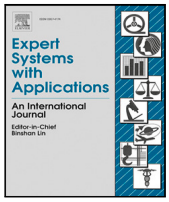
Expert Systems With Applications 235 (2024) 121154

Contents lists available at ScienceDirect

Expert Systems With Applications

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



CAGM: A communicability-based adaptive gravity model for influential nodes identification in complex networks

Guiqiong Xu, Chen Dong *

Department of Information Management, School of Management, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Keywords:
Complex networks
Influential nodes
Adaptive gravity model
Influence radius
Communicability network matrix

ABSTRACT

Identifying influential nodes is a very hot and challenging issue in the field of complex system and network. A great deal of algorithms have been developed to address the influential nodes identification problem, but most of previous studies are compromising between result accuracy and time cost. In this work, we propose a communicability-based adaptive gravity model (CAGM) for influential nodes identification. The key idea of CAGM algorithm is that the importance of each node is evaluated by comprehensively considering the influence probability and influence intensity information of neighbor nodes located in its influence radius. More specifically, the communicability network matrix is introduced to depict the influence probability between each pair of nodes. By integrating k-shell, degree and distance information, the influence radius of each node can be determined uniquely so as to portray the inherent heterogeneity of nodes in complex networks. To verify the effectiveness and applicability of CAGM, several groups of simulated experiments on twelve real and artificial datasets are concluded. Experimental results show that CAGM performs better than eight popular algorithms in terms of top-10 nodes, discrimination ability, imprecision function and ranking accuracy.

1. Introduction

In the real world, numerous complex systems can be modeled as networks of varying sizes with entities as vertices and associations between entities as edges, such as social system, power system, biological system and transportation system (Arebi et al., 2022; Berner et al., 2021; Goos et al., 2022; Lambiotte et al., 2019; Lü, Chen et al., 2016). Influential nodes identification is one of the most significant and essential research issues in network science since it is crucial to explore network structure and function in various applications. For instance, the failures of a small number of influential nodes may induce a cascading failure of majority nodes, and even lead to the collapse of the whole network (Liu et al., 2022; Wang et al., 2019). From the perspective of spreading dynamics, the spread originated from influential nodes will cover a large proportion of nodes in the whole system in a short time (Magdaci et al., 2022; Mi et al., 2018). The study on influential nodes identification attributes to the seed selection in social marketing, disease transmission and control, as well as rumor containment and so on (Dong et al., 2023; Li, Liu et al., 2021).

Up to now, a great deal of algorithms have been developed to address the problem of influential nodes identification. Among them, betweenness centrality (Freeman, 1977), k-shell decomposition centrality (Kitsak et al., 2010), degree centrality (Freeman, 1978), closeness centrality (Freeman, 1978) and H-index centrality (Lü, Zhou et al., 2016) are widely used. However, these classic algorithms merely consider part of topological characteristics and there exist some limitation in influential nodes identification. In order to further enhance the precision and reduce the computation cost, a series of novel algorithms have been put forward for finding crucial spreaders in complex networks. For instance, Wen and Deng (2020) designed the local information dimensionality algorithm for detecting influential spreaders, where Shannon entropy is employed to measure the local structural properties of central nodes in a network. Wang et al. (2022) presented an effective semi-local measure to evaluate the importance of nodes, which not only takes into account the contribution of nodes around the source node, but also considers the distribution of other nodes within optimal range. Meng et al. (2022) put forward a novel measure to evaluate the influence of edges, which makes full use of the k-shell centrality, H-index and clustering coefficient of its connection nodes.

In reality, for evaluating the influence of nodes accurately, we may consider both the information about the source node itself and its nearest neighbors, as well as the interactions between them. For example, gravity law is a classical physical rule for portraying the interaction between two objects, which core idea is that the association between two objects is inversely proportional to the square of the distance of each pair of objects and proportional to the masses between

* Correspondence to: Room 209, School of Management, Shanghai University, No. 333 Nanchen Road, Baoshan District, Shanghai 200444, China.
E-mail addresses: xugq@staff.shu.edu.cn (G. Xu), dongchen199508@shu.edu.cn (C. Dong).

https://doi.org/10.1016/j.eswa.2023.121154
Received 28 February 2023; Received in revised form 6 April 2023; Accepted 7 August 2023
Available online 23 August 2023

pairs of objects. More recently, researchers also apply gravity model in the field of network science to discover vital nodes (Shang et al., 2021; Tu et al., 2021; Yan et al., 2020). According to gravity law, entropy weight method and multiple characteristics of nodes, Yan et al. (2020) presented an effective measure to identify super communicators. Tu et al. (2021) defined a novel gravity model to rank the nodal influence, and the vital nodes are identified according to the local, global and path features of networks. Shang et al. (2021) put forward an effective model, which makes full use of both global and local characteristics of nodes according to the multi-source information fusion.

As discussed before, gravity model-based measures provide a unified framework for evaluating the importance of nodes. In the existing researches, there are two bottleneck problems that need to be addressed. On the one hand, the influence radius of nodes is always set to a fixed value, which is obviously inconsistent with the real situations. Due to the inherent heterogeneity of nodes, the influence scope of nodes is different in a given network. On the other hand, the location of nodes is also an important factor need to be considered. The mutual attraction between each pair of nodes is related with their location, and nodes in the central part of the network are much more likely to establish connection with each other than nodes located in the periphery.

To address these two questions, this work designs a novel gravity based model for detecting influential nodes in networks, called "Communicability-based Adaptive Gravity Model" (CAGM). Two core ideas are included in the CAGM algorithm by reducing time complexity while maintaining accuracy. Firstly, to solve the problem of traditional gravity model is time consuming in some large networks, we calculate the influence radius of each node by considering the gravitational coefficient and mutual attraction capability between each pair of nodes. Secondly, for depicting the inherent heterogeneity in complex networks, we introduce communicability network matrix to evaluate the influence probability that a node imposes to another through utilizing all possible paths between these two nodes. Based on the above analysis, the proposed CAGM algorithm intends to utilize influence probability and influence intensity information of neighbor nodes located in influence scope. In brief, the contributions of this work are summarized as follows:

- The proposed CAGM algorithm utilizes location information of nodes to define the gravitational coefficient between node pairs, and the influence radius of each node can be determined uniquely so as to portray the inherent heterogeneity of nodes in complex networks.
- The communicability network matrix is introduced to depict the influence probability between each pair of nodes, which measures the probability that one node establishes connection with another under all possible paths rather than the shortest path only.
- CAGM calculates the importance of nodes by integrating influence probability and influence intensity information of neighbor nodes located within the influence radius. The experimental results over twelve real and artificial datasets indicate that CAGM is superior to several existing gravity models and popular centrality measures.

The rest of this paper is organized as follows. In the next section, we briefly review some related research. In Section 3, the proposed CAGM algorithm is described in details. Next, the experimental setup is presented in Section 4, including datasets, compared algorithms, spreading model and evaluation methods. Then, Section 5 presents the experimental results and analysis. Finally, some brief conclusion and discussion of this work are given in Section 6.

2. Related work

Identifying influential nodes is an important topic in the field of network science, since it plays a crucial role in investigating network structure and function in various applications. Researchers make great efforts in influential node identification and a series of efficient algorithms have been proposed from different viewpoints. Here we roughly divide them into three categories: local-based algorithm, global-based algorithm and location-based algorithm.

2.1. Local-based algorithm

Some centrality measures evaluate the spreading capability of nodes from local structure information, which considers that the influence of a node is largely affected and reflected by the topological structure of the network it belongs to. For instance, degree centrality (Freeman, 1978) measures the nodal influence by the number of its nearest neighbors. Considering the limitation of degree centrality, Chen et al. (2012) proposed semi-local centrality by aggregating the influence of neighbor nodes within four hops, which significantly improves the accuracy of algorithm with low computational cost. Fowler and Christakis (2008) argued that in social networks, the emotion of a user can affect up to its three-order neighbors, which is line with the assumption of semi-local centrality.

In 2005, Jorge Hirsch pioneered the H-index for evaluating the achievements of scholars, which considers both quantity and quality of each scholar's scholarly output (Hirsch, 2005). Based on this, Lü and her coworkers (Lü, Zhou et al., 2016) firstly revealed the close relationship between the degree, H-index and k-shell index, and found that H-index achieves good performance in identifying influential nodes. After that, some variants and extensions of H-index have been put forward (Liu et al., 2018; Lu & Dong, 2020; Zareie & Sheikhahmadi, 2019). Specifically, Liu et al. (Liu et al., 2018) proposed the local H-index centrality to identify influential nodes according to the H-index of the node itself and its nearest neighbors. Zareie and Sheikhahmadi (2019) presented an extended H-index centrality to recognize influential spreaders, which improves the performance of traditional H-index centrality by introducing the cumulative function to distinguish the importance of neighbors with different degree. Lu and Dong (2020) designed an extended hybrid H-index centrality to quantify spreading capability of nodes by considering the neighborhood topological structure of each node.

2.2. Global-based algorithm

The original intention of the above algorithms is to evaluate the importance of nodes according to the local information of networks. In reality, the controlling capability of each node for traffic and information flow during spreading also has great implications in its importance. From the perspective of spreading dynamics, the faster and wider that one node can enable information to spread, the more important it is. Inspired by this, some methods have been put forward to rank the nodal importance on the basis of the global information of the network. Specifically, betweenness centrality (Freeman, 1977) assumes that the probability of a node locating on the shortest path of all node pairs is proportional to the importance of nodes. Closeness centrality (Freeman, 1978) measures the influence of nodes by quantifying the average shortest distance between a source node and the remaining nodes. Unlike closeness centrality, Katz centrality (Leo, 1953) takes into account both the shortest path and other possible paths between node pairs in networks to identify influential nodes.

More recently, several studies are devised for further improving the performance of traditional global-based algorithms (Ullah et al., 2021; Zareie et al., 2020; Zhong et al., 2022). For instance, Zareie et al. (2020) developed an improved cluster rank algorithm to recognize influential spreaders by considering the common hierarchy of nodes and neighborhood sets. Ullah et al. (Ullah et al., 2021) introduced a local-and-global-centrality to identify the most valuable nodes, which handles local as well as global topological aspects of a network simultaneously. Zhong et al. (2022) put forward a novel algorithm called local degree dimension, which analyzes the rising rate and declining rate of

the numbers of neighbors in each layer of central nodes to assess the propagation capability of each node.

In addition, some researches try to rank influential nodes by considering the local and global information of network simultaneously (Agneessens et al., 2017; Ibnoulouafi et al., 2018; Qiu et al., 2021; Tong et al., 2023). Agneessens et al. (2017) put forward a generalized centrality measure to evaluate the importance of nodes by considering both the information of direct neighbors and the distance between each pair of nodes. Ibnoulouafi et al. (2018) presented a multi-attribute based centrality measure, which combines the information on the position of the node in the network with the local information on its nearest neighborhood. Qiu et al. (2021) proposed an efficient measure to rank the influence of nodes, which assesses the influence of nodes by considering the degree, clustering coefficient and k-shell decomposition method. Tong et al. (2023) presented a novel recognition method to identify vital spreaders in networks, which takes into account both the local property, global information and hierarchical structure.

