.0	0.4544	0.2307	0.4544	0.0

three basic measures. The whole algorithm consists of two key steps: partial preference analysis and random walk on the complete preference graph.

To perform the analysis on partial preference relations, for each basic measure, it needs to judge the preferences between nodes in a network.

Definition 1 (preference relation). Given a measure of node influence, the preference on a pair of nodes can be modeled in the form of function $\Psi : N \times N \rightarrow \mathbb{R}$, where $\Psi(i, j) > 0$ means node i has stronger influence than node j . The preference function $\Psi(i, j)$ is defined as the difference in influence measure between node i and node j .

Take the measure C_D for example, the preference function can be expressed as $\Psi(i, j) = C_D(i) - C_D(j)$. For nodes 1 and 2 in Figure 2, $\Psi(1, 2) = C_D(1) - C_D(2) = 3 - 2 = 1$; thus, node 1 is more preferable than node 2 for information spreading. However, for nodes 2 and 4, they are equal to each other because $\Psi(2, 2) = 0$. Of course, the above definition can also be applied to other two measures C_B and C_C in a similar way.

The value of $\Psi(i, j)$ indicates the strength of preference, and a value of zero means that there is no preference between two nodes. Here, we set $\Psi(i, i) = 0$ for all $i \in N$.

Based on the above definition of preference relation, the partial preference graph for a given influence measure can be further defined as below.

Definition 2 (partial preference graph, PPG). Given a measure of node influence, if the preferences of all node pairs are analyzed, the partial preference graph $PPG = (N, E_p)$ with respect to the given measure can be constructed as follows: N is the set of nodes in the original social network, and E_p represents the set of preference relations between nodes. Specifically, if $\Psi(i, j) > 0$, there is an edge from node i to node j and the weight of the edge is assigned with the value of $\Psi(i, j)$. For the sake of simplicity, if the preference of a node pair is zero, the corresponding edge is omitted in the graph.

Since the ranges of different measures are not identical, it is hard to merge them into a comprehensive model for ranking node influences. Thus, we normalize the values of each measure for all nodes firstly and then generate the corresponding partial preference graph (PPG). Finally, based on the PPG related to each kind of basic measures, the comprehensive preference graph can be built.

In our work, we adopt *min-max normalization* [32] to perform a transformation on the original data of each measure in Table 1. The normalization formulation can be referred as

$$v' = \frac{v - v_{\min}}{v_{\max} - v_{\min}}, \quad (5)$$

where v is the original measure of a given node, and $v_{\max}$ and $v_{\min}$ are the maximum and minimum of such measure for all nodes in the network, respectively. Based on the above transformation, the original metrics in Table 1 can be converted to the new data shown in Table 2.

For the measure of degree centrality (C_D), the partial preference graph and the corresponding matrix of this simple network are modeled in Figure 3 according to the normalized data in Table 2. In the preference matrix (see Figure 3(b)), if two nodes have no preference relation, the element related to these two nodes in the matrix is set to zero.

In a similar way, the partial preference graphs with respect to the other two measures (i.e., betweenness centrality and closeness centrality) are built and demonstrated in Figure 4. Since the concerns of three basic measures are different, the generated PPGs according to these measures are not very consistent. Each measure has its own sense to judge the node's importance for information spreading and also ignores some aspects of influence measurement. For this reason, comprehensively considering the above three measures seems to be more rational than the separate application.

3.2.2. Comprehensively Ranking for Node Influences. To perform the comprehensive evaluation of node influences, it is necessary to construct a model by combining three

(a) The PPG w.r.t. degree centrality (PPG_D)

(b) The matrix of PPG w.r.t. degree centrality

FIGURE 3: The PPG and corresponding matrix about degree centrality for the example network.

(a) The PPG for betweenness centrality (PPG_B)(b) The PPG for closeness centrality (PPG_C)

FIGURE 4: The PPGs w.r.t. other two measures for the example network.

partial preference graphs together. Here, we define this model as a comprehensive preference graph.

Definition 3 (comprehensive preference graph, CPG). For several PPGs about the same social network, the comprehensive preference graph $CPG = (N, E_{cp})$ can be generated by the following rules: N is the same node set as the PPG and the original social network, and the preference on edge $(i, j) \in E_{cp}$ is formed by summing the preferences of this edge in all PPGs together. Specifically, for each edge $(i, j) \in E_{cp}$, its preference $\Psi_{cp}(i, j)$ is calculated as follows: $\Psi_{cp}(i, j) = \sum_{k=1}^{|PPGs|} \Psi_k(i, j)$, where $|PPGs|$ represents the number of PPGs, and $\Psi_k(i, j)$ is the preference of edge (i, j) in the k th PPG.

To deeply understand the definition of CPG, we use the PPGs about three different measures to illustrate the

construction of CPG. Here, we denote the preferences of edge (i, j) in three PPGs as $\Psi_D(i, j)$, $\Psi_B(i, j)$, and $\Psi_C(i, j)$, respectively. If the above preferences are concordant, $\Psi_{cp}(i, j)$ can be directly achieved by addition operation. For example, $\Psi_{cp}(1, 2) = \Psi_D(1, 2) + \Psi_B(1, 2) + \Psi_C(1, 2) = 0.3334 + 0.2632 + 0.1456 = 0.7422$. However, for the edge $(1, 6)$, the corresponding preference in PPG_D and PPG_C are discordant to that in PPG_B. Thus, it needs to build one edge from node 1 to node 6 and another edge in the opposite direction, that is, $\Psi_{cp}(1, 6) = \Psi_D(1, 6) + \Psi_C(1, 6) = 0.3334 + 0.1456 = 0.4790$, and $\Psi_{cp}(6, 1) = \Psi_B(6, 1) = 0.2361$.

It is not hard to find, during the above construction procedure of CPG, that all three PPGs are regarded as equally important. In the real application scenarios, if some basic measures need to be considered differently, the overall preference of edge in CPG can be defined as $\Psi_{cp}(i, j) = \sum_{k=1}^{|PPGs|} w_k \cdot \Psi_k(i, j)$, where w_k is the importance

FIGURE 5: The regularized CPG and the corresponding matrix for the example network.

weight of the k th basic measure (or PPG). In general, if a measure needs to be considered as a key indicator, a high value is assigned to w_k accordingly, otherwise a small value.

Obviously, in the generated CPG, the sum of the outgoing edges from a node may be not equal to 1.0. To facilitate the latter operations, we firstly regularize the CPG according to the following regularization.

