Data-Driven Influential Nodes Identification in Dynamic Social Networks

Ye Qian[1] and Li Pan[1,2(✉)]

[1] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
panli@sjtu.edu.cn

[2] National Engineering Laboratory for Information Content Analysis Technology, Shanghai, China

Abstract. The identification of influential nodes in social networks has significant commercial and academic value in advertising, information management, and user behavior analysis. Previous work only studies the simple topology of the network without considering the dynamic propagation characteristics of the network, which does not fit the actual scene and hinders wide application. To solve the problem, We develop a data-driven model for the identification of influential nodes in dynamic social networks. Firstly, we introduce an influence evaluation metric BTRank based on user interaction behavior and topic relevance of the information. Combining BTRank, LH-index, and betweenness centrality, we construct a multi-scale comprehensive metric system. Secondly, in order to optimize the metric weights calculated by entropy weight method, we use simulation data to train a regression model and obtain the metric weights by Gradient Descent Algorithm. Thirdly, the weights obtained from training are used in weighted TOPSIS to sort the influence of nodes and identify influential nodes among them. Finally, We compare our model with existing models on four real-world networks. The experimental results have demonstrated significant improvement in both accuracy and effectiveness achieved by our proposed model.

Keywords: Social networks · Influential nodes · Data-driven

1 Introduction

With the development of the mobile internet, the online social network plays a vital role in information dissemination among a vast number of users. The social network can be abstractly regarded as a network topology model composed of nodes and edges, which can reflect the social relations among social individuals. Compared with other nodes, influential nodes in social networks are more likely to affect the state of nearby nodes, which makes the information spread more widely. If we identify influential nodes quickly and accurately, it will be better for the government to achieve the guidance and control of major public feeling's affairs. For business, influential nodes identification applied to precision marketing will effectively improve merchant marketing efficiency and reduce promotional costs. Therefore, the research on influential nodes has excellent theoretical and practical values.

H. Gao and X. Wang (Eds.): CollaborateCom 2021, LNICST 406, pp. 592–607, 2021.
https://doi.org/10.1007/978-3-030-92635-9_34

Previous work proposes a series of models to measure the influence of nodes, which can be summarized as network structure-based methods, topic-based methods, transfer entropy-based methods, etc. [1]. Although these models provide detailed calculation methods for node influence evaluation, all of them use a single metric to measure the influence, which cannot comprehensively and accurately measure the influence of nodes. More recently, researchers regard identifying influential nodes in complex networks as a multi-attribute decision-making problem. Considering the uncertain information fusion, text variables, and other factors, some mathematical tools such as evidence theory and fuzzy set theory are also applied to the identification of influential nodes [2]. According to the similarity between the value of each metric and the ideal value, Liern et al. [3] propose the TOPSIS decision model. While Lu et al. [4] take degree centrality, betweenness centrality, and structural hole as the input of the decision model to measure the node's propagation influence in the complex network. However, these influential node identification models still have three main disadvantages. Firstly, the selection of metrics for the evaluation model is unreasonable. The reason is that they only focus on the metrics related to the network structure without considering the dynamic propagation characteristics of the network, such as degree centrality, betweenness centrality, closeness centrality, and other simple traditional centrality indicators. Secondly, existing models do not maximize the integration of multiple metrics that reflect various aspects of node characteristics. Some decision models assign the same weight to the input metrics, which is unreasonable. Different metrics are obtained by different algorithms, and they play different roles in the network. Other models calculate the internal correlation of indicator value to obtain weights. They only consider the relative importance of the indicators and do not consider the difference between the calculated influence of the nodes and the actual results. Thirdly, the selection of indicators for the decision model lacks theoretical support and ignores the logical derivation process.

To address the above problems, we propose a data-driven influential nodes identification model (DINI). Firstly, we propose a topic-level influence evaluation metric BTRank, which is used to describe the topics of information spread by users and the frequency of interaction between users in the time interval. BTRank is an indicator that takes the dynamic propagation characteristics of social networks into account. In order to comprehensively measure the influence of nodes, we combine BTRank, LH-index, and betweenness centrality to establish a comprehensive metric system, where LH-index reflects node's influence in the local network, and betweenness centrality reflects node's influence in the global network. Then, we propose a data-driven weight optimization algorithm to assign reasonable weights to different metrics, facilitating the integration of multiple metrics. Different from the average distribution method or entropy weight method, the data-driven weight optimization algorithm not only considers the information entropy of each metric but also adds the prior knowledge of ground truth, which optimizes the weight and makes it more reasonable and adaptive. Specifically, we first use the entropy weight method to obtain the initial weights and then set the loss function between the influence results obtained by the proposed model and the actual influence value. Finally, we use Gradient Descent Algorithm to optimize the weight of each metric gradually. We apply the optimized weights to the weighted

TOPSIS model to get the ranking of nodes' influence and select the top 1% nodes as influential nodes. The main contributions of this paper can be summarized as follows:

- We propose a topic-level influence evaluation metric BTRank, which describes the interaction behavior among users and the topic relevance of information dissemination. BTRank is a node influence evaluation metric that takes into account the dynamic propagation characteristics of networks.
- We establish a comprehensive metric system, which considers not only the static and dynamic network characteristics but also the local and global network structure, to make the quantitative analysis of social networks more accurate.
- We propose a data-driven weight optimization algorithm. After using the entropy weight method to calculate the weight of each metric, we add the prior knowledge of ground truth to establish the function of the difference between the estimated value and the actual value. Gradient Descent Algorithm is used to optimize the weights to make them more reasonable and more adaptive.
- We apply the weights obtained by the data-driven weight optimization algorithm to weighted TOPSIS and complete the identification of influential nodes. Experiments conducted on four real-world networks indicate that our proposed DINI model outperforms some baselines in both accuracy and effectiveness.