Furthermore, the community structure of network is pervasive in nature and the existing researches show that the network communities is crucial in explaining epidemiological patterns and identifying influential nodes (Samir et al., 2021; Tulu et al., 2017; Zhao et al., 2020). For example, Tulu et al. (2017) proposed a community-based method for important nodes identification, which considers both the number and density of communities and the degree of neighbor nodes. By combining the community structure and the closeness of networks, Zhao et al. (2020) presented a novel method for identifying influential nodes of networks. Samir and his coworkers (Samir et al., 2021) designed a scalable and fast approach to identify influential nodes, which divides the entire network into communities by using the classic Louvain algorithm, and employs k-shell decomposition algorithm to detect the seed nodes in the communities.

2.3. Location-based algorithm

In complex networks, the position information of nodes also plays a crucial role in information dissemination, while nodes with smaller degree but locating at the core of the network tend to have higher spreading capability than that nodes with larger degree at the edges. In 2010, Kitsak et al. (2010) innovatively designed the k-shell centrality to evaluate the nodal influence in networks through decomposing networks iteratively. However, the k-shell centrality is so rough that numerous nodes have the same k-shell value, while these nodes are differ in the importance of accelerating information spreading and maintaining network robustness. To fill this gap, researchers have proposed its variants or improved versions (Bae & Kim, 2014; Wang et al., 2016; Zareie & Sheikhahmadi, 2018; Zeng & Zhang, 2013). Zeng et al. (Zeng & Zhang, 2013) defined the mixed degree decomposition centrality to distinguish the influence of nodes, which takes into account the difference in the number of neighbors during each iteration. Bae and Kim (2014) argued that the influential spreaders are those whose neighbors are located near the core position of networks. Wang et al. (2016) designed the k-shell iteration factor to recognize influential spreaders, which considers the information of iteration during the decomposition process. Zareie and Sheikhahmadi (2018) designed an extended version of the k-shell centrality, where the influence of nodes is associated with the topological location of each node and its neighbors.

In the field of physics, classical gravity formula indicates that any two objects in nature are mutually attracted, and the attraction force is inversely proportional to the quadratic of distance between pairs of objects and proportional to the product of the masses. In 2016, Ma et al. (2016) first devised the gravity centrality model in influential nodes identification, which provides a unified framework in influential spreaders identification. In order to further enhance the accuracy of algorithm, several recent studies extend the gravity centrality model through redefining the mass of nodes or the shortest distance between node pairs (Namtirtha et al., 2018; Wang et al., 2018; Yang & Xiao, 2021; Zhao et al., 2022). Namtirtha et al. (2018) designed the k-shell hybrid method to measure the spreading ability of nodes in networks, which utilizes a hybrid method to calculate the mass of each node and takes into account the relative position relationship between each pair of nodes. Wang et al. (2018) put forward an improved gravitational centrality to recognize influential nodes, where the mass of focal nodes is defined via k-shell value while the mass of their neighbors is measured by the degree value. Yang and Xiao (2021) defined the attraction coefficient between node pairs by using k-shell values of nodes, and presented the k-shell-based gravity centrality to identify influential spreaders. The algorithm comprehensively takes into account the local and global features of networks and thus the accuracy is enhanced to some extent. Zhao et al. (Zhao et al., 2022) constructed a novel model for identifying vital nodes, in which the nodal influence is measured by the sum of gravity between itself and the nodes locating all the walks.

Meanwhile, the influence range of each node is also an important constraint in this model, and a series of researches have been proposed to validate this idea (Li et al., 2019; Li, Shang et al., 2021; Li & Xiao, 2021; Liu et al., 2020). For enhancing the feasibility of the traditional gravity model, Li et al. (2019) put forward an improved gravity-based model for finding influential nodes, which evaluates the nodal importance by considering the influence of each node's neighborhood within a pre-defined influence range. Similarly, Liu et al. (Liu et al., 2020) presented a novel mechanics model to identify influential spreaders, which incorporates both local and global information within the truncation radius. Li and his coworkers (Li, Shang et al., 2021) designed a generalized gravity model to identify the most valuable nodes, which comprehensively measures the local information of nodes in terms of both the clustering coefficient and degree, and the importance of each node is determined by considering the influence of neighbors within a specified radius. More recently, Li and Xiao (2021) introduced a precise influence radius and designed an effective gravity model to recognize influential spreaders, which considers the degree of a node and its neighbors' degree distribution information in the network.

3. Proposed algorithm

As mentioned before, there are still two bottleneck problems in the existing gravity model-based algorithms. On the one hand, the influence radius of all nodes is always set to a fixed value, which is obviously inconsistent with the inherent heterogeneity. On the other hand, the location of nodes is also an important factor in influential nodes identification, but it is neglected in the existing gravity model measures.

To address these two problems, this work proposes the communicability-based adaptive gravity model (CAGM) to search for influential nodes, which is composed of three phases. In the first phase, to achieve good balance between algorithm accuracy and time cost, we estimate the influence range of each node by comprehensively considering the location difference and mutual attraction between each pair of nodes. In the second phase, we introduce the communicability network matrix to portray the influence probability between each two nodes, which can well depict the possibility of information propagation between node pairs by considering all possible paths. In the third phase, the communicability-based adaptive gravity model is constructed to identify influential nodes, which takes into account both influence probability and influence intensity information of neighbor nodes located in influence scope. The process of proposed CAGM algorithm is schematically shown in Fig. 1.

Let $G = (V, E)$ be a simple network, where V and E are the node set and the edge set, respectively. Here we denote $N = |V|$ and $M = |E|$. Suppose that $A = (a_{ij})_{N\times N}$ is the adjacency matrix of G, where $a_{ij} = 1$ if the corresponding $i, j \in E$, otherwise $a_{ij} = 0$. k_i denotes the degree of node i. The proposed CAGM consists of three phases, i.e., estimating the influence radius of nodes, calculating the influence probability between node pairs, and getting the importance scores of nodes. The details of the CAGM algorithm will be described as below, while all the key parameters and symbols are presented in Table 1.

Table 1
Notations.

Parameter	Description
G	An unweighted and undirected network
V	Set of nodes in G
E	Set of edges in G
N	Number of nodes in G
M	Number of edges in G
$A = (a_{ij})_{N\times N}$	Adjacency matrix of G
k_i	Degree of node i
$GC(i)$	Gravity model centrality value of node i
$ks(i)$	K-shell value of node i
d_{ij}	Shortest distance between node i and node j
ψ_i	Influence range of node i in gravity model
$F(i,j)$	Gravitational force exerted by node i to node j
$g_{i,j}$	Gravitational coefficient between node i and node j
$Iks(i)$	Iteration k-shell value of node i
R_i	Exact influence range of node i
δ	Adjustable parameter in communicability-based adaptive gravity model
$X = (x_{ij})_{N\times N}$	Communicability network matrix of G
P_{ij}	Influence probability between node i and node j
$CAGM(i)$	Communicability-based adaptive gravity model value of node i
$\Gamma_t(i)$	The t-order neighborhood node set of node i

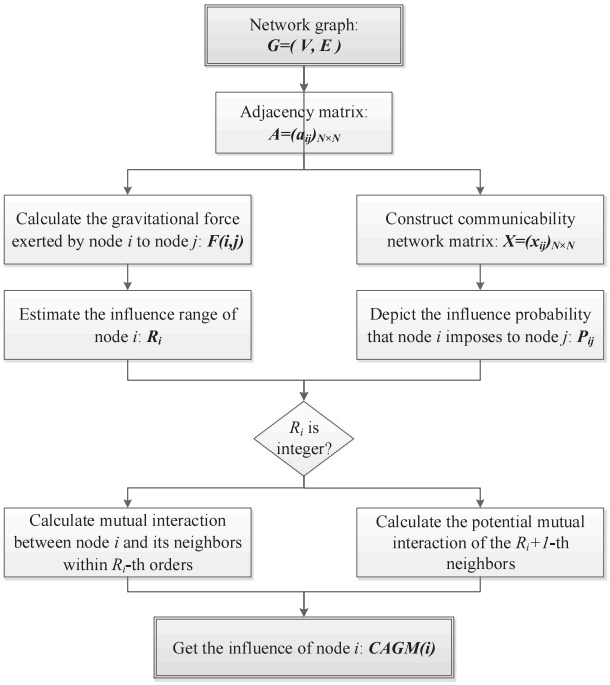


Fig. 1. The flowchart of the proposed CAGM algorithm.

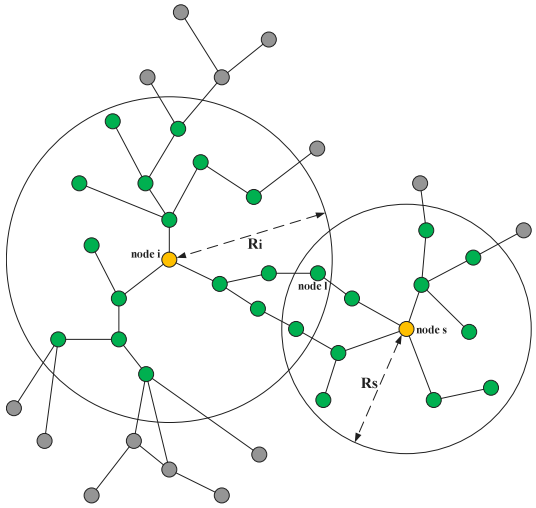


Fig. 2. An example graph for the influence radius of nodes, where $R_i = 3$ and $R_s = 2$.

3.1. Estimating the influence radius of nodes

Inspired by the idea of Newton's gravity formula, Ma and his coworkers (Ma et al., 2016) constructed a gravity centrality model to identify influential nodes, which is exactly the sum of interactions between a node and other nodes within its influence range in the network. The mathematical formula of gravity centrality model is given by

$$GC(i) = \sum_{j\in\psi_i} \frac{ks(i)\cdot ks(j)}{d_{ij}^2}, \tag{1}$$

where ψ_i denotes the neighborhood set whose distance to node i is less than or equal to influence radius r. $ks(i)$ and $ks(j)$ are the k-shell values of node i and node j, and d_{ij} is the distance between nodes i and j.

In the gravity model-based algorithms, the setting of the influence radius attracts much attention of scholars. In Refs. Namtirtha et al. (2018), Wang et al. (2018), Yang and Xiao (2021), researchers only consider the interaction influence of neighbors within three hops and thus the influence radius of nodes is fixed as 3. Considering the difference of networks, influence radius of all nodes is set to half of the network diameter in several existing gravity centrality algorithms (Li et al., 2019; Li, Shang et al., 2021; Liu et al., 2020). In fact, the influence range of each node is related to itself and network structure. In social networks, a few celebrities have a strong voice and can influence multiple neighbor nodes with large influence range, while great majority common users only influence limited neighbor nodes. Given a network, it is unreasonable that the influence range of all nodes are assumed to be a fixed value. Very recently, Li and Xiao (2021) designed an effective measure to recognize important spreaders, where the core idea of this algorithm is to calculate the precise influence range for every node by investigating the association between a node and its furthest node. Inspired by their enlightening idea, we propose a novel adaptive calculation framework for estimating the influence radius of each node in a network with low computational cost.

Specifically, as shown in Fig. 2, node i is randomly selected in the network and its influence range is denoted as R_i, which means that

the influence of node i can affect up to its R_i-th order neighbors. We assume that there exists a node s in the network and it is located outside the influence range of node i. Meanwhile, nodes i and s have an overlapping influence range. In the overlapping scope, the nodes are not only the R_i-th order neighbors of node i, but also the R_s-th order neighbors of node s. Then, we randomly select a node in the overlapping influence region and denote it as l. It is obvious that node l is located on the shortest path between nodes i and s. In other words, if node t is located between nodes i and l, i.e., the distance between nodes i and t is less than the distance between nodes i and l, then node t falls into the influence radius of node i and it cannot affected by node s. Similarly, when node t is located between nodes s and l, node t falls into the influence radius of node s and it cannot be affected by node i. In this sense, node l can be considered as the demarcation point for node pair i and s. By utilizing the information of demarcation points, the value of influence radius for each node can be estimated as follow.