Definition 4 (regularized CPG, CPG_r). Given a CPG, suppose there are τ outgoing edges from node i , then the preference of any edge (i.e., $\Psi_{cp}(i, j) (1 \leq j \leq \tau)$) from such node can be regularized as below. Accordingly, the transformed CPG is denoted as *regularized CPG*.

$$\Psi'_{cp}(i, j) = \frac{\Psi_{cp}(i, j)}{\sum_{k=1}^{\tau} \Psi_{cp}(i, k)}. \quad (6)$$

Based on the above definition, the final regularized and comprehensive preference graph (i.e., CPG_r) of the running example can be constructed in Figure 5(a). Meanwhile, the matrix form of CPG_r is illustrated in Figure 5(b). Obviously, in each row of the comprehensive preference matrix M , the sum of preference values is equal to 1.0. In other words, the matrix M satisfies the basic property of the Markov chain.

As mentioned earlier, the random walk model can be applied to the directed graph to make scientific decisions about ranking. In this paper, we apply random walk to the comprehensive preference graph to rank the node influences

in a social network. In the application, each node is attached with an important factor, and the preference between two nodes is considered as the transition probability of node importance. Here, our goal is to obtain a relatively stable probability distribution over nodes through the iterations in random walk, where the probability is interpreted as the importance or influence of each node. Since the probability of node reflects its importance or influence, the final stable probability distribution can be used to rank nodes.

For the example social network, the rank of seven nodes can be generated through applying random walk to its regularized CPG or the corresponding matrix M (refer to Figure 5). Initially, seven nodes in the graph are considered equally; that is, the importance of each node is set to 1/7. Thus, the initial vector reflecting nodes' importance (or influences) can be represented as $\pi_0[1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7]$. Then, the vector of node influences can be iteratively updated as below.

$$\begin{aligned} \pi_1 &= \pi_0 M \\ \pi_2 &= \pi_1 M \\ &\dots \\ \pi_t &= \pi_{t-1} M. \end{aligned} \quad (7)$$

When step number t reaches 10, the achieved importance distribution of seven nodes is $\pi_{10} = [2.2246e-11, 7.1884e-09, 0.0000e+00, 6.0979e-11, 1.2118e-08, 2.4453e-11, 1.2488e-07]$. For the generated vector, the

Input: (1) the degree measures (C_D) of all nodes in the given network;
 (2) the betweenness measures (C_B) of all nodes;
 (3) the closeness measures (C_C) of all nodes;
 (4) the step number (T) of iterations in random walk.

Output: The rank (R) of nodes in the social network.

1. **for each** basic measure in $\{C_D, C_B, C_C\}$ **do**
2. apply the *min-max* normalization to convert the current measure vector;
3. analyze the preference relation of each node pair;
4. build a partial dependence graph (*PPG*) based on the above preference relations;
5. represent current *PPG* to the corresponding matrix, i.e., $\mathbf{M}_D$, $\mathbf{M}_B$ or $\mathbf{M}_C$;
6. **end for**
7. combine three basic *PPGs* together to form a *CPG*, its matrix ($\mathbf{M}_{sum}$) is the weighted sum of $\mathbf{M}_D$, $\mathbf{M}_B$ and $\mathbf{M}_C$;
8. apply the regularization on each row in $\mathbf{M}_{sum}$ to form a regularized matrix $\mathbf{M}$;
9. set $\pi_0 = [(1/n), (1/n), \dots, (1/n)]$ as the initial probabilities for n nodes in network;
10. **for** t is from 1 to T **do**
11. apply rule $\pi_t = \pi_{t-1}\mathbf{M}$ to update the vector of probability;
12. **end for**
13. generate the ranking (R) of node influences by sorting π_T in ascending order;
14. **return** R ;

ALGORITHM 1: PARW-Rank Algorithm.

node with a lower value means it has a better rank position. Therefore, the rank of seven nodes with respect to the influences is $R = (3, 1, 6, 4, 2, 5, 7)$, i.e., $3 \rightarrow 1 \rightarrow 6 \rightarrow 4 \rightarrow 2 \rightarrow 5 \rightarrow 7$. With reference to the social network in Figure 2, the above ranking result reasonably reflects the influences or importance of nodes in it.

3.3. Algorithm Description. Based on the above technical framework and example illustration, here we further address the rank algorithm for node influences based on preference analysis and random walk (here abbreviated as PARW-Rank (Algorithm 1)).

The algorithm takes three basic measures and step number of random walk iteration as the input data and outputs the sorted sequence of nodes about their influences. In lines 1–6, the partial preference graph is generated according to the preference relation in each basic measure. For three basic measures C_D , C_B , and C_C , the corresponding *PPGs* are represented in the form of matrixes, that is, $\mathbf{M}_D$, $\mathbf{M}_B$, and $\mathbf{M}_C$, respectively. Subsequently, three *PPGs* are combined together to form a comprehensive preference graph (on line 7). In fact, this mergence procedure can be implemented by matrix operations. The matrix of *CPG* (i.e., $\mathbf{M}_{sum}$) can be viewed as the weighted sum of $\mathbf{M}_D$, $\mathbf{M}_B$, and $\mathbf{M}_C$. In this paper, the weights of three basic measures are equal to each other. Of course, the *CPG* (or $\mathbf{M}_{sum}$) should be regularized before the application of random walk (line 8). During the process of random walk, an initial vector π_0 is prepared firstly. In this vector, the probability of each node is set to $1/n$, where n is the total number of nodes in the given social network. Subsequently, the iterations of random walk are performed on the regularized *CPG* (in lines 10–12). Finally, on line 13, the rank of nodes (R) is achieved by sorting the probability vector, which is output from the random walk procedure.

TABLE 3: The basic statistical features of six real-life networks.

Network	n	e	$\langle k \rangle$	k_{max}
ARPA	21	26	2.48	4
ChenNet	23	40	3.48	8
Karate	34	78	4.47	15
PolBooks	105	441	8.4	25
Airlines	235	1297	11.04	130
Email	1133	5451	9.62	71

4. Experimental Analysis

4.1. Experimental Data and Setup. In order to validate the effectiveness of our proposed algorithm for evaluating node influences, six public social networks are adopted in the experimental analysis. The basic features of these networks are shown in Table 3, where n and e represent the node number and edge number in the network, respectively. In addition, k and k_{max} are the average and maximum degrees of node, respectively. These networks are available on two public dataset websites (<http://wiki.gephi.org/index.php/datasets>, <http://www.cs.bris.ac.uk/~steve/networks/peacockpaper>).