The rest of the paper is organized as follows. Section 2 reviews related work. We describe our data-driven model for identifying influential nodes in social networks in Sect. 3. Section 4 presents details and results of the experiments. Finally, we conclude our work and describe future research directions in Sect. 5.

2 Related Work

The measurement of nodes' influence in social networks is mainly used to identify influential nodes in social networks. Commonly used methods for studying the influence of nodes are network structure-based method, behavior-based method, transfer entropy-based method, and so on [1].

In the previous study, the centrality of nodes is used to measure the influence of nodes in complex networks [5]. Hu et al. [6] use the concept of degree centrality, which calculates the influence of a node with the node's out-degree. Degree centrality is intuitive and efficient. However, it only reflects the local structural characteristics of nodes and ignores the global characteristics. Considering the global characteristics, Singh et al. [7] propose betweenness centrality and closeness centrality based on a faster algorithm. The difference between the two algorithms is that the former investigates the number of shortest paths through the node, while the latter considers the average length of the shortest paths from the node to other nodes. However, in the actual scene, the scale of social networks is huge, and the computational complexity of these two kinds of centrality is high, which are not very efficient methods to identify influential nodes in social networks. Wang et al. [8] attach importance to those nodes that are more central in the network, even if their degree centrality is not high. Based on such an idea, they propose a K-shell algorithm. In the early stage, social network reflects homogeneity. In other words, the nodes with high out-degree tend to have neighbors

with high out-degree. With the gradual evolution of social networks, the number of nodes keeps increasing, and some social networks begin to reflect heterogeneity [9]. Therefore, the simple method based on network topology cannot solve the problem of identifying influential nodes in social networks.

In recent years, some scholars combine information content with network topology. Yu et al. [10] use the sum of three factors, the quality of the user's tweets, the frequency of user's tweets being forwarded, and the similarity of interests between users, as the social influence of the user. Sun et al. [11] combine the amount of information in users' text content, users' emotions, and fans' behaviors to measure user's influence. Some scholars obtain user's influence by studying the behavior of users spreading information. The greedy algorithm proposed by Ren et al. [12] solves the problem of influence transmission in the network and makes the information of users as the initial nodes spread most widely in the network. Zheng et al. [13] consider that users who log on social media frequently and the number of whose neighbor nodes is increasing should have a stronger influence. In actual social media, the user's influence is related to the topics of information [14, 15]. The EIRank algorithm proposed by Bo et al. [16] introduces the topic factor into the measurement of node's influence. EIRank analyzes the information spread by users, summarizes the topics that users are interested in, and finds out the relation between the topics. This relation is combined with the static network structure to obtain the influence of nodes.

The above methods define and measure the influence of nodes in social networks from different perspectives. However, social networks in the real world are more complex than experimental settings, including network topology, user attributes, user behavior, topics of interactive information, and so on. The process of information dissemination among nodes is affected by the above factors. Therefore, in order to make the model more applicable to real-world social networks, it is necessary to combine the network topology and social network propagation elements effectively. Based on this key point, we establish a comprehensive metric system that includes not only static and dynamic network characteristics but also local and global network structure. Moreover, the influential nodes in social networks can be identified more accurately through a more reasonable weight allocation method.

3 Data-Driven Model for Influential Nodes Identification in Social Networks

Considering the dynamic propagation characteristics of real-world social networks, we establish a data-driven influential nodes identification model (DINI), which mainly contains two modules: multi-scale comprehensive metric system and data-driven metric weight optimization, as shown in Fig. 1.

Firstly, we propose an indicator BTRank to describe the influence of nodes at the topic level based on three elements: the topics of blogs posted by users, the structure of social networks, and the interaction behavior among users. Then, we combine BTRank with LH-index and betweenness centrality, which are based on local and global network structure, respectively, to form a comprehensive multi-scale metric system. In the metric weight optimization module, we use the entropy weight method to get the initial weight

of each metric and then use the data-driven way to optimize the metric weights. Finally, the metric weights are used in the weighted TOPSIS model to obtain the node influence ranking.

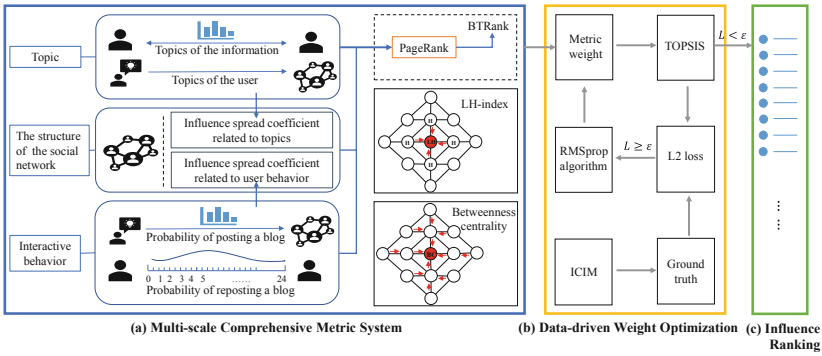

Fig. 1. The framework of our data-driven influential nodes identification model in dynamic social networks