In addition, the location of nodes is also a vital feature in the field of network science, which is rarely considered in the existing gravity model measures. In reality, the location is an important feature in influential nodes identification, and nodes located in the central part of the network are likely to be more attractive than those located in the periphery (Yang & Xiao, 2021). On this basis, we introduce the iteration k-shell algorithm (Iks) (Wang et al., 2016) for representing the relative position of each node.

Inspired by the classical gravity based model (Ma et al., 2016), the gravitational force exerted by node i to node j is defined as

$$F(i,j) = g_{ij} \cdot \frac{k_i \cdot k_j}{d_{ij}^2}, \tag{2}$$

where k_i and k_j represent the degree of nodes i and j, respectively. d_{ij} denotes the shortest distance between node i and node j, which is obtained by Dijkstra's algorithm (Dijkstra, 1959). The gravitational coefficient g_{ij} is given by

$$g_{ij} = e^{(Iks(i)-Iks(j))/(Iks_{max}-Iks_{min})}, \tag{3}$$

where Iks_{max} and Iks_{min} are the maximum and minimum iteration k-shell values of nodes in the network, respectively. In Eq. (2), the gravitational coefficient g_{ij} is introduced to portray the mutual attraction level between nodes i and j.

Each node has its distinct influence range in a network. If the node l is located at the outermost layer neighborhood of node i, then the distance between node i and node l is justly the influence range of node i. In this case, node l is called the demarcation point of node i. According to Eq. (2), we have

$$F(i,l) = g_{il} \cdot \frac{k_i \cdot k_l}{d_{il}^2} \equiv g_{il} \cdot \frac{k_i \cdot k_l}{R_i^2}. \tag{4}$$

In a similar manner, the gravitational force exerted by node s to demarcation node l can be computed as

$$F(s,l) = g_{sl} \cdot \frac{k_s \cdot k_l}{d_{sl}^2}. \tag{5}$$

As observed in Fig. 2, node l is located on the periphery of the influence range of node i, and it is also located on the periphery of the influence range of node s. Since node l falls into the overlapping influence range of node i and node s, it is subject to the attraction of both nodes i and s. Following the principle of force balance, the gravitational force $F(i,l)$ exerted by node i to node l is equal to the gravitational force $F(s,l)$ exerted by node s to node l. Under this assumption, together with Eqs. (4) and (5), we require that

$$F(i,l) = F(s,l) \equiv g_{il} \cdot \frac{k_i \cdot k_l}{R_i^2} = g_{sl} \cdot \frac{k_s \cdot k_l}{(d_{is}-R_i)^2}. \tag{6}$$

If there are more central nodes associated with node l, Eq. (6) can be redefined as

$$g_{il} \cdot \frac{k_i \cdot k_l}{R_i^2} = \overline{g_{sl}} \cdot \frac{\overline{k_s} \cdot k_l}{(\overline{d_{is}}-R_i)^2}, \tag{7}$$

where $\overline{g_{sl}}$ and $\overline{d_{is}}$ represent the average gravitational coefficient and average shortest distance between node l and multiple central nodes, $\overline{k_s}$ denotes the average degree of the multiple central nodes.

Transposing terms and extracting the square root in the right hand of Eq. (6), we get

$$d_{is} - R_i = R_i \sqrt{\frac{k_s \cdot g_{sl}}{k_i \cdot g_{il}}}, \tag{8}$$

solving the above equation with respect to R_i, the explicit formula of influence radius for node i can be derived as

$$R_i = \frac{d_{is}}{1+\sqrt{\frac{k_s}{k_i} \cdot \frac{g_{sl}}{g_{il}}}}. \tag{9}$$

We know from Eq. (9) that the value of R_i is dependent on two unknown nodes s and l. Fortunately, we may introduce a parameter in Eq. (9) to estimate a small value range for R_i. The detailed calculation is proceeded as follows.

First, we take

$$\delta = \sqrt{\frac{g_{sl}}{g_{il}}}, \tag{10}$$

together with Eqs. (3) and (10), the parameter δ can be rewritten as

$$\delta = \sqrt{e^{\frac{(Iks(s)-Iks(l))-(Iks(i)-Iks(l))}{Iks_{max}-Iks_{min}}}} = \sqrt{e^{\frac{Iks(s)-Iks(i)}{Iks_{max}-Iks_{min}}}}, \tag{11}$$

which indicates that the value range of the parameter δ is from $\sqrt{1/e}$ to $\sqrt{e}$. Since $\sqrt{1/e} \approx 0.6065$, $\sqrt{e} \approx 1.6487$, the value of the parameter δ approximately lies in the range $[0.6, 1.7]$.

Next, in order to save computational cost, we consider twelve different cases for the parameter δ in its range with each step of 0.1. For each specified value of parameter δ in $[0.6, 0.7, 0.8, \ldots, 1.5, 1.6, 1.7]$, from Eq. (11) we may obtain the value of $Iks(s)$, and then the values of k_s and d_{is} can be derived. If there exist multiple nodes s corresponding to the value of $Iks(s)$, the values of k_s and d_{is} in (9) are the average degree and average distance of these nodes. Having the values of k_i, k_s, d_{is} and δ, the influence radius R_i of node i can be derived using Eq. (9).

Third, taking a particular value for δ, the influence radius for all nodes in the network can be calculated. By employing Eq. (16) or Eq. (17), one can compute the importance score of each node and thus finally obtains twelve ranking lists corresponding to different δ values. Kendall's τ correlation coefficient is used to measure the accuracy between each ranking list and real spreading capability under SIR model (Maji et al., 2020; Namtirtha et al., 2022). The δ value is determined by the ranking list with the largest Kendall's τ correlation coefficient. The parameter experiment on different datasets will be conducted in Section 5.1.

3.2. Calculating the influence probability between pairs of nodes

It is common for individuals to establish contact through mediums, such as common friends, hobbies, interests and business transactions. Similarly, there are still some associations between unconnected nodes in a complex network. In other words, with the assistance of neighbor nodes, information can be propagated between each pair of unconnected nodes. As is well known, a number of topological and dynamical properties of complex networks are established on the assumption that most of the transport on the network flows along the shortest paths (Estrada & Hatano, 2008). As a matter of fact, both shortest paths and non-shortest paths play important roles during information propagation process.

Considering all possible paths between node pairs, Estrada et al. (2012) designed communicability network matrix with the following

form

$$X = e^{A} = \sum_{y=0}^{\infty} \frac{1}{y!} A^{y} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1N} \\ X_{21} & X_{22} & \cdots & X_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ X_{N1} & X_{N2} & \cdots & X_{NN} \end{bmatrix}, \tag{12}$$

where X_{ij} characterizes the global communicability between node i and node j. In order to construct the communicability network matrix with low computational cost, we adopt the eigenvalue decomposition of the adjacency matrix $A = Q \wedge Q^{-1}$, and $\wedge = diag(\lambda_1(A), \ldots, \lambda_i(A), \ldots, \lambda_N(A))$, where $\lambda_i(A)$ is the ith eigenvalue of A, and Q is an orthogonal matrix consisting of a corresponding standard orthogonal basis. On this basis, we can obtain that

$$A^{y} = Q \wedge Q^{-1} \cdot Q \wedge Q^{-1} \cdots Q \wedge Q^{-1} = Q \wedge Q^{-1}. \tag{13}$$

Note that the symbol ∞ in (12) means the largest distance between node pairs in the network, Eq. (12) can be simplified as

$$X = \sum_{y=0}^{L} \frac{1}{y!} Q \wedge Q^{-1}, \tag{14}$$

where L denotes the diameter of network.

As defined by Eq. (12), communicability network matrix can effectively depict the probability that pairs of nodes establish connection with each other under all the possible paths. The larger the X_{ij} value is, the more likely that one node shares resource with another. To facilitate the subsequent calculation, we normalize the matrix X and the normalized element is denoted as

$$P_{ij} = \frac{X_{ij} - X_{ij}^{min}}{X_{ij}^{max} - X_{ij}^{min}}, \tag{15}$$

where X_{ij}^{max} is the maximum value in matrix X, X_{ij}^{min} is the maximum value in matrix X. The value of P_{ij} means the possibility that node i spreads information to node j in the network. Meanwhile, P_{ij} can also be considered as the influence probability between node i and node j in a certain sense.

3.3. Computing the importance of nodes

As discussed in Sections 3.1 and 3.2, we propose some solutions to address the bottleneck problems mentioned above in the existing gravity model-based algorithms. Considering the location information of nodes in the network, we define the gravitational force exerted by a source node to target node by introducing the gravitational coefficient between node pair. Subsequently, we provide a computing procedure for estimating the influence range for each node, which can well depict the inherent heterogeneity of nodes in complex networks. Furthermore, using the communicability network matrix, the influence probability between each pair of nodes is calculated through considering all possible paths rather than the shortest path only. Based on the above analysis, we propose the Communicability-based Adaptive Gravity Model (CAGM) to identify influential nodes, which intends to make full use of influence probability and influence intensity information of neighbor nodes located in the influence scope.

If the influence radius R_i is an integer, the proposed algorithm only takes into account the mutual interaction between node i and its neighbors within R_i-th orders. The influence of node i is calculated as

$$\begin{aligned} CAGM(i) = &\sum_{j \in \Gamma_1(i)} P_{ij} \cdot F(i,j) + \sum_{k \in \Gamma_2(i)} P_{ik} \cdot F(i,k) + \cdots \\ &+ \sum_{w \in \Gamma_{R_i}(i)} P_{iw} \cdot F(i,w) = \sum_{d_{iy} \le R_i} P_{iy} \cdot F(i,y), \end{aligned} \tag{16}$$

where d_{iy} is the shortest distance between nodes i and y.

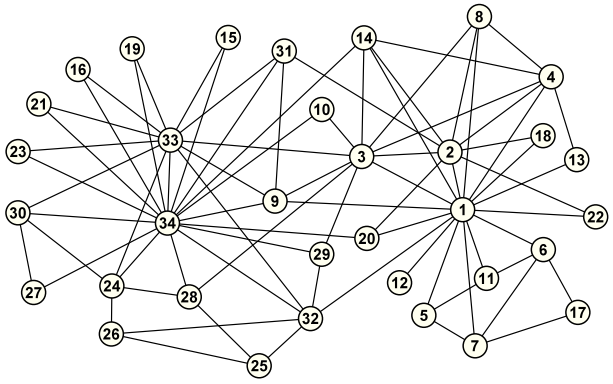


Fig. 3. The Karate network with 34 nodes and 78 edges.