The ARPA (Advanced Research Projects Agency) network [33] is a distributed computer network system, in which there are 21 computer terminals and 26 links between them. The second network (denoted as ChenNet here) is provided by Chen et al. in [11]. Their work is also on the topic of evaluating node influences, so the network is suitable for the experimental analysis in the paper. Karate is the well-known Zachary karate club network. The network captures 34 members of a karate club and contains 78 pairwise links between members of the club [34]. PolBooks (with 105 nodes and 441 edges) is a network of books about US politics published around the 2004 presidential election and sold by the

online bookseller Amazon. The edges in the network represent the frequent copurchasing of books by the same buyer. The Airlines dataset is a network (with 235 nodes and 1297 edges) of US domestic airline traffic, in which nodes represent the airports and edges are the airline routes. The Email network is generated using email data from a large European research institution. It has 1133 nodes and 5451 edges in total.

In order to perform comparative analysis, our algorithm and other three algorithms based on basic measures were all implemented in Java programming language on the Eclipse platform with JDK1.7. The experiments were employed on an Intel Core i5 CPU 3.2 GHz machine with 4 GB RAM running Windows 7.

4.2. SIR Model. To verify the correctness and rationality of our algorithm, it needs a reference rank of nodes about their influences to perform the evaluation. Here, we also use the results of the Susceptible-Infected-Recovered (SIR) epidemic model [35] as the expected rank for evaluation. In the SIR model, each node is in one of three statuses, i.e., susceptible (S), infected (I), and recovered (R). During the epidemic spreading on networks, set S contains the individuals (or nodes) susceptible (not yet infected) to the disease; set I includes the nodes that have been infected and are able to spread the disease to susceptible individuals; and set R contains the nodes that have been recovered and will never be infected again.

At each step, for each infected node, one of its susceptible neighbors will be randomly infected with probability α (in our experiments, we set $\alpha = 1$). At the same time, every infected node has a chance (β) to be recovered. Once the node is recovered, it will not be infected again and no longer infect other susceptible nodes. In our work, the recovering probability $\beta = 1/\langle k \rangle$, where $\langle k \rangle$ is the average degree of a given network. The spreading process will terminate if there is not any infected node in the network.

For each time of simulation, the total number of infected and recovered nodes of a given initially infected node can be counted. After T_{rep} repeated trials, the influences of each initial node can be collected, and the expected rank of all nodes in the network can be further achieved. In our experiments, T_{rep} is set to 1000 to calculate the statistical influences of each node.

4.3. Evaluation Metrics. While evaluating the node influences, both our proposed algorithm and three basic methods produce a ranked list of nodes. Hence, the precision of influence evaluation should be measured by analyzing the similarity between the generated rank and the SIR model-based simulation result. Here, we refer to two metrics in the field of information retrieval to show the precision of each evaluation method.

Suppose R is the ranked list generated by an influence evaluation method, and $\hat{R}$ is the expected rank list of nodes according to the SIR model, the $P@k$ metric for ranking problem can be redefined as follows.

$$P@k(R, \hat{R}) = \frac{|R(k) \cap \hat{R}(k)|}{k}, \quad (8)$$

where $R(k)$ is the sublist of the former k elements of R , $|R(k) \cap \hat{R}(k)|$ returns the size of the common elements appearing in both $R(k)$ and $\hat{R}(k)$, and k can be set to a value within the range from 1 to the length of the ranked list.

Further, the *average precision* (AP) can be defined as the average of $P@k$, that is,

$$AP(R, \hat{R}) = \frac{\sum_{k=1}^n P@k(R, \hat{R})}{n}, \quad (9)$$

where n is the total number of elements in the ranked list. For the problem in this paper, it refers to the number of nodes in the given social network.

Besides the measure of AP, we also adopt the *Kendall rank correlation coefficient* (KRCC) to judge the consistency of two ranked lists, i.e., the generated rank and the expected rank.

$$KRCC(R, \hat{R}) = \frac{n_c - n_d}{n(n-1)/2}, \quad (10)$$

where n_c is the number of concordant pairs between the two ranked lists, and n_d is the number of discordant pairs—if the preference relation of the two nodes in the pair is consistent for both lists, it is called a concordant pair, otherwise it is discordant.

In our experiments, the rank of the SIR model-based simulation was used as the expected result, and the above two metrics between the ranked list of each method and the expected rank were calculated, respectively.

4.4. Experimental Results. The results of two relatively simple networks are addressed firstly in details, and then the ranking results of the other four networks are discussed. Finally, the effectiveness of our proposed evaluation algorithm for node influences is summarized.

4.4.1. The Result of the ARPA Network. For the ARPA network shown in Figure 6, the ranking results of both three basic methods and our algorithm are all generated and illustrated in Table 4. To facilitate the comparison, the expected rank based on the SIR model is also listed here.

Based on the results shown in Table 4, we can find that both the C_C -based method and our PARW-Rank algorithm can achieve the highest AP value (i.e., 0.73). For the other metric, KRCC, although the corresponding value of the C_C -based method is the best (0.59), the result on such metric of our algorithm is still higher than those of the C_D -based and C_B -based methods.

Briefly speaking, the result of the PARW-Rank algorithm is as good as or worse than that of the C_C -based method for two metrics, respectively. However, its performance is obviously better than those of the other two basic methods for the ARPA network. Therefore, PARW-Rank can still be considered as an effective method for evaluating node influences.

4.4.2. The Result of the ChenNet Network. The network in Figure 7 is designed by Chen et al. in [11]; here, we use it as a benchmark network to evaluate the effect of node influence

FIGURE 6: Topological structure of the ARPA network.

TABLE 4: The ranking results and precisions of four algorithms and SIR-model for the ARPA network.

Algorithm	Ranking result	AP	KRCC
SIR model	3, 14, 2, 15, 17, 19, 12, 16, 18, 13, 1, 4, 6, 20, 5, 11, 21, 7, 10, 8, 9	—	—
Degree (C_D)	2, 3, 14, 6, 12, 15, 19, 1, 4, 5, 7, 8, 9, 10, 11, 13, 16, 17, 18, 20, 21	0.71	0.36
Betweenness (C_B)	3, 12, 19, 6, 4, 14, 13, 5, 11, 2, 18, 10, 7, 20, 9, 21, 8, 17, 15, 16, 1	0.64	0.24
Closeness (C_C)	3, 19, 12, 18, 4, 13, 14, 17, 2, 20, 5, 6, 11, 15, 16, 21, 1, 7, 10, 9, 8	0.73	0.59
PARW-Rank	3, 12, 19, 14, 2, 6, 4, 13, 18, 5, 20, 11, 21, 10, 7, 17, 15, 16, 9, 1, 8	0.73	0.45

FIGURE 7: The network referred from literature [11].