3.1 Multi-scale Comprehensive Metric System

In order to deal with the inaccurate quantification of nodes' influence in previous models, we establish a multi-scale metric system, as shown in Fig. 1(a). The metric system consists of three metrics. The specific description of each metric is as follows:

(1) Topic-level Influence Measurement Based on User Behavior

In social networks, users are influenced by the content of information from the source users, and users also influence the spread of information from the source users. Then, We can calculate the influence of the users through mining the behavioral characteristics of users. The dissemination of information between users is mainly affected by two factors. The first factor is the user's interest in this information, which varies with users and information. Therefore, it is necessary to model the user's topics and the content of information spread among users. LDA is used to obtain the topic distribution vectors of users and information [17]. According to the distribution vectors, the probability of the user responding to other users will be calculated. The second factor is the interaction level between users. If user v often forwards blogs published by user u, it indicates that the interaction level between user u and user v is high, and user v is likely to continue forwarding blogs published by user u in the future. These two factors are critical in calculating topic-level influence based on user behavior. The influence of user u is denoted by $BR(u)$.

Given a social network $G = (V, E)$, where V represents the set of users and E represents the edges of nodes due to their interaction behavior. The topic-level influence measurement algorithm based on user behavior (BTRank) is divided into three modules:

Based on topic mining technology, the influence propagation coefficient $p(u, v)$ related to topics is calculated. Firstly, the content of blogs published by users is collected, and the unsupervised LDA model is used to obtain the topic distribution vectors of users and blogs. Each user has a topic distribution vector denoted by $\overrightarrow{p_u} = (p_u^1, p_u^2, \ldots, p_u^z)$, where p_u^i denotes the interest level of user u in the ith topic. Information is propagated through the edge, which also has a topic distribution vector denoted by $\overrightarrow{p_{uv}} = (p_{uv}^1, p_{uv}^2, \ldots, p_{uv}^z)$, where p_{uv}^k represents the kth topic's proportion of the information propagated through user u to user v. Finally, the similarity measurement method is used to calculate the influence propagation coefficient $p(u, v)$ related to topics among users.

Based on the temporal information of users publishing or forwarding blogs, the influence propagation coefficient $p(u, v)^{'}$ related to behavior is calculated. We use a statistical method to calculate the probability of users publishing or forwarding blogs in each period. Users publish and forward blogs with obvious periodicity. A day is divided into 24 disjoint time intervals, $p_u(t_i)$ represents the probability of user u publishing blogs in the ith time interval, $\widetilde{p_v}(t_j)$ represents the probability of user v forwarding blogs in the jth time interval. Then, the cumulative probability method is used to obtain influence propagation coefficient $p(u, v)^{'}$ related to behavior.

The influence propagation coefficient $p(u, v|\overrightarrow{p_{uv}})$ is calculated using formula (1). Based on the influence propagation coefficient $p(u, v|\overrightarrow{p_{uv}})$, the PageRank algorithm is used to measure the influence of each user.

$$p(u, v|\overrightarrow{p_{uv}}) = \sum_{i=1}^{z} p_v^i p_{uv}^i + \sum_{i=1}^{23} p_u(t_i) \sum_{j=i+1}^{24} \widetilde{p_v}(t_j) \tag{1}$$

In the PageRank algorithm, the propagation range of a user's information depends on the number of times the information is forwarded by other users. If user v forwards the information of user u, it can be seen as a voting process of user v to user u. User v will contribute its influence to user u with the probability of $p(u, v|\overrightarrow{p_{uv}})$. The influence of u is the accumulation of several users' influence, which is an iterative process. In each iteration, user u's influence is calculated using formula (2).

$$BR(u) = (1 - \lambda) + \lambda \sum_{v \in N(u)} p(u, v|\overrightarrow{p_{uv}}) \cdot BR(v) \tag{2}$$

where $N(u)$ represents the set of user u's neighbor users, $p(u, v|\overrightarrow{p_{uv}})$ is the probability of user v responding to user u, $BR(u)$ is the influence of user u. λ is the damping coefficient used to ensure the convergence of the calculation results, which is generally 0.85.

After calculating the influence propagation coefficient $p(u, v|\overrightarrow{p_{uv}})$ by combining the two influence propagation coefficients, we further measure the influence of users using the PageRank algorithm. We show the BTRank algorithm in Algorithm 1.

(2) Influence Measurement Based on Local Network Structure

For each node in the network, the influence of the node can be measured by calculating the influence of its neighbor nodes. The higher the influence of its neighbor nodes are, the higher the influence of the node is [18].

Algorithm 1. Influence Measurement Based on User Behavior and Topics

Input: Social network $G(V, E)$, $\overrightarrow{p_u}$, $\overrightarrow{p_{uv}}$, $p_u(t_i)$, $\widetilde{p_v}(t_j)$, damping coefficient λ
Output: User k's influence: $BR(k), k = 1, 2, \ldots, n$

```
for (u, v) ∈ E do
    Compute coefficient p(u, v) related to topics
    Compute coefficient p(u, v)' related to behavior
    Compute influence propagation coefficient p(u, v|p_uv) using formula (1)
end for
repeat
    for k ∈ V do
        BR(k) = (1 − λ) + λ Σ_{v∈N(k)} p(k, v|p_kv) · BR(v)
    end for
until Convergence
return BR(k)
```

$$LH(i) = h(i) + \sum_{v \in N(i)} h(v) \tag{3}$$

where $h(i)$ is the H-index of node i, that is, node i has at most h neighbors whose degree exceeds h. $N(i)$ is the set of node i's neighbor nodes.