On the contrary, if the influence radius R_i is a non-integer, the potential mutual interaction of the $\lceil R_i \rceil$-th neighbors also needs to be considered. In this case, the influence of node i is calculated as

$$\begin{aligned} CAGM(i) = &\sum_{j \in \Gamma_1(i)} P_{ij} \cdot F(i,j) + \sum_{k \in \Gamma_2(i)} P_{ik} \cdot F(i,k) + \cdots + \\ &\sum_{x \in \Gamma_{\lfloor R_i \rfloor}(i)} P_{ix} \cdot F(i,x) + (R(i) - \lfloor R_i \rfloor) \cdot \sum_{d_{iw} = \lceil R_i \rceil} P_{iw} \cdot F(i,w) \\ = &\sum_{d_{iy} \le \lfloor R_i \rfloor} P_{iy} \cdot F(i,y) + (R(i) - \lfloor R_i \rfloor) \cdot \sum_{d_{iw} = \lceil R_i \rceil} P_{iw} \cdot F(i,w). \end{aligned} \tag{17}$$

To illustrate the proposed CAGM algorithm more clearly, the Karate network shown in Fig. 3, is chosen as an example. Firstly, by varying the parameter δ from 0.6 to 1.7 with each step of 0.1, the influence radius of each node can be obtained by using Eq. (9). Then, according to Eq. (15), one can easily calculate the influence probability between each pair of nodes. Next, through Eqs. (16)–(17), we can compute the CAGM value for each node in the Karate network. Twelve ranking lists of nodes in the Karate network can be obtained, which correspond to specific values for the parameter δ. The optimal parameter δ is determined according to the Kendall tau correlation coefficient. For the Karate network, the proposed CAGM algorithm reaches the highest Kendall tau correlation coefficient when $\delta = 0.7$. Finally, with the optimal parameter value in hand, one can easily get the influence radius and CAGM value for each node as shown in Table 2, here D_i denotes the maximum distance form node i to the others.

The outline of communicability-based adaptive gravity model is presented in **Algorithm 1**. By comprehensively considering the gravitational coefficient and mutual attraction between a node and its farthest node, and taking a particular value for δ, lines 1–8 compute the exact influence range of each node to accurately portray each node's influence scope according to the assumption mentioned above. Line 9 outputs the final influence radius list corresponding to the largest Kendall's τ correlation coefficient. Lines 10–14 construct the communicability network matrix of the network and elements in this matrix represent the influence probability between each pair of nodes. To facilitate the subsequent calculation, line 15 normalizes the communicability network matrix. Lines 16–22 evaluate the importance of each node by considering two different cases. If the exact influence radius of node i is an integer, all nodes locating in its influence sphere need to be taken into account to evaluate the importance of nodes. If the exact influence radius of node i is a non-integer, the potential influence of the $\lceil R_i \rceil$-th neighbors also needs to be taken into account. Finally, line 23 returns the node influence ranking list F.

3.4. Computational complexity

The proposed CAGM algorithm consists of three phases. Its computational complexity is analyzed as follows. **In the first phase**, calculating the distance between node pairs needs $O(N \cdot \log N)$ time, the

Table 2
The maximum distance D_i, influence radius R_i, and the $CAGM(i)$ value for all nodes in Karate network, where the parameter δ in Eq. (11) is equal to 0.7.

Node	D_i	R_i	$CAGM(i)$	Node	D_i	R_i	$CAGM(i)$	Node	D_i	R_i	$CAGM(i)$
1	3	1.0696	701.0502	13	4	0.5455	5.4364	25	4	2.0229	5.6502
2	3	1.4371	220.4516	14	3	1.0889	131.8263	26	4	2.0229	6.0051
3	3	0.8312	16.1856	15	5	0.3859	4.6751	27	5	0.2199	0.5377
4	3	1.3229	114.1344	16	5	0.3859	4.6751	28	4	1.5422	29.7997
5	4	1.3486	7.6638	17	5	0.9545	0.4965	29	4	2.0229	26.5136
6	4	1.4101	11.1727	18	4	0.5455	5.4364	30	5	2.8202	40.2290
7	4	1.4101	11.1727	19	5	0.3859	4.6751	31	4	1.5119	59.7766
8	4	1.3229	49.6841	20	3	1.3486	24.3452	32	3	1.4099	94.0165
9	3	1.0889	142.8658	21	5	0.3859	4.6751	33	4	1.6193	386.8807
10	4	0.1364	4.0716	22	4	0.5455	5.4364	34	4	1.5404	592.4136
11	4	1.3486	7.6638	23	5	0.3859	4.6751				
12	4	0.2199	0.3294	24	5	1.8430	57.4370				

Algorithm 1 Communicability-based adaptive gravity model

Input: Network $G = (V, E)$, $|V(G)| = N$.
Output: Node influence ranking list $CAGM = \{CAGM(1), ..., CAGM(N)\}$.

for $i = 1$ *to* N **do**
  for $\delta = 0.6$ *to* 1.7(with each step of 0.1) **do**
    Assume virtual node s is far away enough form node i
    Calculate the $Iks(s)$ value according to Eq. (11)
    Derive the values of k_s and d_{is}
    Calculate the influence radius R_i according to Eq. (9)
  end for
end for
Determine the influence radius list based on Kendall's τ correlation coefficient
for $i = 1$ *to* N **do**
  for $j = 1$ *to* N **do**
    Calculate the influence probability X_{ij} based on Eq. (12)
  end for
end for
Normalize the communicability network matrix according to Eq. (15)
for $i = 1$ *to* N **do**
  if R_i is an integer **then**
    Calculate the CAGM value of node i using Eq. (16)
  else
    Calculate the CAGM value of node i using Eq. (17)
  end if
end for
return Node influence ranking list $CAGM = \{CAGM(1), \cdots, CAGM(N)\}$

time cost of calculating degree centrality and iteration k-shell algorithm for every node are $O(N)$. Thus, calculating the influence radius for all nodes requires $O(N \cdot \log N + 2 \cdot N)$ time. **In the second phase**, both the construction and normalization of communicability network matrix need $O(N)$ time, thus it takes $O(2 \cdot N)$ time to calculate the influence probability between pairs of nodes. **In the third phase**, the time complexity of computing the importance of nodes is related to topological structure of networks. If all nodes in the network can only influence nearest neighbors, calculating the $CAGM(i)$ values for all nodes will cost $O(N)$ time. The worst case is that the influence radii for all nodes are equal to the diameter of the network. In this case, calculating the $CAGM(i)$ values will cost $O(N^2)$ time. In addition, for determining the optimal parameter, the process of computing the importance of nodes will be repeated 12 times. To sum up, the computational complexity of the CAGM algorithm in the best case is $O(N \cdot \log N + 2 \cdot N + 2 \cdot N + 12 \cdot N) \approx O(N \cdot \log N)$, and the computational complexity of the CAGM algorithm in the worst case is $O(N \cdot \log N + 2 \cdot N + 2 \cdot N + 12 \cdot N^2) \approx O(N^2)$. For most networks, the influence radius for nodes is much less than the diameter of network, so the proposed algorithm has low computational complexity and it is applicable to large-scale networks.

Table 3
The basic statistical features of twelve networks with N nodes and M edges. $\langle k \rangle$ represents the average degree, β_{th} denotes the threshold of the network and λ is a tunable parameter in the LGC (Ullah et al., 2021) method.

Network	N	M	$\langle k \rangle$	β_{th}	λ
Facebook	4039	88234	43.6910	0.0094	0.2
Power	4941	6594	2.6691	0.3483	0.6
BA-6000	6000	29982	9.9940	0.0379	0.9
LFR-6000	6000	17962	5.9873	0.1178	0.9
WS-6000	6000	29804	9.9547	0.1004	0.8
Erdos	6100	7515	2.4639	0.0613	0.5
WV	7066	100736	28.5129	0.0069	0.5
Ca-hepth	9877	25998	5.2593	0.0798	0.4
PGP	10680	24316	4.5536	0.0530	1.0
DBLP	12591	49620	7.8818	0.0229	0.6
Sex	15810	38540	4.8754	0.0365	0.5
Condmat	23133	93439	8.0784	0.0453	1.0

4. Experimental setup

4.1. Datasets

To portray the performance of the CAGM algorithm, nine real-world networks are choosed varying from small to large, and with different characteristics and functions. (i) Facebook: a network of Facebook social cycles; (ii) Power: a network of power grid system in western United States; (iii) Erdos: a co-authorship network around Paul Erdos; (iv) WV: a social network of Wikipedia; (v) Ca-hepth: a collaboration network in Arixv; (vi) PGP: a social network between users of pretty good privacy algorithm; (vii) DBLP: a collaboration network in the field of scientific publications; (viii) Sex: a social network of prostitution; (ix) Condmat: a collaboration network of the condensed matter section (Cond-mat) in Arixv.

In addition, three artificial networks are chosen as datasets in this work, including Scale-Free Network (BA-6000) (Barabasi & Albert, 1999), Lancichinetti–Fortunato–Radicchi network (LFR-6000) (Lancichinetti et al., 2008) and Small-World network (WS-6000) (Watts & Strogatz, 1998). The BA-6000 artificial network with parameters $N = 6000$, initial node's number $m_0 = 20$ and number of nodes per join $m = 4$. The LFR-6000 artificial network with parameters $N = 6000$ and probability of forming edges between each pair of nodes $p = 0.2$. The WS-6000 artificial network with parameters $N = 6000$, average degree of initial network $\langle k_0 \rangle = 2$ and rewiring probability $p = 0.001$. Both sets of artificial network datasets are generated by software Gephi. Table 3 lists the main features of these twelve networks.

4.2. Algorithms for comparison

To illustrate the performance of the CAGM algorithm, eight popular and competitive algorithms are chosen as baseline algorithms.

- ECRM (Zareie et al., 2020): A novel cluster rank algorithm to detect the influential nodes, which is based on local clustering coefficient and the correlation degree between the node itself and neighbor nodes.
- EGM (Li & Xiao, 2021): An improved gravity model to detect influential nodes, which utilizes degree distribution information and shortest distance to define exact influence range for each node in the network.
- GC (Ma et al., 2016): GC is the first gravity model to evaluate the influence of nodes by using the classical gravity formula, where the k-shell value is regarded as the mass of a node and its influence is defined by summing attraction forces within three level neighbors.
- GGC (Li, Shang et al., 2021): A generalized gravity model to find influential spreaders, which utilizes both local clustering coefficient and degree information of nodes to refine the mass of each node.
- KSGC (Yang & Xiao, 2021): An extended gravity model assesses the importance of nodes by comprehensively considering both the location, local and global information of nodes.
- Ksh (Namtirtha et al., 2018): A k-shell hybrid method to evaluate the propagation ability of nodes in a complex network, which integrates multiple dimensional information including degree, k-shell, contact distance as well as neighborhood set.
- LGC (Ullah et al., 2021): An effective algorithm to measure the influence of nodes on the basis of local and global characteristics of a network simultaneously.
- LKG (Samir et al., 2021): A scalable and fast approach to detect top-k influential nodes, which divides the entire network into communities by using the classic Louvain algorithm, and employs k-shell decomposition algorithm to identify the seed nodes in the communities.

4.3. The spreading model

The susceptible–infected–recovered (SIR) model, one of the most classical propagation dynamic models in complex networks, is widely used to simulate the spreading process (such as the spread of diseases, ideas, cultures, or rumors) (Dong et al., 2022; Xu & Meng, 2023). Owing to its simplicity and proper simulation capability in contagious spreading processes, SIR model has been frequently employed to evaluate the performance of algorithms in influential nodes identification. SIR model has three states: susceptible (S), infected (I) and removed (R). Initially, the source spreading node v_i is initially set to I, while all other nodes are set to S. Then, each infected node can infect its susceptible neighbors independently with probability β, and itself transforms into the recovered state with probability μ, where we set $\mu = 2\beta$. The propagation process is repeated until there are no nodes in state I. The threshold of the network is $\beta_{th} = \langle k\rangle/\langle k^2\rangle$. Table 3 lists the β_{th} values for the twelve datasets. To improve the accuracy, every node simulates the spreading process 1000 times and the real spreading capability of nodes is the average value.

