Decision Support Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

Spreading the information in complex networks: Identifying a set of top-*N* influential nodes using network structure

Mukul Gupta, Rajhans Mishra *

Indian Institute of Management Indore, Information Systems Area, Indore, India

ARTICLE INFO

Keywords:
Complex networks
Information propagation
Network structure
Top-*N* influential nodes
Node ranking

ABSTRACT

The real world contains many complex networks, including research networks, social networks, biological networks, and transport networks. Real-world complex networks are unconstrained and can be characterized as undirected and unweighted. Understanding and controlling the process of information propagation in such networks is significant for decision-making activities and has many uses, such as disease control, market advertising, rumor control, and innovation propagation. Identifying the influencers in complex networks is an important activity, as influencers play a key role in spreading information to aid the decision-making process. In this study, we consider the problem of identifying a set of top-*N* influential nodes for spreading the information in undirected and unweighted networks using the network structure in the absence of domain-specific knowledge. In this study, we propose a novel method that computes the ranking scores of the nodes in the network and considers the influence of other nodes simultaneously when forming the set of top-*N* influential nodes. The proposed method is different from other methods of identification of influential nodes in the network, in that it takes into consideration the position of the nodes in the network while computing the ranking score, thereby preventing the clustering of important nodes, which hampers the information flow. Experiments are performed using several real-world complex networks to demonstrate the effectiveness of the proposed method.

1. Introduction

The decision-making process in a network structure depends on the information flow in the network. In real-world complex networks, influencers moderate the flow of information, and identifying the influencers is a critical task for decision support activities [8,20]. Most real-world phenomena, including friendship, biological interaction, transport infrastructure, and research collaboration, can be represented easily and naturally using complex network structures [4,5,21]. The objects in these real-world phenomena are represented as nodes and their interactions as links between those nodes [4,21]. Most of the time, these interactions are bidirectional and have no weightage as such, leading to undirected and unweighted networks like Facebook friendship networks, gene–gene interaction networks, co-authorship networks, and many more [8,9,11]. In these real-world complex networks, understanding and controlling information propagation is significant for many real-world applications, such as disease control, market advertising, rumor control, and innovation propagation [12,16,18].

To control and understand the information propagation behavior of nodes in real-world complex networks, we need to identify influential/important nodes in the network [13,14]. The spread of information is important for decision-making in a complex network. A faster and greater information flow will lead to more accuracy in decision-making in the network [12,15]. Identifying the right set of influencers can support faster information flow, which in turn supports the decision-making process. Different networks have domain-specific properties that result in domain-specific influencer identification processes [17]. However, having a domain-independent method of influencer identification provides versatility; the availability of domain-specific information for various real-world complex networks, like gene–gene interaction networks and behavioral interaction networks, may not be present or may be difficult to obtain, thus limiting the applicability of methods that require domain-specific information to find the influential nodes [11,16].

In this study, the proposed method considers only the network structure for identifying the top-*N* influential nodes in the network and does not require the domain-specific knowledge of the network for the identification of influential nodes. Collection of domain-specific

* Corresponding author.
E-mail addresses: mukulg@iimidr.ac.in (M. Gupta), rajhansm@iimidr.ac.in (R. Mishra).

https://doi.org/10.1016/j.dss.2021.113608
Received 2 January 2021; Received in revised form 12 May 2021; Accepted 29 May 2021
Available online 1 June 2021

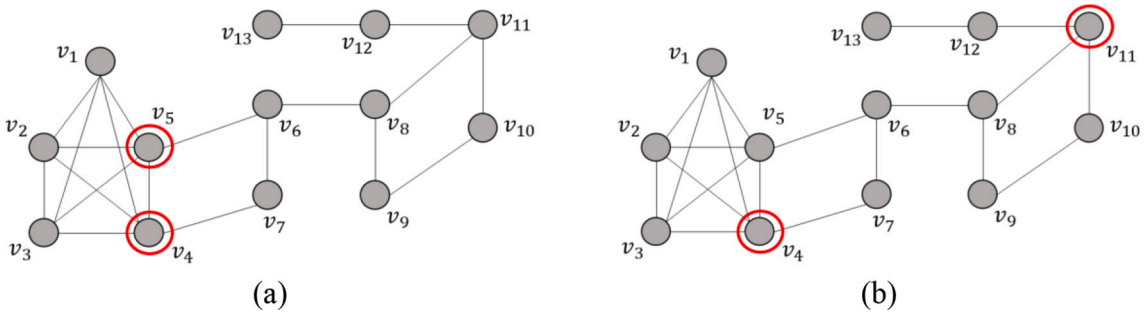

Fig. 1. Finding the top two ($N = 2$) influential nodes in a toy network.

knowledge of a network is cumbersome and may limit the applicability of the domain-specific methods. The primary research aim is to find the influential nodes in a real-world complex network using a method that is domain-independent and therefore can easily be used in real-world complex networks. Identifying the top-N influential nodes is an important activity for controlling and analyzing the information flow in the network.