(3) Influence Measurement Based on Global Network Structure

Betweenness centrality can be used to mine the nodes that play a key role in the transmission of information. The betweenness centrality of node i is denoted by $BC(i)$, which is calculated as follows:

$$BC(i) = \sum_{p \neq i \neq q}^{n} \frac{g_{pq}(i)}{g_{pq}} \tag{4}$$

where g_{pq} represents the number of shortest paths from node p to node q. $g_{pq}(i)$ is the number of paths passing through node i among the g_{pq} shortest paths.

Using the above three indicators, the metric system includes not only static and dynamic network characteristics but also local and global network structure, which makes the quantitative analysis of social networks more accurate.

3.2 Data-Driven Weight Optimization Algorithm

Most of the existing comprehensive evaluation models use uniform distribution or an entropy weight method to assign weight to each metric. Uniform distribution ignores the differences among indicators, which is not in line with the actual situation. The entropy weight method uses the information entropy of indicators to determine the weight [19], ignoring the prior knowledge of ground truth. Therefore, we propose a data-driven metric weight optimization algorithm, as shown in Fig. 1(b). Based on the weights obtained by the entropy weight method, we add the prior knowledge of ground truth to optimize

the metric weights. The influence of nodes can be evaluated more accurately with the weighted TOPSIS algorithm. The specific steps are as follows:

The three metrics are combined to generate metric matrix $D(x_{n\times 3})$. After normalization, the normalized metric matrix $D_1 = (a_{ij})$ is obtained.

$$a_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}, j = 1, 2, 3 \tag{5}$$

The entropy weight method is used to get the initial metric weights. In general, if the information entropy of a metric is small, which means that the value of the metric changes greatly, provides more information, and plays an important role in a comprehensive evaluation. Then, the weight of the corresponding metric is large. On the contrary, it indicates that the metric value does not change much, provides little information, and plays a small role in the comprehensive evaluation. Then, the weight of the corresponding indicator is small.

Calculate the weight $p_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{n} a_{ij}}}$ of the ith user under the jth indicator. The entropy value $e_j = \frac{-p_{ij} \cdot lnp_{ij}}{ln(n)}$ of the jth indicator is obtained with the weight p_{ij}. Then the entropy weight of the jth metric is calculated:

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{3} 1 - e_j} \tag{6}$$

After using the entropy method to obtain the initial weights, we add the prior knowledge of ground truth to optimize the initial weights. The optimization variable is the weights of the three indicators, which are represented by w_1, w_2 and w_3 in turn. The optimization goal is to calculate the difference between ground truth T_{gt} and nodes' influence T_{cal} obtained by the proposed model. We use the $L2$ distance to measure the difference. The optimization problem is defined as follows:

$$\begin{aligned} &\min_{w_1,w_2,w_3} L(w_1, w_2, w_3) = \|T_{gt} - T_{cal}\|_2 \\ &s.t. \begin{cases} 0 < w_1, w_2, w_3 < 1 \\ w_1 + w_2 + w_3 = 1 \end{cases} \end{aligned} \tag{7}$$

We choose the commonly used adaptive learning rate RMSprop algorithm that is one of the gradient descent algorithms [20] to solve the optimization problem. The simulation results of independent cascade model based on the influence propagation coefficient $p\left(u, v|\overrightarrow{p_{uv}}\right)$ are used to form ground truth. Along the gradient direction of the loss function, each metric is updated as follows:

$$w_i^{t+1} = w_i^t - \frac{\eta}{\sqrt{S_t + \phi}} \cdot \frac{\partial L}{\partial w_i^t}, i = 1, 2 \tag{8}$$

$$S_t = \beta S_{t-1} + (1 - \beta) \left[\frac{\partial L}{\partial w_t}\right]^2 \tag{9}$$

where S_t is the sum of the current and past square gradients, the learning rate η is 0.001, the floating-point ϕ is 10^{-7}, and the weight β is 0.9.

In Algorithm 2, we show the method to optimize each indicator's weight. After traversing all the ground truth and updating the weight, the results are the optimized metric weights. The metric weights optimized in this way learn the prior knowledge of ground truth and can be used for comprehensive evaluation more reasonably. The data-driven metric weight optimization algorithm proposed here is a metric weight optimization method that is not only applicable to node influence evaluation models but can also be extended to other comprehensive evaluation models with multiple metrics.

Algorithm 2. Metric Weight Optimization

Input: Influence metric $LH(k), BC(k), BR(k), k = 1, 2, \ldots, n$, threshold ϵ, ground truth T_{gt}
Output: Metric weights w_1, w_2, w_3

for $k = 1$ to n **do**
  Generate metric matrix $D\left(x_{k\times 3}\right) = (LH(k), BC(k), BR(k))$
end for
Normalize D to D_1 using formula (5)
Generate entropy weight vector $W = (w_1, w_2, w_3)$ using formula (6)
Compute T_{cal} by weighted TOPSIS
$t = 1$
while $\left|L^{t+1} - L^t\right| \geq \varepsilon$ **do**
  $L\left(w_1, w_2, w_3\right) = \left\|T_{gt} - T_{cal}\right\|_2$
  Update W by RMSprop algorithm
  $t = t + 1$
end while
return w_1, w_2, w_3

3.3 Influential Nodes Identification Based on Data-Driven Weighted TOPSIS

In order to integrate the above three dimensions of metrics for measuring nodes' influence, we use a weighted TOPSIS algorithm with the metric weights obtained in Sect. 3.2 to rank nodes.