4.4. Evaluation methods

The Kendall's tau correlation coefficient (Maji et al., 2020; Namtirtha et al., 2022) is introduced to evaluate the association of two sequences X and Y. Suppose that list $X = \{x_1, x_2, \ldots, x_N\}$ is the node spreading capability list obtained by SIR model, and list $Y = \{y_1, y_2, \ldots, y_N\}$ represents the node centrality score list obtained by different algorithms. If $(x_i - x_j)(y_i - y_j) > 0$, the node pairs (x_i, y_i) and (x_j, y_j) are said to be consistent. Otherwise, they are inconsistent. The Kendall's correlation coefficient τ is calculated as,

$$\tau = \frac{2(X_a - X_b)}{|X|(|X|-1)}, \tag{18}$$

where X_a and X_b are the number of consistent and inconsistent node pairs in the two lists X and Y, and $|X|$ denotes the number of nodes in the network.

To evaluate the performance of important node identification methods on measuring super-spreaders, the imprecision function is introduced to evaluate the difference between the influence ranges of nodes calculated by a specified method to those simulated by SIR model (Kitsak et al., 2010), which is calculated as,

$$T(p) = 1 - \frac{M(p)}{M_{eff}(p)}, \tag{19}$$

where $M(p)$ and $M_{eff}(p)$ denote the average influence ranges of the top-p ($p = q \times N$, q is node proportion and $q \in [0,1]$) nodes in the two ranked lists given by the specified algorithm and SIR model, respectively. Obviously, the closer the $T(p)$ value is to 0, the better method performs in identifying top-ranked important nodes.

Monotonicity (Maji et al., 2021) is employed to compare the distinguishing ability of algorithms in measuring the importance of nodes, which can be calculated as,

$$Mon(Z) = \left(1 - \frac{\sum_{z\in Z} |B|_z \times (|B|_z - 1)}{|B| \times (|B| - 1)}\right)^2, \tag{20}$$

where Z denotes the ranking score list calculated by different algorithms mentioned above. $|B|$ is the volume of the ranking sequence Z, and $|B|_z$ represents the number of nodes ranked the zth. The monotonicity score lies in the continuous range $[0,1]$. When Mon is 1, it means that there is only one node per rank in the ranking sequence Z derived by a specific method, indicating that the method has the best distinguishing ability. In contrast, $Mon = 0$ (i.e., all nodes in the sequence Z have the same rank) means that the ranking method has the poorest distinguishing ability.

5. Experimental results

In this section, the proposed CAGM algorithm is compared with eight popular algorithms for solving important nodes identification problem, including ECRM (Zareie et al., 2020), EGM (Li & Xiao, 2021), GC (Ma et al., 2016), GGC (Li, Shang et al., 2021), KSGC (Yang & Xiao, 2021), Ksh (Namtirtha et al., 2018), LGC (Ullah et al., 2021) and LKG (Samir et al., 2021). In the following, we conduct five groups of experiments on nine real-world datasets and three artificial datasets. The first group of experiments is to determine the proper value of δ for each dataset. The other four groups of experiments illustrate the performance of CAGM in terms of ranking accuracy, size of influence spread, imprecision function and discrimination ability.

5.1. Experiments for the parameters

In contrast to existing gravity model-based algorithms, the advantage of the proposed CAGM algorithm is that the influence radius of each node can be uniquely determined and thus the higher ranking accuracy can be achieved. According to Eq. (9), the value of influence radius R_i is dependent on the parameter δ. It follows from Eq. (11) that the value of δ lies in the range $[0.7, \ldots, 1.7]$. The first group of experiments is to determine the proper value of δ for each dataset. For this purpose, we investigate the effect of the parameter δ on the accuracy of the proposed CAGM algorithm. To reduce the involved computational cost, the parameter δ is variable from 0.7 to 1.7 with each step of 0.1. In this case, twelve ranking lists are obtained which correspond to a specific value of δ. Finally, we can calculate the Kendall's tau correlation coefficients between these twelve ranking lists and the real spreading capability list generated by the SIR model. Finally, the δ value is determined by the ranking list with the largest Kendall's τ correlation coefficient. Generally speaking, δ is variable with the change of network structures. Table 4 represents the experimental results over twelve datasets with twelve different values for

Table 4
Effect of variation of parameter δ on the accuracy of the proposed CAGM algorithm based on Kendall tau correlation coefficient over the twelve datasets.

Network	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7
Facebook	**0.6473**	0.6365	0.5116	0.4766	0.5234	0.4551	0.4059	0.4229	0.4297	0.4501	0.4702	0.4814
Power	0.4779	0.4962	0.5820	0.5869	**0.6419**	0.6342	0.6292	0.6166	0.6042	0.4783	0.4768	0.4604
BA-6000	0.3905	0.3923	0.3955	0.4149	0.5490	**0.5544**	0.5287	0.4631	0.2308	0.2814	0.4317	0.4269
LFR-6000	0.6018	0.6106	0.6577	**0.7908**	0.7883	0.7817	0.7086	0.5841	0.4113	0.4069	0.3755	0.2325
WS-6000	0.2444	**0.2527**	0.2515	0.2512	0.1852	0.1281	0.0300	0.0112	0.0158	0.0386	0.0736	0.2167
Erdos	0.7252	0.7253	0.7258	0.7312	**0.7947**	0.7522	0.7269	0.7258	0.7569	0.6702	0.6735	0.7176
WV	**0.7250**	0.7247	0.7245	0.7153	0.6576	0.5066	0.4066	0.3787	0.2912	0.2326	0.1940	0.1475
Ca-hepth	0.5787	0.5126	0.3689	0.2117	**0.6109**	0.3655	0.3747	0.3858	0.3837	0.3914	0.3714	0.3765
PGP	**0.6013**	0.5831	0.5497	0.4002	0.4674	0.4647	0.4742	0.4670	0.4484	0.4517	0.4899	0.4804
DBLP	0.6925	0.6922	0.6927	**0.6965**	0.6611	0.6329	0.5668	0.5480	0.4923	0.4243	0.4218	0.3541
Sex	0.5495	0.5583	0.5638	**0.5770**	0.5366	0.5172	0.5099	0.4600	0.4183	0.4465	0.4429	0.4359
Condmat	0.7655	0.7751	0.7931	**0.8180**	0.7970	0.7760	0.7688	0.7471	0.7197	0.6863	0.6416	0.5972

parameter δ. As can be observed in Table 4, the Kendall's τ coefficient is variable when the parameter δ changes. More specifically, the highest τ value can be obtained for $\delta = 0.6$ over Facebook, WV and PGP datasets, for $\delta = 0.7$ over WS-6000 dataset, for $\delta = 0.9$ over LFR-6000, DBLP, Sex and Condmat datasets, for $\delta = 1.0$ over Power, Erdos and Ca-hepth dataset, and for $\delta = 1.1$ over BA-6000 dataset.

5.2. Ranking accuracy

In the second group of experiments, we investigate the effect of the infection probability β of the SIR model on the ranking accuracy of different algorithms. As an important parameter in SIR model, the value of β will affect the accuracy of the entire experiment. If the β value is too large, we cannot precisely distinguish the influence of nodes due to the spreading of infected nodes will quickly extend to the whole network. In contrast, if the β value is too small, the spreading of nodes will be limited to a small scope and the influence of nodes cannot be accurately assessed. For the networks with $\beta_{th} \approx 0.10$, we choose 20 different probabilities in the range $[0, 0.2]$ with each step of 0.01. For the networks with high propagation threshold, such as the Power dataset with $\beta_{th} = 0.3483$, we choose 20 different probabilities at 0.01 intervals from the range $[0.15, 0.35]$.

Fig. 4 shows the Kendall's tau correlation coefficients of nine algorithms in nine real-world datasets and three artificial datasets. Experimental results show that the proposed CAGM performs better than the eight compared algorithms in Facebook, Power, WS-6000, Erdos, WV and Condmat datasets. In addition, the ranking accuracy is sensitive to β in some specific datasets. In Ca-hepth, PGP and Sex datasets, when infection rate is small, the ranking accuracy of CAGM algorithm is inferior to contrast algorithms. When the infection rate is around threshold, the ranking accuracy of CAGM increases obviously and outperforms compared algorithms. When the infection rate continues to increase, the ranking accuracy of CAGM is gradually stable and its performance is far superior to that of the contrast algorithms. In Erdos dataset, the CAGM algorithm performs better than other algorithms except when infection probability $0.02 < \beta < 0.11$. In BA-6000 dataset, the τ value of CAGM is superior to other algorithms except when infection probability $0.15 < \beta < 0.20$. In LFR-6000 dataset, the proposed CAGM algorithm is inferior to Ksh algorithm when $0.02 < \beta < 0.10$, and it is obvious that our proposed algorithm and Ksh algorithm are superior to seven other contrast algorithms throughout the whole spreading process. In summary, the ranking list obtained by CAGM is highly correlated with the real spreading capability list obtained by the SIR model. In other words, compared with eight popular and competitive algorithms, the proposed CAGM algorithm can rank and detect influential nodes in complex networks more accurately, and it has a lower error compared to the real spreading capability of nodes.

5.3. Imprecision function

The third group of experiments utilize the imprecision values to further compare the CAGM algorithm with eight contrast algorithms. According to Eq. (19), the imprecision values of different algorithms are employed to assess the propagation capability of top-ranked nodes. In this experiment, the infection probability β is fixed as threshold for all studied datasets, and Fig. 5 shows the experimental simulation results. Here, proportion of top-ranked nodes P varies in the range $[0.01, 0.1]$ with each step of 0.01. In six datasets including Power, BA-6000, WS-6000, Ca-hepth, Sex and Condmat, the $T(p)$ values of CAGM are much lower than those of comparison algorithms, which indicates that CAGM algorithm performs better than other eight algorithms. In addition, the $T(p)$ values of EGM, KSGC and Ksh algorithms are approximate to CAGM in Erdos dataset. In Facebook and PGP datasets, the performance of CAGM, GC and LKG are better than other compared algorithms in terms of $T(p)$ values. In LFR-6000 and WV datasets, the $T(p)$ values of CAGM and Ksh are lower than others. The $T(p)$ values of CAGM is close to those of KSGC, LGC and LKG algorithms in DBLP dataset. To summarize, the proposed CAGM algorithm has a rather low imprecision function value, and it achieves the best performance in all studied networks.

5.4. Top-10 nodes

In various application fields, people are always more interested in super-spreaders (i.e., nodes with the high influence ranking). Inspired by this, we select the top-10 nodes calculated by specific algorithms to simulate the spreading process in SIR model, and the experiments can effectively represent the spreading capability of these nodes. The infection probability $\beta = \beta_{th}$ and recovered probability $\mu = 2\beta$. Fig. 6 shows the influence coverage $F(t)$ of the top-10 ranking nodes on twelve datasets for different methods. As shown in Fig. 6, the $F(t)$ values of CAGM are higher than other algorithms throughout spreading process. In six datasets including Facebook, Power, WS-6000, LFR-6000, Ca-hepth, DBLP and Condmat, the $F(t)$ obtained by CAGM is approximate to the compared algorithms at the beginning of propagation process. However, the $F(t)$ value is always the best when the propagation is stable. In Erdos dataset, the performance of CAGM, KSGC and Ksh are superior to other algorithms in terms of $F(t)$ value. In addition, CAGM is similar to ECRM and LGC algorithms in BA-6000 and Sex datasets, is approximate to ECRM and GGC algorithms in PGP dataset, and is close to GGC algorithm in WV dataset. Notably, the infection scale obtained by CAGM is the largest when the spreading process reaches a steady state. It is concluded that the proposed CAGM is an effective algorithm for detecting the most influential nodes in networks.