TABLE 5: The ranking results and precisions of four algorithms and SIR model for the ChenNet network.

Algorithm	Ranking result	AP	KRCC
SIR model	23, 11, 22, 18, 16, 20, 17, 14, 15, 13, 12, 21, 10, 19, 1, 6, 8, 3, 4, 7, 2, 9, 5	—	—
Degree (C_D)	1, 23, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 3, 8, 10, 19, 2, 4, 6, 7, 9, 5	0.73	0.54
Betweenness (C_B)	1, 10, 6, 23, 11, 22, 21, 20, 12, 14, 16, 18, 17, 15, 13, 19, 3, 8, 2, 4, 5, 7, 9	0.64	0.49
Closeness (C_C)	10, 23, 11, 6, 22, 1, 20, 12, 14, 21, 16, 15, 17, 18, 13, 19, 3, 8, 2, 4, 7, 9, 5	0.73	0.55
PARW-Rank	23, 1, 11, 22, 20, 21, 12, 14, 10, 16, 18, 17, 15, 13, 19, 6, 3, 8, 2, 4, 7, 9, 5	0.82	0.68

ranking algorithms. Taking the rank generated by the SIR model as the expected result (refers to Table 5), our PARW-Rank algorithm obviously outperforms the other three basic methods in the metrics of both AP and KRCC.

For the metric of AP, PARW-Rank's value is 0.82 and is the best one in all four methods. The APs of the C_D -based and C_C -based methods are both 0.73, so the performances

of these two methods for ranking node influences are lower than that of our algorithm. The worst is the C_B -based method; the corresponding AP value is only 0.64.

While considering the other metric, i.e., KRCC, the preference order about ranking precisions of all four methods is generally similar to the case of metric AP. Our PARW-Rank algorithm still wins the best result (i.e., 0.68), and the

TABLE 6: The ranking results and precisions for other four networks.

Network	Algorithm	Ranking result (Top 10 nodes)	AP	KRCC
Karate	Degree (C_D)	1, 34, 33, 3, 2, 4, 31, 9, 14, 23, ...	0.70	0.43
	Betweenness (C_B)	1, 34, 33, 3, 31, 9, 2, 14, 19, 6, ...	0.67	0.40
	Closeness (C_C)	1, 3, 31, 9, 34, 14, 33, 19, 4, 30, ...	0.72	0.42
	PARW-Rank	1, 3, 34, 33, 31, 9, 14, 2, 4, 32, ...	0.72	0.48
PolBooks	Degree (C_D)	8, 12, 3, 87, 68, 70, 77, 29, 11, 38, ...	0.54	0.08
	Betweenness (C_B)	29, 52, 9, 12, 68, 80, 3, 59, 8, 7, ...	0.54	0.06
	Closeness (C_C)	29, 59, 7, 52, 9, 80, 14, 68, 4, 32, ...	0.53	0.06
	PARW-Rank	29, 68, 9, 12, 8, 3, 70, 59, 80, 87, ...	0.55	0.09
Airlines	Degree (C_D)	137, 51, 81, 131, 71, 42, 155, 85, 201, 193, ...	0.60	0.36
	Betweenness (C_B)	137, 51, 81, 201, 131, 71, 19, 174, 42, 119, ...	0.60	0.30
	Closeness (C_C)	137, 51, 81, 131, 71, 42, 155, 201, 85, 193, ...	0.59	0.33
	PARW-Rank	137, 51, 81, 131, 71, 201, 42, 119, 193, 155, ...	0.61	0.36
Email	Degree (C_D)	105, 333, 16, 23, 42, 41, 196, 233, 21, 76, ...	0.56	0.22
	Betweenness (C_B)	333, 105, 23, 578, 76, 233, 135, 41, 355, 42, ...	0.55	0.20
	Closeness (C_C)	333, 23, 105, 42, 41, 76, 233, 52, 135, 378, ...	0.56	0.21
	PARW-Rank	333, 105, 23, 42, 41, 233, 76, 135, 134, 52, 378, ...	0.56	0.22

results of the C_C -based and C_D -based methods are 0.55 and 0.54, respectively. The worst one, the result of the C_B -based method, is only 0.49.

In summary, for the network (denoted as ChenNet) in reference [11], our comprehensive ranking algorithm can generate a more precise result than the other three basic methods for evaluating node influences in a network.

4.4.3. The Results of the Other Four Networks. Similarly, the comparison is also performed on the remaining four social networks and the corresponding results are listed in Table 6.

For network Karate, the APs of our PARW-Rank algorithm and the C_C -based method are both 0.72, and the corresponding values of the C_D -based method and C_B -based method are 0.70 and 0.67, respectively. Obviously, the dominance relation of these four methods with respect to AP is $\text{PARW-Rank} \sim C_C > C_D > C_B$, where symbol “ $\sim$ ” means two methods are comparable and “ $>$ ” represents a superiority relation (i.e., be better than). While considering the other metric (i.e., KRCC), the best value is the result (0.48) of our PARW-Rank algorithm, and the worst is the result of the C_B -based method, i.e., 0.40. In addition, the KRCCs of the C_D -based method and C_C -based method are 0.43 and 0.42, respectively. Hence, the order of four methods with respect to KRCC is $\text{PARW-Rank} > C_D > C_C > C_B$.

For network PolBooks, the best method for evaluating node influences is our PARW-Rank algorithm. The metrics AP and KRCC of this algorithm are 0.55 and 0.09, respectively. The second one is the C_D -based method, whose AP and KRCC are 0.54 and 0.08, respectively. The next is the C_B -based method, and its AP is the same as that of the C_D -based method, but its KRCC is only 0.06. Although the KRCC of the C_C -based method is also 0.06, its AP is the lowest one in all four methods, that is, only 0.53. Therefore, the dominance relation of four methods can be expressed as $\text{PARW-Rank} > C_D > C_B > C_C$.