Weighted TOPSIS follows the following steps to measure nodes' influence [21].

Multiply the normalized metric matrix D_1 with the weight vector to get the weighted metric matrix $D_2 = (b_{ij})$:

$$b_{ij} = w_j \cdot a_{ij}, i = 1, 2, \ldots, n; j = 1, 2, 3 \tag{10}$$

When we establish positive ideal solution A^+ and negative ideal solution A^-, the Euclidean distance S_i between node i's value and the two ideal solutions are calculated respectively. Thus, the proximity $K_i = \frac{S_i^-}{S_i^- + S_i^+}$ to the ideal solution is obtained. Node i with a large K_i value means that it is significant, and the information it posts is more widely spread in the network. Conversely, the influence is small.

Complexity analysis: Using three metrics to measure the influence of n nodes in the network takes $O(nn_\theta + nE + n^2 t(\epsilon_1))$, where n_θ denotes the maximum degree of the node, E is the number of edges, $t(\epsilon_1)$ is the number of iterations, and the number of iterations is related to the threshold ϵ_1 for convergence. It takes $O(nlog(n))$ to rank all

nodes in descending order. Combining the three metrics using weighted TOPSIS needs $O(nt(\epsilon_2))$. Where ϵ_2 is the threshold for convergence. Therefore, the worst-case time complexity of our algorithm is $O(nE)$.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets: We compare the proposed DINI model with other models on four real-world social networks, including Email network, Facebook network, Enron network, and Gowalla network, where nodes denote users and edges denote interactions between users. Several basic statistics are listed in Table 1. The table contains information about the number of nodes, the number of edges, and the average degree.

Table 1. Statistics of four real-world networks.

Network	Email	Facebook	Enron	Gowalla
$\|V\|$	1005	3483	36692	137873
$\|E\|$	16706	65536	183831	661800
Average degree	33.25	37.63	10.02	9.61

Ground Truth: We use the independent cascade model based on the influence propagation coefficient $p\,(u, v|\overrightarrow{p_{uv}})$ mentioned in Sect. 3.1 (ICIM) to simulate the information transmission process in social networks [22]. Each node activates its neighbors with probability $p(u, v|\overrightarrow{p_{uv}})$. The final result is expressed as $F(u)$, representing the total number of nodes that end up in the active state at the end of the propagation process starting from node u.

Evaluation Metric: The influence of nodes is sorted, and the top 1% nodes are selected as influential nodes. We use Spearman's rank correlation coefficient [23] to measure the accuracy of influential nodes identification. Spearman's rank correlation coefficient is defined as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{11}$$

where ρ is the rank correlation coefficient, d_i is the difference between the rankings of the same node obtained by the model and ground truth, and N is the number of influential nodes. The more accurate the influential nodes are identified, the larger Spearman's rank correlation coefficient ρ is.

Settings: There is a constant parameter ϵ in Algorithm 2, which is the threshold for convergence. We set $\epsilon = 0.002$ for Email, $\epsilon = 0.005$ for Facebook, $\epsilon = 0.135$ for Enron, $\epsilon = 0.215$ for Gowalla. Experiments were conducted on a machine with an Intel Core CPU i7-9700 at 3 GHz and 32 GB memory.

4.2 Performance Comparison

To identify influential nodes in social networks, we sort the influence results of nodes calculated by our proposed DINI model and select the top 1% nodes as influential nodes. Five typical methods are chosen to be baselines. We compare DINI with baselines, including Entropy-based ranking measure (ERM) [24], WVoteRank [25], New Evidential Centrality (NEC) [2], NLPCA [26] and TOPSIS [4]. These methods are comprehensive evaluation models, which make full use of the network structure and the information transmitted between users to measure the influence of users. The experiments are conducted on real-world social networks mentioned in Sect. 4.1 to demonstrate the rationality of weight distribution and the effectiveness of our proposed DINI model.

Influential Nodes Identification Performance. We use Spearman's rank correlation coefficient [23] to measure the accuracy of influential nodes identification by different models. Spearman's rank correlation coefficient ρ is the correlation between the model results and the simulation results, which is higher when the identification of influential nodes is more accurate. Figure 2 shows that ρ changes with the mean value α of the influence propagation coefficient $p\,(u, v|\overrightarrow{p_{uv}})$, where (a) - (d) shows the results of