5.5. Discrimination ability

In this subsection, monotonicity is employed to further evaluate the distinguishing ability of different algorithms. Table 5 gives the monotonicity values of nine algorithms on twelve networks, and the maximum monotonicity value of these nine algorithms in each network are highlighted in bold. Experimental results show that the proposed CAGM achieves the highest monotonicity values for all twelve datasets. In addition, the monotonicity values of ECRM, GGC, KSGC and LGC are

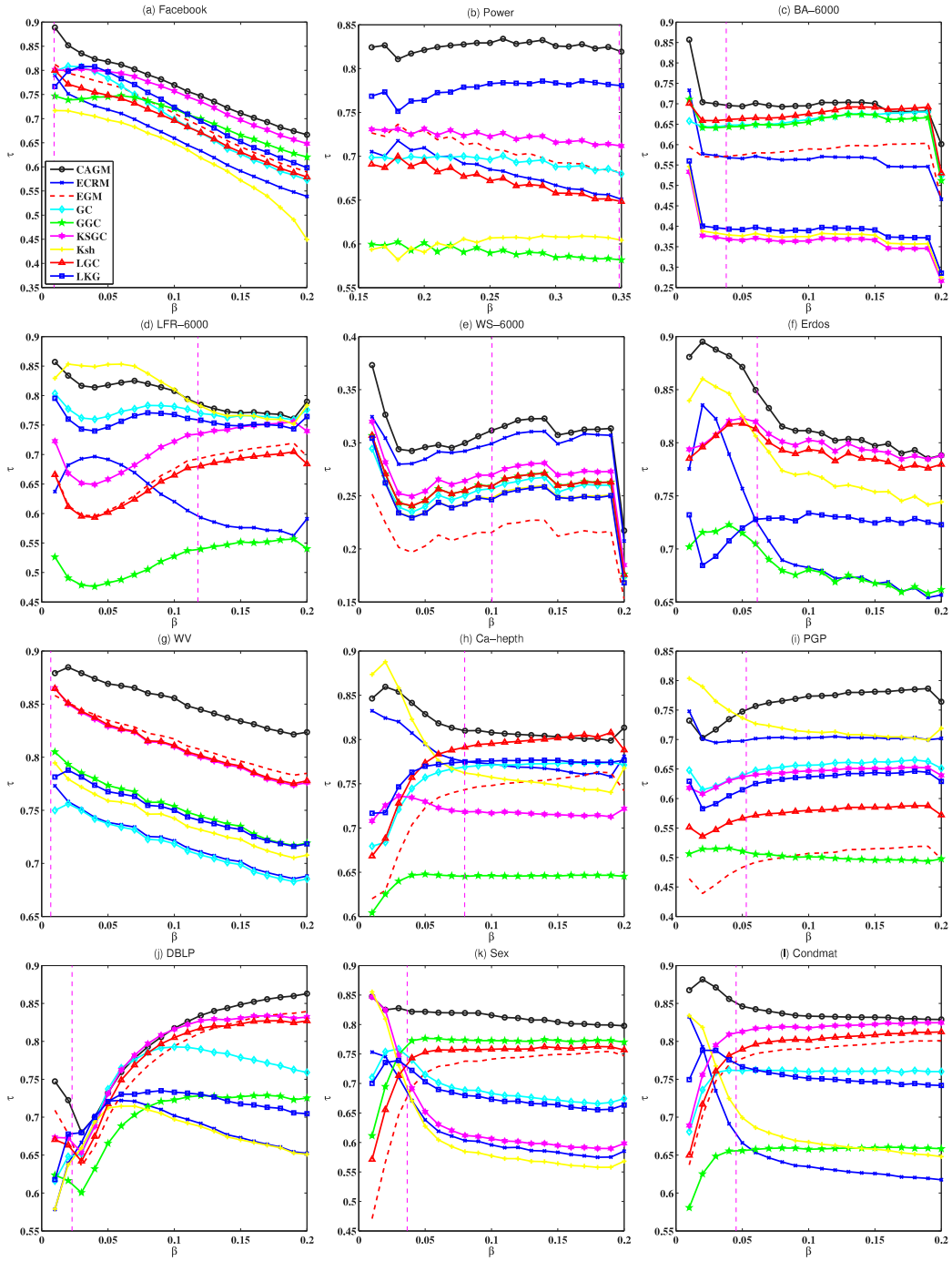


Fig. 4. The Kendall's tau correlation coefficient between the ranking list simulated by the SIR model and the ranking lists generated by nine different algorithms with varying infection probabilities β and recovered probability $\mu = 2\beta$. Note that the blue dotted line indicates the epidemic outbreak threshold β_{th}, and simulation results of the SIR model are obtained through an average of 1000 independent experiments. (a) Facebook, (b) Power, (c) BA-6000, (d) LFR-6000, (e) WS-6000, (f) Erdos, (g) WV, (h) Ca-hepth, (i) PGP, (j) DBLP, (k) Sex, (l) Condmat.

also close to that of CAGM in most of datasets. We note that EGM, GC, Ksh and LKG reach the highest values in a small fraction of networks, but their overall performances are poor. In summary, the proposed CAGM algorithm performs the best in terms of monotonicity and it can assign each node in the network with a unique influence value to some extent.

Furthermore, we introduce the CCDF (complementary cumulative distribution function) to evaluate the variability of different algorithms (Gupta & Mishra, 2021). The mathematical definition of CCDF is as $F(Y) = P(Y > y)$, where $P(Y > y)$ denotes the possibility of elements in Y is larger than y, y is an element in Y. From the view of differentiating nodes, if there are several nodes with the same rank, it is difficult to choose a node as a representative, and the method of ranking a few nodes in the same rank performs better.

Fig. 7 illustrates the CCDF for CAGM algorithm and other eight contrast algorithms in different networks. Experimental results demonstrate that the CAGM algorithm can distinguish the importance of nodes by giving different ranking scores to a large number of nodes. Specifically, these nine algorithms have close performance in terms of distinguishing the importance of the nodes in Facebook and WV datasets. CAGM algorithm can divide the nodes into a maximum number of levels and performs better than other eight compared algorithms

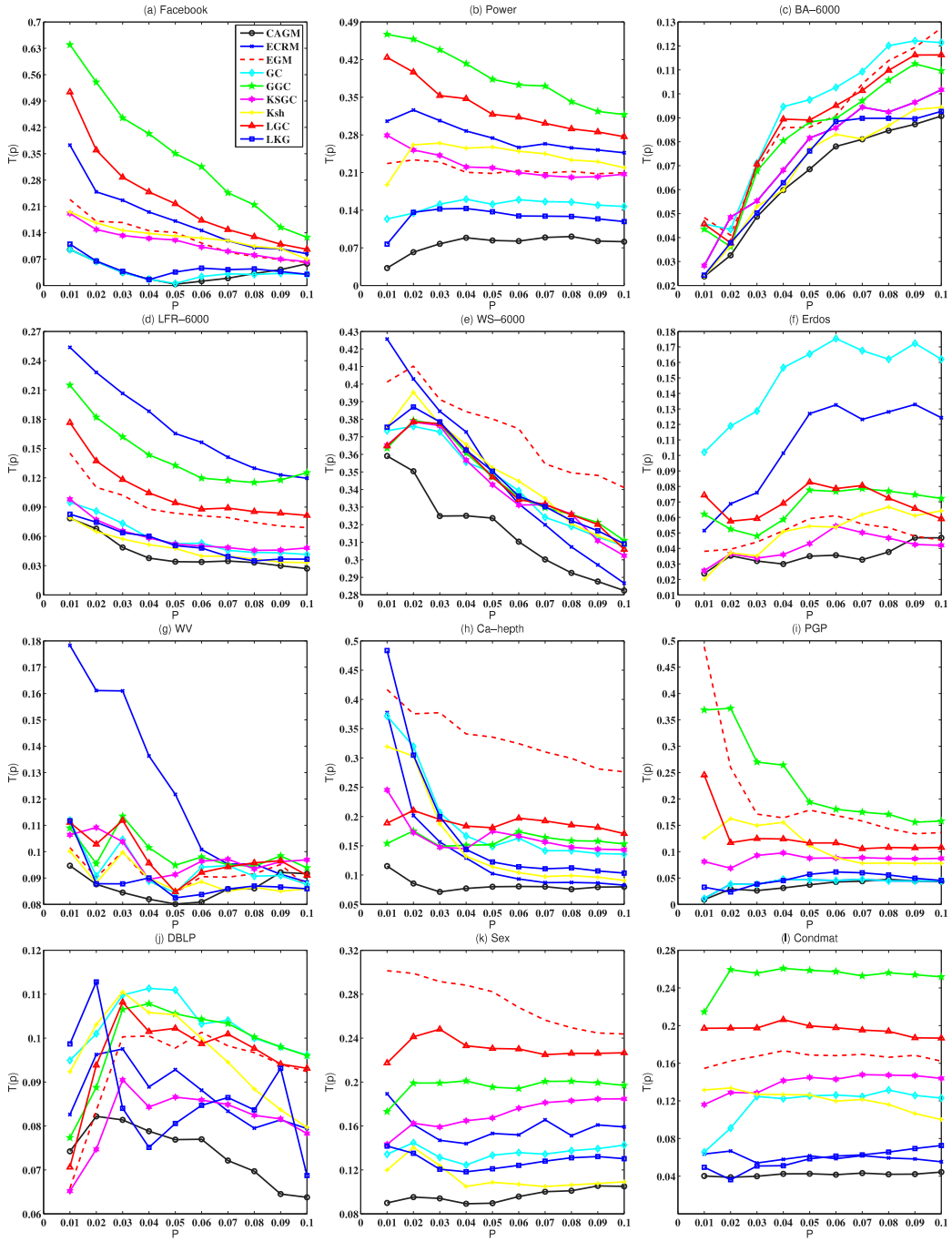


Fig. 5. The imprecision values of nine centrality measures in nine real-world datasets and three artificial datasets. The results are obtained by 1000 independent experiments with the SIR model when $\beta = \beta_{th}$ and $\mu = 2\beta$. P denotes the proportion of nodes in the network whose value ranges from 0.01 to 0.1. (a) Facebook, (b) Power, (c) BA-6000, (d) LFR-6000, (e) WS-6000, (f) Erdos, (g) WV, (h) Ca-hepth, (i) PGP, (j) DBLP, (k) Sex, (l) Condmat.

in Power, Erdos, Ca-hepth, PGP, DBLP, Sex and Condmat datasets. In addition, in three artificial datasets including BA-6000, ER-6000 and WS-6000, GC algorithm is inferior to other eight algorithms. In summary, the CAGM algorithm has the best performance in terms of CCDF on all studied datasets.