For the third network, i.e., Airlines, the best one is still the PARW-Rank algorithm. Two evaluation metrics (i.e., AP and KRCC) of this algorithm are 0.61 and 0.36, respectively. Although the KRCC of the C_D -based method is also 0.36, its AP is 0.60 and is worse than that of PARW-Rank. For metric AP, the corresponding values of the other two methods (C_B -based method and C_C -based method) are 0.60 and 0.59, respectively. Thus, the order of the four methods with respect to this metric is $\text{PARW-Rank} > C_D > C_B > C_C$. For metric KRCC, the values of the C_B -based and C_C -based methods are 0.30 and 0.33, respectively. Therefore, the dominance relation for this metric is $\text{PARW-Rank} \sim C_D > C_C > C_B$.

For the last network Email, the APs of PARW-Rank, the C_D -based method, and the C_C -based method are the same, i.e., 0.56. The corresponding value of the C_B -based method is 0.55. Hence, for metric AP, the relation of these four methods is $\text{PARW-Rank} \sim C_D \sim C_C > C_B$. With regard to the other metric (KRCC), the values of PARW-Rank and the C_D -based method are both 0.22, and the C_C -based method's value is 0.21. The worst is the C_B -based method, and its value is only 0.20. It is easy to see that the order of these methods about KRCC is $\text{PARW-Rank} \sim C_D > C_C > C_B$.

4.4.4. Summary on Experimental Results. Based on the above results, we can summarize the dominance relation of the four methods and demonstrate the results in Table 7. In most cases, PARW-Rank can win the first place or share the first place with C_D or C_C for precisely identifying influential nodes in a social network. The only one exception appears in network ARPA for metric KRCC. In such case, PARW-Rank is worse than the C_C -based method, but is still better than the C_D -based and C_B -based methods.

Specifically, for the metric of AP, the PARW-Rank algorithm is better than the other three methods for networks ChenNet, PolBooks, and Airlines. At the same time, PARW-

TABLE 7: The summary of the effects of four methods for ranking node influences.

Network	Dominance relation of four methods	
	AP	KRCC
ARPA	$\text{PARW-Rank} \sim C_C > C_D > C_B$	$C_C > \text{PARW-Rank} > C_D > C_B$
ChenNet	$\text{PARW-Rank} > C_C \sim C_D > C_B$	$\text{PARW-Rank} > C_C > C_D > C_B$
Karate	$\text{PARW-Rank} \sim C_C > C_D > C_B$	$\text{PARW-Rank} > C_D > C_C > C_B$
PolBooks	$\text{PARW-Rank} > C_D \sim C_B > C_C$	$\text{PARW-Rank} > C_D > C_B > C_C$
Airlines	$\text{PARW-Rank} > C_D \sim C_B > C_C$	$\text{PARW-Rank} \sim C_D > C_C > C_B$
Email	$\text{PARW-Rank} \sim C_D \sim C_C > C_B$	$\text{PARW-Rank} \sim C_D > C_C > C_B$

Rank is comparable to the C_C -based method for networks ARPA and Karate and is also comparable to the C_D -based method in the case of network Email.

For the other metric issue (KRCC), except of the above-mentioned exception case, the PARW-Rank algorithm outperforms the other three methods for networks ChenNet, Karate, and PolBooks. For the rest two networks, PARW-Rank is comparable to the C_D -based method.

Based on the experimental analysis of the real-life six networks, we can conclude that our comprehensive ranking algorithm (PARW-Rank) achieves the better effectiveness than do the three basic methods for evaluating node influences in a social network.

5. Discussions

5.1. Threats to Validity. Threats to construct validity regard the relation between theory and observation. In this study, we focus on the design of a new algorithm for comprehensively evaluating the influences of nodes in a social network. The SIR epidemic model [35] was used to generate the reference rank of nodes. Some other epidemic models, such as SI and SIS [36], have also been adopted for the influential node identification. Thus, the use of them may give different results. Although the average precision (AP) and Kendall rank correlation coefficient (KRCC) are two well-known evaluation metrics for comparing two different ranked sequences, there are still some other metrics available for this purpose. Based on these metrics, the experimental results have the potential to change.

Threats to external validity regard the generalization of our results in other situations. As mentioned in Section 3.1, although only three basic centrality measures are used in our framework to rank influential nodes, the measures in the framework can be flexibly replaced. When applying other basic measures into our algorithm, the experimental results have not been deeply investigated in the current research. On the other hand, six public social networks are adopted in the experiments to validate the effectiveness of our proposed algorithm. Having more social networks, especially the networks with larger sizes, can strengthen the scalability of the algorithm.

Threats to internal validity regard factors that could influence our experimental results. We have carefully inspected the implementation code of our algorithm to ensure the reliability of experimental results. In this study,

we treat the three basic measures equally in the algorithm. In fact, each of these basic measures may play a different role in identifying the influential nodes. Assigning different weights to them may produce different results.

5.2. Potential Extension of the Algorithm. As pointed out in the above subsection, our algorithm faces a potential threat in scalability. The threat comes mainly from two aspects: one is the problem of computation overhead, and the other is the robustness of the computation result.

In our algorithm, both PPG and CPG are represented by the matrix. For a large-sized social network, the corresponding matrix of PPG or CPG has a large dimensional number accordingly. In general, a matrix with high dimensions will lead to the heavy computation overhead about matrix manipulation. Since the subsequent random walk is performed on the matrix of CPG, the computation cost will obviously increase if the size of the social network becomes large. To ensure the lightweight computation in our algorithm, it is necessary to build the reduced versions of PPG and CPG for the large-sized social network.

As shown in (4), for a social network with n nodes, the final distribution vector can be represented as $\pi_t = [p_t(1), p_t(2), \dots, p_t(n)]$. Obviously, for any two elements $p_t(i)$ and $p_t(j)$ in π_t , the difference between them becomes narrow as the size of the social network (i.e., n) increases. When the size becomes very large, the corresponding difference will be very small. As a consequence, the ranking result will become very sensitive. From this perspective, it is also necessary to effectively control the size of CPG.

Here, we provide a preliminary solution for large-scale networks as follows. Suppose f basic measures are adopted in the algorithm, then for each measure $m_i (1 \leq i \leq f)$, we can select the top k nodes from the given social network. Here, the set of the partial nodes is denoted as $S_{\text{top}k}(m_i)$. The top k can set as a ratio of the size of the social network in practice, such as 5% or 10%. Then, for all measures, the union set of $S_{\text{top}k}(m_i) (1 \leq i \leq f)$ can be calculated as $S_{\text{top}k} = \bigcup_{i=1}^f S_{\text{top}k}(m_i)$. Subsequently, the PPG can be built based on $S_{\text{top}k}$ for each basic measure, and the final reduced CPG can be generated accordingly. Obviously, the cardinality of $S_{\text{top}k}$ is far less than the size of the social network. Therefore, the influential node identification based on the reduced CPG can save a lot of computation cost.