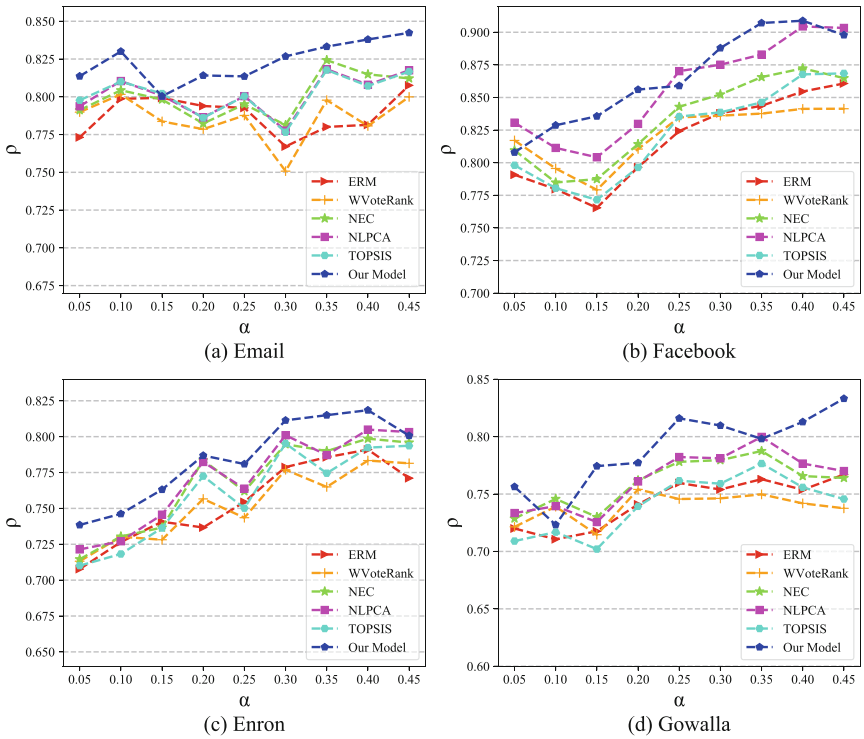

Fig. 2. Correlation between the ranking of influential nodes identified by the influence simulation model and the rankings of the corresponding nodes obtained by our proposed DINI model and other models in (a) Email, (b) Facebook, (c) Enron, (d) Gowalla

various models on four datasets correspondingly. With the increase of α, the overall accuracy of the influential nodes identified by different models trends to increase.

As can be seen from Fig. 2, the model proposed in this paper outperforms other models generally. The most apparent superiority is in (a), when $\alpha = 0.3$, where shows a 5% improvement over the second-place method NEC. Although in rare cases, such as $\alpha = 0.05$ or $\alpha = 0.25$ in (b), The accuracy of our model is 2% lower than that of NLPCA, we still have a far leading efficiency advantage, which is described in Sect. 4.2 in detail. The reason for these cases is that when the value of α is small, the information spreads slowly in the network, and the nodes in the critical position of the network can make the information spread more widely, which makes the model based on the network structure identify influential nodes more effectively. Besides, our proposed model has inherent scalability that is not available in other methods, proving by the consistent advantage across four different data sets. In summary, our model can identify influential nodes more accurately in different data sets and link the propagation characteristics of information with the node's attributes, which can better measure the node's influence.

Comparison on the Activation Capability of Influential Nodes. In general, if a node is considered as an influential node, it should be able to activate more nodes in the independent cascade model based on the influence propagation coefficient $p\left(u, v|\overrightarrow{p_{uv}}\right)$ [22]. Therefore, the top 1% of nodes proposed by each algorithm are selected as the initial