6. Conclusions

How to recognize influential nodes in networks is an interesting and popular topic in the field of network science. In reality, for characterizing the influence of nodes more accurately, we may consider both the information about the source node itself and its nearest neighbors, as well as the interactions between them. Inspired by the idea of universal gravity formula, we propose the communicability-based adaptive gravity model (CAGM) to detect the most influential nodes in networks. Due to the inherent heterogeneity of nodes in complex networks, the influence radius of each node is different. To achieve good balance between algorithm accuracy and computational complexity, we estimate the influence radius of each node by comprehensively considering the location difference and mutual attraction between each pair of nodes. Furthermore, the communicability network matrix is introduced to depict the influence probability between each pair of nodes under all

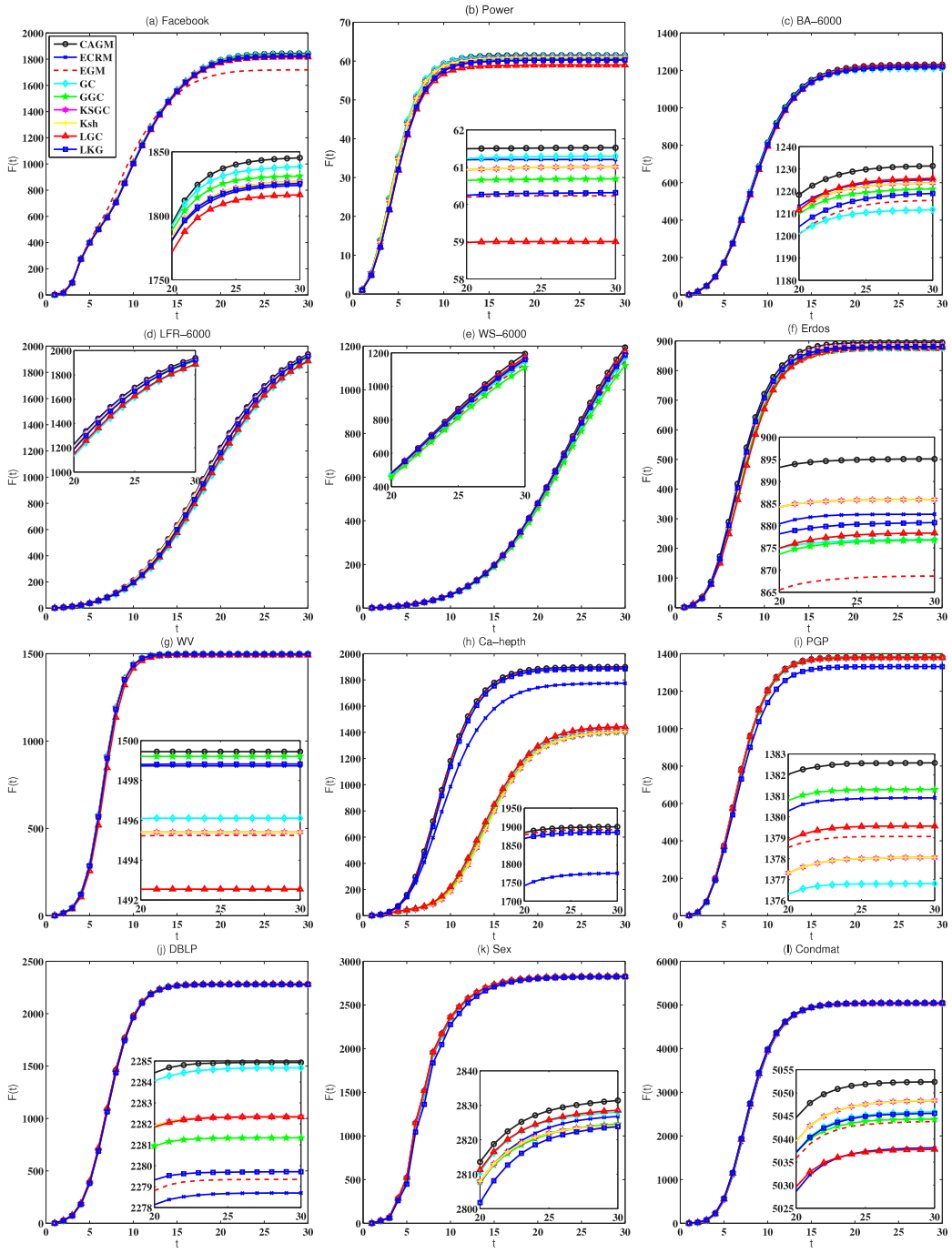


Fig. 6. The influence of propagation for the top-10 ranking nodes using CAGM and other eight algorithms, where F(t) represents the number of infected and recovered nodes at the time (t) in the range between 0 and 30. The F(t) values are obtained by averaging 1000 independent runs of the SIR model. (a) Facebook, (b) Power, (c) BA-6000, (d) LFR-6000, (e) WS-6000, (f) Erdos, (g) WV, (h) Ca-hepth, (i) PGP, (j) DBLP, (k) Sex, (l) Condmat.

possible paths. On this basis, the importance of each node is evaluated by taking into account both the influence probability and influence intensity information of neighbor nodes located in its influence radius. This algorithm has a low computational complexity of $O(N^2)$, which is applicable to large-scale networks.

To verify the effectiveness of CAGM, several groups of experiments are conducted on some real and artificial datasets. The proposed CAGM algorithm is compared with several popular algorithms including ECRM, LGC and LKG, and it is also compared with five existing gravity-based algorithms. Experimental results show that CAGM performs better than eight popular algorithms in terms of top-10 nodes, discrimination ability, imprecision function and ranking accuracy. Specifically, in terms of the ranking accuracy, the proposed CAGM algorithm performs better than other comparison algorithms with a margin as high as 2.90% in each dataset. Furthermore, in terms of the imprecision function, the T(p) values of CAGM are much lower than those of comparison algorithms with a margin as high as 3.71%. In addition, the proposed CAGM algorithm achieves the largest spreading probability when the propagation process is termination. In four datasets including Facebook, WV, DBLP and DBLP, the margin between CAGM and eight comparison algorithms reaches 0.49%. While the proposed CAGM algorithm outperforms other algorithms on the

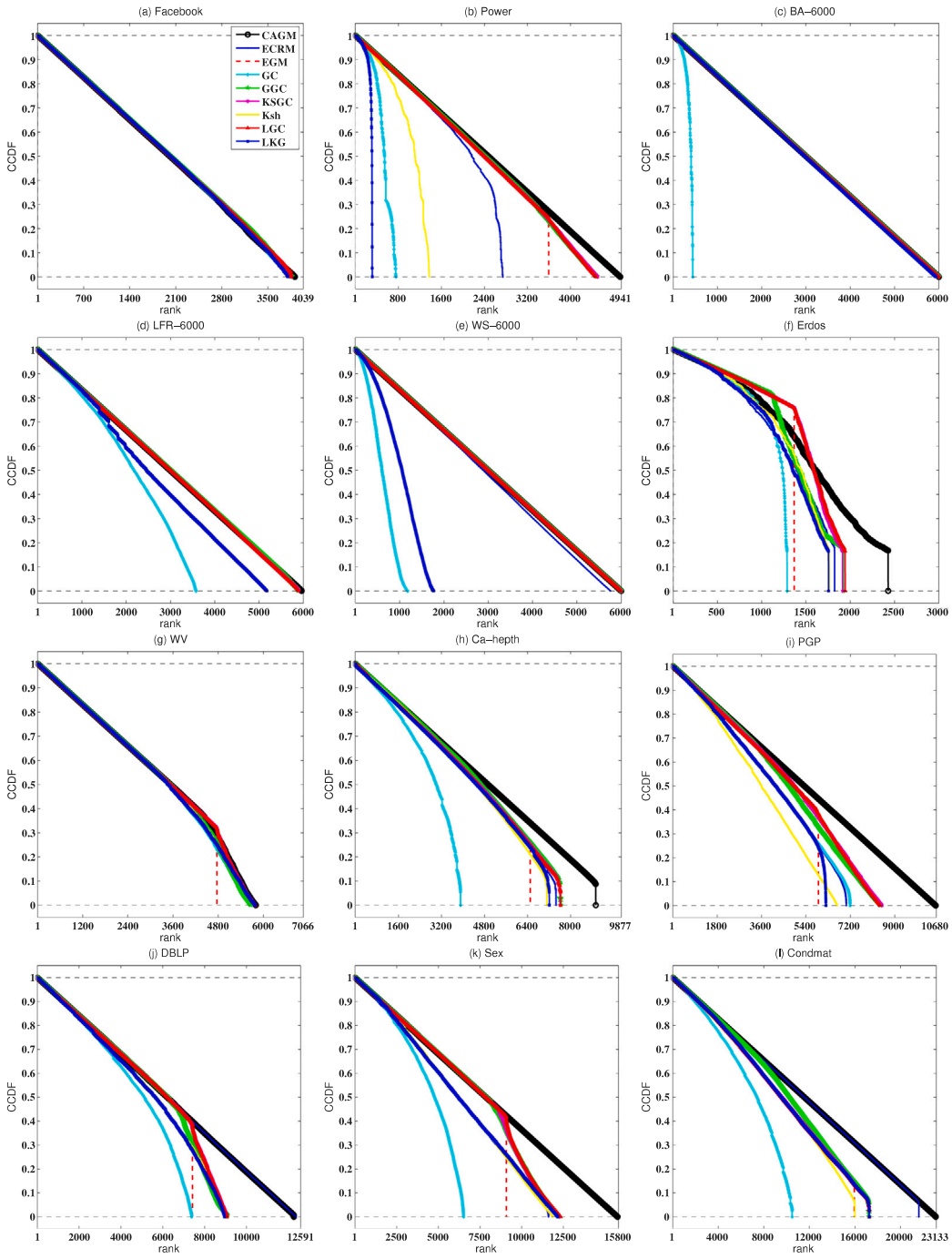


Fig. 7. The complementary cumulative distribution function (CCDF) plots the ranking distributions of ranking list offered by nine algorithms. (a) Facebook, (b) Power, (c) BA-6000, (d) LFR-6000, (e) WS-6000, (f) Erdos, (g) WV, (h) Ca-hepth, (i) PGP, (j) DBLP, (k) Sex, (l) Condmat.

Table 5
The monotonicity value of ranking list generated by nine measures on twelve networks.

Network	CAGM	ECRM	EGM	GC	GGC	KSGC	Ksh	LGC	LKG
Facebook	**0.9999017**	0.9999008	0.9992759	0.9998673	0.9999000	0.9998744	0.9859753	0.9998735	0.9998519
Power	**0.9999813**	0.9915938	0.8807054	0.9731143	0.9998996	0.9998982	0.9769745	0.9998918	0.9000751
BA-6000	**0.9999999**	0.9999914	0.9999989	0.9452974	0.9453488	0.9999998	0.9999996	0.9999998	0.9999125
LFR-6000	**0.9999928**	0.9999839	0.9999828	0.9994585	0.9999817	0.9999840	0.9999797	0.9999832	0.9986433
WS-6000	**0.9999999**	0.9999997	0.9999998	0.9954500	0.9955393	0.9999997	0.9999998	0.9999998	0.9920402
Erdos	**0.9447645**	0.9333520	0.1820281	0.9298875	0.9444052	0.9442183	0.9439524	0.9444563	0.9278455
WV	**0.9997308**	0.9996994	0.8044544	0.9996121	0.9996460	0.9996234	0.9996225	0.9996459	0.9983701

(*continued on next page*)

Table 5 (*continued*).