Meanwhile, the approximate treatment will not cause a significant impact on identifying the most influential nodes.

6. Related Work

Nowadays, the Internet has been applied to all aspects of our lives. Accordingly, the interactions between individuals on the Internet are becoming more and more frequent and plentiful, that is, the so-called *social network* [1]. Since it plays a great role in the economic, social, and even security activities in the real world, it is very necessary to understand the mechanisms such as community, evolution, and information propagation behind the social network.

At the earlier stage, the research concerns mainly focused on the static features and structures of social network [6]. For example, the *power-law distribution* of node degree was discovered as an important property for most networks [37]. On the other side, the distance between each node pair was also analyzed. The results have exhibited a phenomenon of *small world*, which means the individuals in a network can reach to each other in relatively few steps [38]. Furthermore, the effect of node clustering was measured by the metrics such as the clustering coefficient. Meanwhile, some algorithms about community discovery were proposed to understand the connection strength between nodes in a network [39, 40].

Besides the above static features, the dynamic issues, such as network evolution, information diffusion, and cascading failure, can help researchers to better explore the rules behind social networks. In recent years, the problem of identifying influential nodes has attracted wide attention [41–43]. The influence of a node is usually reflected by the ability of spreading information. Remarkably, *centrality* has been viewed as an important indicator for information spreading, and quite a few centralities have been defined [44]. Since the nodes with higher degree usually have stronger ability of information spreading, the degree of node is viewed as a centrality (i.e., *degree centrality* [16]) to characterize the importance of node in a network. The significant advantage of this measure lies in that it can be obtained easily. However, it takes only one step into consideration for evaluating the information-spreading ability of a node. As a result, its precision is limited in most cases. Moreover, the enhanced versions, such as *semilocal centrality* [11] and *node diversity* [45], were also proposed for identifying influential nodes in networks. Although these measures can provide more accurate prediction, they only consider the information spreading from a local point of view. In fact, for evaluating the diffusion or propagation of information, the topological information of whole network (i.e., global metrics) should be taken into account.

Since the nodes with high betweenness often play the role of gateway in a social network, the betweenness is viewed as an important indicator for measuring the information-spreading ability of a node. Thus, the *betweenness centrality* [16, 17] was defined for identifying influential nodes in the past. However, this metric only considers the bridging function of the node, but fails to consider its outward diffusion capability. In order to consider the diffusion speed of

information, the *closeness centrality* [16], which is defined based on the distance from the given node to other nodes, was introduced to make up for the shortcomings of betweenness centrality. In addition, some other centralities such as *Physarum centrality* [46] and *tunable path centrality* [47] were also proposed based on the paths of node pair. It is not hard to find that all above methods have to compute the paths between nodes in the networks, so the overhead is greatly higher than the local metrics such as degree centrality. On the other side, these path-based metrics may be useful to evaluate the information diffusion speed and the importance of the node, but it is not very good at describing the breadth of information spreading. In the paper, we merge the local metrics (e.g., degree centrality) and global metrics (e.g., betweenness centrality and closeness centrality) together to construct a preference relation model (i.e., comprehensive preference graph) for ranking node influences. Because our PARW-Rank algorithm takes the above three typical metrics into account, it can obtain much better performance.

As a classical algorithm for ranking Web pages, PageRank [48] has achieved great success in the applications of the search engine. Intuitively, it can be adopted for evaluating the influences of node in a social network, which has been confirmed by some previous studies [42, 49]. Furthermore, Lü et al. [50] proposed a simple variant of PageRank to identify influential spreaders in directed networks. In the improved model named LeaderRank, a ground node connected with every other node by a bidirectional link is introduced into the original network, and then the random walk process is applied to rank nodes according to their influences [51]. The experimental results show that LeaderRank can produce more stable ranking results and has the faster convergence speed than PageRank does. Although all these PageRank-based methods use the random walk algorithm, they perform the ranking on the original social networks. By contrast, our PARW-Rank algorithm applies the random walk technique on the comprehensive preference graph (CPG) rather than the original social network to rank node influences. In our algorithm, although the CPG has the same node set as the original network, the directional edges in CPG represent the preference relations, which are determined by combining three basic centralities (i.e., *degree centrality*, *betweenness centrality*, and *closeness centrality*) together. In other words, the rank of nodes in our algorithm is not directly based on the topology of the network but according to the new generated preference model.

In recent years, some comprehensive ranking methods for evaluating the influences of node have been presented. Wei et al. [52] proposed a new centrality measure for the weighted network based on the Dempster-Shafer evidence theory [53], in which the degree and strength of every node are both taken into consideration. This method only focuses on the local structure of the network. On the other hand, it is used for the weighted network, whereas our algorithm is for the basic social network. As a multiple-attribute decision-making (MADM) technique, TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) has been successfully applied to solve some typical decision-making problems [19, 20, 28, 54]. To identify influential nodes, this technique

was introduced to rank the nodes in a network according to their influences [25, 26]. Different from the random walk technique, TOPSIS belongs to the category of static ranking. Therefore, in theory, the ranking result of TOPSIS will be not as good as that of the random walk-based method. By and large, combining several centrality measures to generate a comprehensive rank is a promising solution to evaluate node influences. At the same time, our comprehensive algorithm based on preference analysis and random walk has a good theoretical basis. In addition, the experimental results confirm that it achieves much better performance than the single centrality measure does.

The influence maximization is a very relevant problem to the influential node identification. It aims at finding a subset of key users that maximize their influence spread over a social network [24, 55]. At present, a series of algorithms have been developed for this problem. Among them, there are two typical categories: one is from a community-based perspective and the other is from an evolutionary computation perspective. Since the community structure in social networks plays an important role in tracking the local spread of influence, some effective algorithms, such as INCIM [56] and CoFIM [57], have been proposed to model the influence propagation by exploiting the community structure. On the other side, some typical evolutionary algorithms, such as GA [58], SA [59], and PSO [60], are also employed to find the most influential nodes in a social network. However, both community detection and evolutionary search are time-consuming operations in general. As a consequence, the efficiency of these algorithms is a typical weakness. In addition, some other mathematical tools, such as game theory [61] and mathematical programming [62], have been used to solve the influence maximization problem. Strictly speaking, this problem has a difference from ours, but these mathematical models can provide us with potential inspiration to create a new solution for the influential node identification.