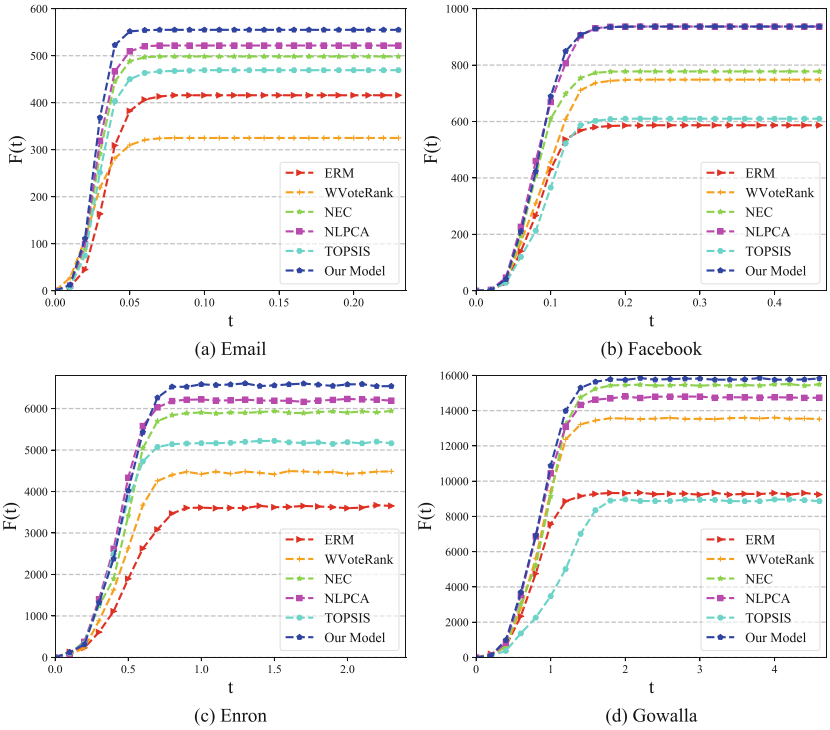

Fig. 3. Average activation capability of influential nodes

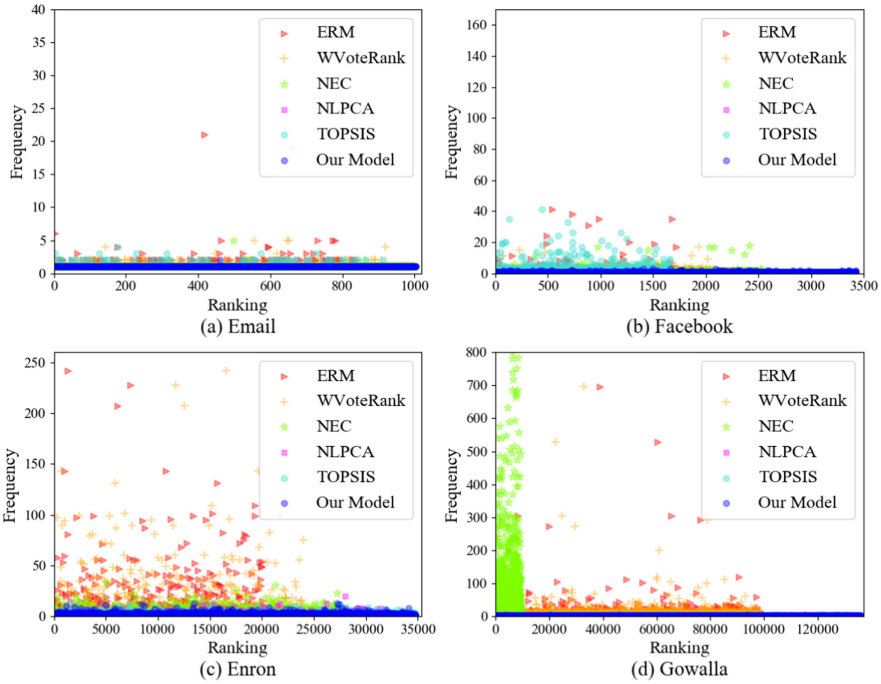

Fig. 4. Frequency of nodes with the same ranking

set of nodes for the independent cascade model based on the influence propagation coefficient. When $\alpha = 0.2$, we plot the relationship between propagation time and the average number of activated nodes, and the result is shown in Fig. 3.

As can be seen from Fig. 3, the influential nodes identified by the DINI model proposed in this paper activate the highest average number of nodes after the propagation process reaches a stable state. Obviously, in four different networks, the influential nodes identified by our proposed model have the best information dissemination and activation ability, showing higher accuracy than other models.

Comparison on the Distinction of Different Models. If the number of nodes with the same ranking is small, it proves that the model has good discrimination in measuring the influence of nodes. When $\alpha = 0.2$, Fig. 4 shows the number of nodes with the same ranking in the four networks. As can be seen from Fig. 4, the results of the DINI model proposed in this paper have the least number of nodes with the same ranking, while other models have at most 800 nodes with the same ranking. In particular, ERM and NEC have the largest number of nodes with the same ranking. The reason is that each node in a high-density network has a high probability of being in a critical position in the network, which causes ERM and NEC methods based on the network structure to be ineffective in ranking nodes. From this perspective, it can be concluded that our proposed DINI model can measure the influence of nodes more effectively.

Comparison on the Running Time of Different Models. The running time of models is a common index to measure the performance of models. When $\alpha = 0.2$, Fig. 5 depicts the running time required to select the influential nodes using six different models on four different networks. It is obvious that the running time of our DINI model is almost the same as that of the WVoteRank, and it is more efficient than NEC and NLPCA. For small networks, such as Email and Facebook, it takes no more than 10 s for our model to complete the identification of influential nodes. For Gowalla network with 137873 nodes and 661800 edges, its running time can also be controlled within 200 s. Though our model is slightly less efficient compared with ERM and WVoteRank, our model shows superiority over them in the accuracy of identifying influential nodes. Therefore, our model has a high time efficiency on the whole.

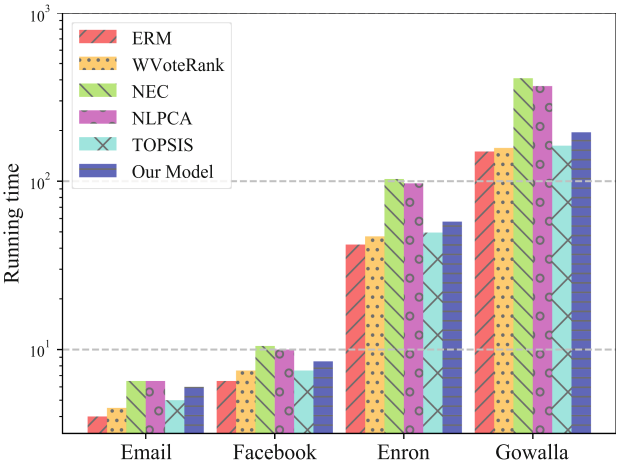

Fig. 5. Running time of six different models

5 Conclusion and Future Work

In this paper, we propose a data-driven model for influential nodes identification in dynamic social networks. To make the measurement of nodes' influence reflect the nodes' real situation in the social network, we mainly focus on three aspects, including the quantitative indicators, the metric evaluation system, and experiment settings. Firstly, we propose a topic-level indicator BTRank to quantify the influence based on user interaction behavior and topics of the information. Combining BTRank, which reflects the dynamic propagation characteristics of the network, with LH-index and betweenness centrality, which reflect the static topology of the network, we establish a multi-scale integrated evaluation metric system. Secondly, We introduce a weight optimization algorithm that computes a set of weights containing prior knowledge and apply the weights to weighted TOPSIS to identify influential nodes. Finally, to simulate the actual propagation process of information in social networks, we use an independent cascade model based on the influence propagation coefficient $p\left(u, v|\overrightarrow{p_{uv}}\right)$ in our

experiments. We do experiments on four real-world social networks and compare the proposed DINI model with five existing advanced models: Entropy-based ranking measure, WVoteRank, New Evidential Centrality, NLPCA, and TOPSIS. Besides, we also compare the effect of weighted TOPSIS and TOPSIS in terms of balancing metrics. The experimental results verify that our model can achieve better performance on both accuracy and effectiveness. In future research, we will make use of spatial and temporal characteristics of information dissemination in social networks to identify influential nodes.