Network	CAGM	ECRM	EGM	GC	GGC	KSGC	Ksh	LGC	LKG
Ca-hepth	**0.9947113**	0.9934908	0.9076678	0.9893532	0.9936360	0.9936275	0.9935072	0.9936351	0.9827589
PGP	**0.9998547**	0.9996248	0.7110069	0.9993602	0.9997951	0.9997747	0.9995117	0.9997733	0.9919440
DBLP	**0.9999121**	0.9999080	0.7182698	0.9992569	0.9994383	0.9994377	0.9994350	0.9994393	0.9994173
Sex	**0.9998742**	0.9997189	0.6764315	0.9987315	0.9997214	0.9997041	0.9996702	0.9997324	0.9825111
Condmat	**0.9999999**	0.9969649	0.9779095	0.9960961	0.9982163	0.9982161	0.9981697	0.9982149	0.9967155

remaining datasets with a margin as high as 3.38%. Moreover, CAGM obtains the highest score in uniqueness of ranking, achieving almost unique ranking with a monotonicity relation score of more than 0.9997 on ten datasets. As the research continues, we will continue to explore in this direction and strive to investigate the performance of algorithm on other spreading dynamics models, or various types of networks such as temporal networks or heterogeneous networks.

CRediT authorship contribution statement

Guiqiong Xu: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Chen Dong:** Methodology, Validation, Writing – original draft, Revising.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Guiqiong Xu reports financial support was provided by National Natural Science Foundation of China. Guiqiong Xu reports financial support was provided by National Social Science Fund of China.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to sincerely and deeply thank the editor and the anonymous referees for their helpful comments and constructive suggestions. This work was supported by grants from the National Natural Science Foundation of China (Project No. 11871328) and the National Social Science Fund of China (Project No. 22BGL240). All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.121154.

References

Agneessens, F., Borgatti, S. P., & Everett, M. G. (2017). Geodesic based centrality: Unifying the local and the global. *Social Networks*, *49*, 12–26.
Arebi, P., Fatemi, A., & Ramezani, R. (2022). Event stream controllability on event-based complex networks. *Expert Systems With Applications*, *213*, 118886.
Bae, J., & Kim, S. (2014). Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A*, *395*, 549–559.
Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.
Berner, R., Vock, S., Schoell, E., & Yanchuk, S. (2021). Desynchronization transitions in adaptive networks. *Physical Review Letters*, *126*, 028301.
Chen, D., Lü, L., Shang, M., Zhang, Y., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A*, *391*, 1777–1787.
Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik*, *1*, 269–271.
Dong, C., Xu, G., Meng, L., & Yang, P. (2022). CPR-TOPSIS: A novel algorithm for finding influential nodes in complex networks based on communication probability and relative entropy. *Physica A*, *603*, 127797.
Dong, C., Xu, G., Yang, P., & Meng, L. (2023). TSIFIM: A three-stage iterative framework for influence maximization in complex networks. *Expert Systems With Applications*, *212*, 118702.
Estrada, E., & Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, *77*, 036111.
Estrada, E., Hatano, N., & Benzi, M. (2012). The physics of communicability in complex networks. *Physics Reports*, *54*, 89–119.
Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *The BMJ*, *337*, a2338.
Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*, 35–41.
Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*, 215–239.
Goos, H., Kinnunen, M., Salokas, K., Tan, Z., Liu, X., Yadav, L., Zhang, Q., Wei, G., & Varjosalo, M. (2022). Human transcription factor protein interaction networks. *Nature Communications*, *13*, 766.
Gupta, M., & Mishra, R. (2021). Spreading the information in complex networks: Identifying a set of top-N influential nodes using network structure. *Decision Support Systems*, *149*, 113608.
Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 16569–16572.
Ibnoulouafi, A., El Haziti, M., & Cherifi, H. (2018). M-centrality: identifying key nodes based on global position and local degree variation. *Journal of Statistical Mechanics: Theory and Experiment*, *7*, 073407.
Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, *6*, 888–893.
Lambiotte, R., Rosvall, M., & Scholtes, I. (2019). From networks to optimal higher-order models of complex systems. *Nature Physics*, *15*, 313–320.
Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, *78*, 046110.
Leo, K. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *19*, 39–43.
Li, M., Liu, R., Lü, L., Hu, M., Xu, S., & Zhang, Y. (2021). Percolation on complex networks: Theory and application. *Physics Reports*, *907*, 1–68.
Li, Z., Ren, T., Ma, X., Liu, S., Zhang, Y., & Zhou, T. (2019). Identifying influential spreaders by gravity model. *Scientific Reports*, *9*, 8387.
Li, H., Shang, Q., & Deng, Y. (2021). A generalized gravity model for influential spreaders identification in complex networks. *Chaos Solitons Fractals*, *143*, 110456.
Li, S., & Xiao, F. (2021). The identification of crucial spreaders in complex networks by effective gravity model. *Information Science*, *578*, 725–744.
Liu, Y., Song, A., Shan, X., Xue, Y., & Jin, J. (2022). Identifying critical nodes in power networks: A group-driven framework. *Expert Systems With Applications*, *196*, 116557.
Liu, F., Wang, Z., & Deng, Y. (2020). GMM: A generalized mechanics model for identifying the importance of nodes in complex networks. *Knowledge-based Systems*, *193*, 105464.
Liu, Q., Zhu, Y., Jia, Y., Deng, L., Zhou, B., Zhu, J., & Zou, P. (2018). Leveraging local h-index to identify and rank influential spreaders in networks. *Physica A*, *512*, 379–391.
Lü, L., Chen, D., Ren, X., Zhang, Q., Zhang, Y., & Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, *650*, 1–63.
Lu, P., & Dong, C. (2020). EMH: Extended mixing H-index centrality for identification important users in social networks based on neighborhood diversity. *Modern Physics Letters B*, *34*, 2050284.
Lü, L., Zhou, T., Zhang, Q., & Stanley, H. E. (2016). The H-index of a network node and its relation to degree and coreness. *Nature Communications*, *7*, 10168.
Ma, L., Ma, C., Zhang, H., & Wang, B. (2016). Identifying influential spreaders in complex networks based on gravity formula. *Physica A*, *451*, 205–212.
Magdaci, O., Matalon, Y., & Yamin, D. (2022). Modeling the debate dynamics of political communication in social media networks. *Expert Systems With Applications*, *206*, 117782.
Maji, G., Dutta, A., Malta, M. C., & Sen, S. (2021). Identifying and ranking super spreaders in real world complex networks without influence overlap. *Expert Systems With Applications*, *179*, 115061.
Maji, G., Namtirtha, A., Dutta, A., & Malta, M. C. (2020). Influential spreaders identification in complex networks with improved k-shell hybrid method. *Expert Systems With Applications*, *144*, 113092.
Meng, L., Xu, G., Yang, P., & Tu, D. (2022). A novel potential edge weight method for identifying influential nodes in complex networks based on neighborhood and position. *Journal of Computational Science*, *60*, 101591.

Mi, J., Li, Y., Peng, W., & Huang, H. (2018). Reliability analysis of complex multi-state system with common cause failure based on evidential networks. *Reliability Engineering and System Safety*, *174*, 71–81.

Namtirtha, A., Dutta, A., & Dutta, B. (2018). Identifying influential spreaders in complex networks based on kshell hybrid method. *Physica A*, *499*, 310–324.

Namtirtha, A., Dutta, B., & Dutta, A. (2022). Semi-global triangular centrality measure for identifying the influential spreaders from undirected complex networks. *Expert Systems With Applications*, *206*, 117791.

Qiu, L., Zhang, J., & Tian, X. (2021). Ranking influential nodes in complex networks based on local and global structures. *Applied Intelligence*, *51*, 4394–4407.

Samir, A. M., Rady, S., & Gharib, T. F. (2021). LKG: A fast scalable community-based approach for influence maximization problem in social networks. *Physica A*, *582*, 126258.

Shang, Q., Deng, Y., & Cheong, K. H. (2021). Identifying influential nodes in complex networks: Effective distance gravity model. *Information Science*, *577*, 162–179.

Tong, T., Dong, Q., Sun, J., & Jiang, Y. (2023). Vital spreaders identification synthesizing cross entropy and information entropy with kshell method. *Expert Systems With Applications*, *224*, 119928.

Tu, D., Xu, G., & Meng, L. (2021). GPN: A novel gravity model based on position and neighborhood to identify influential nodes in complex networks. *International Journal of Modern Physics B*, *35*, 2150183.

Tulu, M. M., Hou, R., & Younas, T. (2017). Finding important nodes based on community structure and degree of neighbor nodes to disseminate information in complex networks. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 269–273). IEEE.

Ullah, A., Wang, B., Sheng, J., Long, J., Khan, N., & Sun, Z. (2021). Identification of influential spreaders in complex networks. *Expert Systems with Applications*, *186*, Article 115778.

Wang, J., Li, C., & Xia, C. (2018). Improved centrality indicators to characterize the nodal spreading capability in complex networks. *Applied Mathematics and Computation*, *334*, 388–400.

Wang, W., Liu, Q., Liang, J., Hu, Y., & Zhou, T. (2019). Coevolution spreading in complex networks. *Physics Reports*, *820*, 1–51.

Wang, X., Slamu, W., Guo, W., Wang, S., & Ren, Y. (2022). A novel semi local measure of identifying influential nodes in complex networks. *Chaos Solitons Fractals*, *158*, 112037.

Wang, Z., Zhao, Y., Xi, J., & Du, C. (2016). Fast ranking influential nodes in complex networks using a k-shell iteration factor. *Physica A. Statistical Mechanics and its Applications*, *461*, 171–181.

Watts, D., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.

Wen, T., & Deng, Y. (2020). Identification of influencers in complex networks by local information dimensionality. *Information Science*, *512*, 549–562.

Xu, G., & Meng, L. (2023). A novel algorithm for identifying influential nodes in complex networks based on local propagation probability model. *Chaos Solitons Fractals*, *168*, 113155.

Yan, X., Cui, Y., & Ni, S. (2020). Identifying influential spreaders in complex networks based on entropy weight method and gravity law. *Chinese Physics B*, *29*, 048902.

Yang, X., & Xiao, F. (2021). An improved gravity model to identify influential nodes in complex networks based on k-shell method. *Knowledge-based Systems*, *227*, 107198.

Zareie, A., & Sheikhahmadi, A. (2018). A hierarchical approach for influential node ranking in complex social networks. *Expert Systems With Applications*, *93*, 200–211.

Zareie, A., & Sheikhahmadi, A. (2019). EHC: Extended H-index centrality measure for identification of users' spreading influence in complex networks. *Physica A*, *514*, 141–155.

Zareie, A., Sheikhahmadi, A., Jalili, M., & Fasaei, M. S. K. (2020). Finding influential nodes in social networks based on neighborhood correlation coefficient. *Knowledge-based Systems*, *194*, 105580.

Zeng, A., & Zhang, C. (2013). Ranking spreaders by decomposing complex networks. *Physics Letters A*, *377*, 1031–1035.

Zhao, Z., Guo, Q., Yu, K., & Liu, J. (2020). Identifying influential nodes for the networks with community structure. *Physica A*, *551*, 123893.

Zhao, J., Wen, T., Jahanshahi, H., & Cheong, K. H. (2022). The random walk-based gravity model to identify influential nodes in complex networks. *Information Science*, *609*, 1706–1720.

Zhong, S., Zhang, H., & Deng, Y. (2022). Identification of influential nodes in complex networks: A local degree dimension approach. *Information Science*, *610*, 994–1009.