7. Conclusion

With the rapid development of Internet technology and Web media, social network, as a new communication platform, has penetrated into our lives and played an important role. It has affected all aspects of our lives, especially in the aspects of information diffusion, public sentiment analysis, and so on. Accordingly, it is very necessary to investigate the social network from both aspects of static structure and dynamic behavior. While considering the dynamic behaviors of a network, information diffusion between nodes is an important exemplification [63]. Therefore, the evaluation on node influences becomes a key and challenging problem.

In order to identify the influential nodes in a social network with high precision, a comprehensive evaluation model is proposed in the paper. In our model, three basic and representative centralities are taken into consideration. For each basic centrality measure, a partial preference graph (PPG) is built according to the preference relations of node pairs. Then, the comprehensive preference graph (CPG) is generated by merging the above three PPGs together. Thus, the linkage between two nodes in CPG can reflect the overall

preference information of three representative centralities. Subsequently, the random walk technique is performed on the CPG to rank the nodes in network according to their influences. Besides the running example, six public social networks, such as Arpa, Karate, and PolBooks, are taken as benchmarks to validate the effectiveness of our proposed evaluation algorithm. The experimental results confirm that our comprehensive algorithm based on preference relation and random walk has the obvious advantages than the three basic ranking methods.

Although our PARW-Rank algorithm has exhibited its good performance and robustness for identifying influential spreaders in a social network, there are still some valuable and interesting problems that deserve further exploration. For example, we will adapt our algorithm to rank the spreaders in the weighted social network. In addition, how to analyze the influences of nodes in a dynamic (or mobile) social network is also an attractive research topic.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61462030 and 61762040), Jiangxi Social Science Research Project (Grant No. TQ-2015-202), Natural Science Foundation of Jiangxi Province (Grant Nos. 20162BCB23036 and 20171ACB21031), Science Foundation of Jiangxi Educational Committee (Grant No. GJJ150465), and the Education Science Project of Jiangxi Province (Grant No. YB-2015-026).

References

- [1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, London, UK, 1994.
- [2] J. Scott, *Social Network Analysis: a Handbook*, Sage Publication, London, 2nd edition, 2002.
- [3] Tech Crunch, "Facebook now has 2 billion monthly users," June 2017, <http://techcrunch.com/2017/06/27/facebook-2-billion-users/>.
- [4] Tencent, "Tencent announces 2017 third quarter results," November 2017, <http://www.tencent.com/en-us/articles/15000651510741924.pdf>.
- [5] K. Musial, P. Bródka, and P. de Meo, "Analysis and applications of complex social networks," *Complexity*, vol. 2017, Article ID 3014163, 2 pages, 2017.
- [6] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

- [7] C. Thovex and F. Trichet, "Static, dynamic and semantic dimensions: towards a multidisciplinary approach of social networks analysis," in *Knowledge Science, Engineering and Management. KSEM 2010. Lecture Notes in Computer Science*, vol. 6291, Y. Bi and M. A. Williams, Eds., Springer, Berlin, Heidelberg, 2010.
- [8] F. Bodendorf and C. Kaiser, "Detecting opinion leaders and trends in online social networks," in *Proceeding of the 2nd ACM workshop on Social web search and mining - SWSM '09*, pp. 65–68, Hongkong, China, November 2009.
- [9] X. Liu, C. Liu, and X. Zeng, "Online social network emergency public event information propagation and non-linear mathematical modeling," *Complexity*, vol. 2017, Article ID 5857372, 7 pages, 2017.
- [10] A. E. Motter and Y.-C. Lai, "Cascade-based attacks on complex networks," *Physical Review E*, vol. 66, no. 6, article 065102, 2002.
- [11] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1777–1787, 2012.
- [12] M. Zhao, T. Zhou, B.-H. Wang, and W.-X. Wang, "Enhanced synchronizability by structural perturbations," *Physical Review E*, vol. 72, no. 5, article 057102, 2005.
- [13] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: a survey," *Journal of Network and Computer Applications*, vol. 106, pp. 17–32, 2018.
- [14] K. Li, L. Zhang, and H. Huang, "Social influence analysis: models, methods, and evaluation," *Engineering*, vol. 4, no. 1, pp. 40–46, 2018.
- [15] N. Sumith, B. Annappa, and S. Bhattacharya, "Influence maximization in large social networks: heuristics, models and parameters," *Future Generation Computer Systems*, vol. 89, pp. 777–790, 2018.
- [16] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978–1979.
- [17] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
- [18] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, USA, 1992.
- [19] M. Jalayer, M. Azheian, and M. A. M. A. Kermani, "A hybrid algorithm based on community detection and multi attribute decision making for influence maximization," *Computers and Industrial Engineering*, vol. 120, pp. 234–250, 2018.
- [20] A. Zareie, A. Sheikahmadi, and K. Khamforoosh, "Influence maximization in social networks based on TOPSIS," *Expert Systems with Applications*, vol. 108, pp. 96–107, 2018.
- [21] Y.-Y. Ko, K.-J. Cho, and S.-W. Kim, "Efficient and effective influence maximization in social networks: a hybrid-approach," *Information Sciences*, vol. 465, pp. 144–161, 2018.
- [22] X.-Y. Zhao, B. Huang, M. Tang, H. F. Zhang, and D.-B. Chen, "Identifying effective multiple spreaders by coloring complex networks," *EPL (Europhysics Letters)*, vol. 108, no. 6, article 68005, 2014.
- [23] I. Mesgari, M. A. M. A. Kermani, R. Hanneman, and A. Aliahmadi, "Identifying key nodes in social networks using multi-criteria decision-making tools," *Mathematical Technology of Networks*, vol. 128, pp. 137–150, 2015.
- [24] T. Bian, J. Hu, and Y. Deng, "Identifying influential nodes in complex networks based on AHP," *Physica A: Statistical Mechanics and its Applications*, vol. 479, pp. 422–436, 2017.
- [25] H. Yu, Z. Liu, and Y.-J. Li, "Key nodes in complex networks identified by multi-attribute decision-making method," *Acta Physica Sinica*, vol. 62, no. 2, article 020204, 2013.
- [26] Y. Du, C. Gao, Y. Hu, S. Mahadevan, and Y. Deng, "A new method of identifying influential nodes in complex networks based on TOPSIS," *Physica A: Statistical Mechanics and its Applications*, vol. 399, pp. 57–69, 2014.
- [27] C. Kang, S. Kraus, C. Molinaro, F. Spezzano, and V. S. Subrahmanian, "Diffusion centrality: a paradigm to maximize spread in social networks," *Artificial Intelligence*, vol. 239, pp. 70–96, 2016.
- [28] M. A. M. A. Kermani, A. Badiie, A. Aliahmadi, M. Ghazanfari, and H. Kalantari, "Introducing a procedure for developing a novel centrality measure (sociability centrality) for social networks using TOPSIS method and genetic algorithm," *Computers in Human Behavior*, vol. 56, pp. 295–305, 2016.
- [29] C. Salavati, A. Abdollahpouri, and Z. Manbari, "BridgeRank: a novel fast centrality measure based on local structure of the network," *Physica A: Statistical Mechanics and its Application*, vol. 496, pp. 635–653, 2018.
- [30] L. Lovász, "Random walks on graphs: a survey," in *Combinatorics, Paul Erdős is Eighty (Volume 2)*, pp. 1–46, János Bolyai Mathematical Society, Budapest, Hungary, 1993.
- [31] N. N. Liu and Q. Yang, "Eigen Rank: a ranking-oriented approach to collaborative filtering," in *Proceeding of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pp. 83–90, Singapore, Singapore, July 2008.
- [32] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, California, 3rd edition, 2011.
- [33] W. Fan, Z. Liu, and P. Hu, "Identifying node importance based on information entropy in complex networks," *Physica Scripta*, vol. 88, no. 6, article 065201, 2013.
- [34] A.-L. Barabási, *Network Science, Chapter 9*, Cambridge University Press, UK, 2015.
- [35] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.
- [36] L. J. S. Allen, "Some discrete-time SI, SIR, and SIS epidemic models," *Mathematical Biosciences*, vol. 124, no. 1, pp. 83–105, 1994.
- [37] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [38] J. Guare, *Six Degree of Separation: a Play*, Vintage Books, New York, 1990.
- [39] M. E. J. Newman, "Detecting community structure in networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.
- [40] U.-U. Narantsatsralt and S. Kang, "Social network community detection using agglomerative spectral clustering," *Complexity*, vol. 2017, Article ID 3719428, 10 pages, 2017.
- [41] M. Kitsak, L. K. Gallos, S. Havlin et al., "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.
- [42] X. Ren and L. Lü, "Review of ranking nodes in complex networks," *Chinese Science Bulletin (Chinese Version)*, vol. 59, no. 13, pp. 1175–1197, 2014.