Acknowledgements. This work is supported by National Key Research and Development Plan in China (2018YFC0830500), National Natural Science Foundation of China (62172278).

References

1. Hafiene, N., Karoui, W., Romdhane, L.B.: Influential nodes detection in dynamic social networks: a survey. Expert Syst. Appl. **159**, 113642 (2020)
2. Bian, T., Deng, Y.: A new evidential methodology of identifying influential nodes in complex networks. Chaos Solitons Fractals **103**, 101–110 (2017)
3. Liern, V., Pérez-Gladish, B.: Multiple criteria ranking method based on functional proximity index: un-weighted TOPSIS. Ann. Oper. Res. 1–23 (2020). https://doi.org/10.1007/s10479-020-03718-1
4. Lu, M.: Node importance evaluation based on neighborhood structure hole and improved TOPSIS. Comput. Networks **178**, 107336 (2020)
5. Martin, T., Zhang, X., Newman, M.E.: Localization and centrality in networks. Phys. Rev. E **90**(5), 052808 (2014)
6. Hu, R.J., Li, Q., Zhang, G.Y., Ma, W.C.: Centrality measures in directed fuzzy social networks. Fuzzy Inf. Eng. **7**(1), 115–128 (2015)
7. Singh, R.R., Goel, K., Iyengar, S., Gupta, S.: A faster algorithm to update betweenness centrality after node alteration. Internet Math. **11**(4–5), 403–420 (2015)
8. Wang, Z., Zhao, Y., Xi, J., Du, C.: Fast ranking influential nodes in complex networks using a k-shell iteration factor. Phys. A Stat. Mech. Appl. **461**, 171–181 (2016)
9. Lee, J.K., Choi, J., Kim, C., Kim, Y.: Social media, network heterogeneity, and opinion polarization. J. Commun. **64**(4), 702–722 (2014)
10. Yu, D., Chen, N., Ran, X.: Computational modeling of Weibo user influence based on information interactive network. Online Inf. Rev. (2016)
11. Sun, X., Xie, F.: The three-degree calculation model of microblog users' influence (short paper). In: Gao, H., Wang, X., Yin, Y., Iqbal, M. (eds.) CollaborateCom 2018. LNICST, vol. 268, pp. 151–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12981-1_10
12. Ren, T., et al.: Identifying vital nodes based on reverse greedy method. Sci. Rep. **10**(1), 1–8 (2020)
13. Zheng, Z., Gao, X., Ma, X., Chen, G.: Predicting hot events in the early period through Bayesian model for social networks. IEEE Trans. Knowl. Data Eng. (2020)
14. Riquelme, F., González-Cantergiani, P.: Measuring user influence on twitter: a survey. Inf. Process. Manag. **52**(5), 949–975 (2016)
15. Drakopoulos, G., Kanavos, A., Tsakalidis, A.K.: Evaluating twitter influence ranking with system theory. In: WEBIST (1), pp. 113–120 (2016)
16. Bo, H., McConville, R., Hong, J., Liu, W.: Social network influence ranking via embedding network interactions for user recommendation. In: Companion Proceedings of the Web Conference 2020, pp. 379–384 (2020)

17. Sapul, M.S.C., Aung, T.H., Jiamthapthaksin, R.: Trending topic discovery of twitter tweets using clustering and topic modeling algorithms. In: 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6. IEEE (2017)
18. Liu, Q., et al.: Leveraging local h-index to identify and rank influential spreaders in networks. Phys. A Stat. Mech. Appl. **512**, 379–391 (2018)
19. Chen, P.: Effects of the entropy weight on TOPSIS. Expert Syst. Appl. **168**, 114186 (2021)
20. Ning, Z., Iradukunda, H.N., Zhang, Q., Zhu, T.: Benchmarking machine learning: how fast can your algorithms go? arXiv preprint arXiv:2101.03219 (2021)
21. Yuan, B., Chang, J.E., Zhang, F.: Influential node identification method of assembly system based on TOPSIS and topology. J. Phys. Conf. Ser. **1605**, 012019 (2020)
22. Li, P., Liu, K., Li, K., Liu, J., Zhou, D.: Estimating user influence ranking in independent cascade model. Phys. A Stat. Mech. Appl. **565**, 125584 (2021)
23. Batyrshin, I.Z., Ramirez-Mejia, I., Batyrshin, I.I., Solovyev, V.: Similarity-Based correlation functions for binary data. In: Martínez-Villaseñor, L., Herrera-Alcántara, O., Ponce, H., Castro-Espinoza, F.A. (eds.) MICAI 2020. LNCS (LNAI), vol. 12469, pp. 224–233. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60887-3_20
24. Guo, C., Yang, L., Chen, X., Chen, D., Gao, H., Ma, J.: Influential nodes identification in complex networks via information entropy. Entropy **22**(2), 242 (2020)
25. Sun, H., Chen, D., He, J., Ch'ng, E.: A voting approach to uncover multiple influential spreaders on weighted networks. Phys. A Stat. Mech. Appl. **519**, 303–312 (2019)
26. Basu, S., Maulik, U.: Mining important nodes in complex networks using nonlinear PCA. In: 2017 IEEE Calcutta Conference (CALCON), pp. 469–473. IEEE (2017)