- [43] P. Kumaran and S. Chitrakala, "A survey on influence spreader identification in online social network," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 1–7, Chennai, India, February 2016.
- [44] A. Landherr, B. Friedl, and J. Heidemann, "A critical review of centrality measures in social networks," *Business & Information Systems Engineering*, vol. 2, no. 6, pp. 371–385, 2010.
- [45] L. Liu, F. Zhu, M. Jiang, J. Han, L. Sun, and S. Yang, "Mining diversity on social media networks," *Multimedia Tools and Applications*, vol. 56, no. 1, pp. 179–205, 2012.
- [46] C. Gao, X. Lan, X. Zhang, and Y. Deng, "A bio-inspired methodology of identifying influential nodes in complex networks," *PLoS One*, vol. 8, no. 6, article e66732, 2013.
- [47] C.-L. Pu, W. Cui, and J. Yang, "Tunable path centrality: quantifying the importance of paths in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 405, pp. 267–277, 2014.
- [48] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank citation ranking: bringing order to the Web," Technical Report 1999–66, Stanford Info Lab, 1999.
- [49] Z.-L. Luo, W.-D. Cai, Y.-J. Li, and D. Peng, "A PageRank-based heuristic algorithm for influence maximization in the social network," in *Recent progress in data engineering and internet technology*, Lecture notes in Electrical Engineering, vol. 157, pp. 485–490, Springer, Berlin, Heidelberg, 2012.
- [50] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS One*, vol. 6, no. 6, article e21202, 2011.
- [51] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Physica A: Statistical Mechanics and its Applications*, vol. 404, pp. 47–55, 2014.
- [52] D. Wei, X. Deng, X. Zhang, Y. Deng, and S. Mahadevan, "Identifying influential nodes in weighted networks based on evidence theory," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 10, pp. 2564–2575, 2013.
- [53] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, 1976.
- [54] K. Yoon and C. Hwang, *Multiple Attribute Decision Making: an Introduction*, Vol. 104, Sage, 1995.
- [55] A. Bozorgi, S. Samet, J. Kwisthout, and T. Wareham, "Community-based influence maximization in social networks under a competitive linear threshold model," *Knowledge-Based Systems*, vol. 134, pp. 149–158, 2017.
- [56] A. Bozorgi, H. Haghighi, M. S. Zahedi, and M. Rezvani, "INCIM: a community-based algorithm for influence maximization problem under the linear threshold model," *Information Processing and Management*, vol. 52, no. 6, pp. 1188–1199, 2016.
- [57] J. Shang, S. Zhou, X. Li, L. Liu, and H. Wu, "CoFIM: a community-based framework for influence maximization on large-scale networks," *Knowledge-Based Systems*, vol. 117, pp. 88–100, 2017.
- [58] K. Zhang, H. Du, and M. W. Feldman, "Maximizing influence in a social network: improved results using a genetic algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 478, pp. 20–30, 2017.
- [59] D. Li, C. Wang, S. Zhang, G. Zhou, D. Chu, and C. Wu, "Positive influence maximization in signed social networks based on simulated annealing," *Neurocomputing*, vol. 260, pp. 69–78, 2017.
- [60] M. Gong, J. Yan, B. Shen, L. Ma, and Q. Cai, "Influence maximization in social networks based on discrete particle swarm optimization," *Information Sciences*, vol. 367–368, pp. 600–614, 2016.
- [61] M. A. M. A. Kermani, S. F. F. Ardestani, A. Aliahmadi, and F. Barzinpour, "A novel game theoretic approach for modeling competitive information diffusion in social networks with heterogeneous nodes," *Physica A: Statistical Mechanics and its Applications*, vol. 466, pp. 570–582, 2017.
- [62] M. A. M. A. Kermani, A. Aliahmadi, and R. Hanneman, "Optimizing the choice of influential nodes for diffusion on a social network," *International Journal of Communication Systems*, vol. 29, no. 7, pp. 1235–1250, 2016.
- [63] P. Basaras, D. Katsaros, and L. Tassioulas, "Detecting influential spreaders in complex, dynamic networks," *Computer*, vol. 46, no. 4, pp. 24–29, 2013.

Submit your manuscripts at
www.hindawi.com