Most previous studies on finding the top-N influential nodes have computed the ranking scores of nodes in isolation, meaning that the ranking of a node is performed independently of other nodes [18,21]. However, in reality, this is not the case. If high-ranked nodes are in proximity (i.e., clustered in the network and not scattered over the whole network), then the selection of top-N nodes may not be effective, as the coverage from those nodes for information propagation in the network will not be high. However, if the method selects the influential nodes in such a way that the nodes are scattered over the whole network, then the information propagation may be faster [9].

For example, consider the network shown in Fig. 1(a). In the network, there are thirteen nodes and twenty edges. If we find a set of the top two ($N = 2$) influential nodes, then it may be possible that many algorithms would suggest nodes v_4 and v_5 due to the network structure and structural properties of these two nodes. However, it is obvious that information propagation in the network using these two nodes would not be faster, as the nodes are clustered, and therefore the coverage may not be high. In another case, shown in Fig. 1(b), the two nodes that are selected for information propagation are v_4 and v_{11}. Since these two nodes are important from the point of view of network structure and location in the network, the information propagation from these two nodes will be faster than in the previous case. It follows that the selection of the nodes for information propagation in a set should not be carried out in isolation, but that each node should be considered in the set.

We propose a novel method for finding the set of top-N influential nodes in complex real-world undirected and unweighted networks. To find the influential nodes, the proposed method requires only the network structure and thus has wide applicability in different domains without the need for domain-specific knowledge. It performs the k-core decomposition [10] of the network and then computes the normalized iteration multiplier (NIM) to compute the iteration number at which the k-core decomposition is performed for all the nodes in the networks. This gives the global importance of the nodes in the network and how they are surrounded by the other nodes in the network. After computing the NIM, the proposed method computes the normalized global importance (NGI), which considers the degree of the nodes, k-core decomposition values, and NIM to compute the NGI of the nodes. It also computes the relative local–global importance (RLGI) of nodes by considering the immediate neighborhood. Finally, the normalized ranking scores of nodes are computed.

The expected contribution of this study is a novel method for finding a set of top-N influential nodes in real-world complex networks. The proposed method is domain-independent and does not require node- or network-specific information. It uses only the network structure and a parameter-free technique, efficiently handling both local and global information for the network to compute the ranking scores of the nodes and form a set of top-N influential node nodes. The proposed method is novel in the way it uses the immediate neighborhood of nodes to compute their local importance, thereby preventing the clustering of influential nodes in the network. This will enhance the information propagation speed in the network from the set of top-N influential nodes.

The rest of the paper has the following sections. In Section 2, relevant work from the literature is reviewed. In Section 3, the proposed method is presented. The experimental setup is presented in Section 4. Section 5 gives the results of the experiments, which are then discussed in Section 6. Finally, the paper concludes in Section 7 with recommendations for future research directions.

2. Related work

Finding the important/influential nodes in real-world complex networks is an important data mining task, as it has various real-world applications in different domains ranging from social networks to biological networks [9,10,13,20]. Various real-world applications, such as disease control, market advertising, rumor control, and innovation propagation, require the identification of a set of influential nodes in the network so that the information can be propagated quickly and cover the maximum number of nodes in the network [14,16]. In this study, we address the problem of finding a set of top-N influential nodes in undirected and unweighted networks.

Various approaches have been used to find the influential nodes in networks according to their assumptions, and different measures of centrality have been proposed to find the influential nodes in the network according to various criteria. Measures such as degree centrality (DC) [1], closeness centrality (CC) [3], betweenness centrality (BC) [15], and eigenvector centrality (EC) [2] can be used to find the influential nodes. Some of these measures are local and some are global. The local measures use the neighborhood of the node to compute the ranking score, while in the case of global measures the whole network structure is considered for the purposes of computing the ranking score. To apply these measures, knowledge of domain-specific assumptions is required. For example, if a node is considered influential when it is connected to many other nodes in the network, then we can apply the DC measure to rank the nodes [1,19]. However, if a node is considered influential when it is connected to other important nodes, then EC can be applied [2,19]. Thus, these measures are applied effectively if domain-specific information for the network is available.

However, for many real-world networks, domain-specific information is not available or is difficult to acquire, making it difficult to choose the appropriate conventional centrality measures to find the influential nodes [5,19]. Various approaches have been proposed to find the influential nodes in real-world complex networks considering only the network structure. Lü et al. [13] developed a method called LeaderRank

Table 1
Summary of related work.

Method name	Network type				Summary of method	Time complexity (m = number of edges, n = number of nodes in the network)	Limitation(s)
	Directed	Undirected	Weighted	Unweighted			
LeaderRank (LR), Lü et al. [13]	✓			✓	Based on the random walker and utilizes the stochastic matrix to determine the importance of a node	$O((m+2n)\cdot T)$, where T is the number of iterations	This method is a variant of the PageRank algorithm, and it emphasizes the incoming links to the node only. It does not consider the outgoing links of a node while computing its importance.
Extended Gravity Centrality (EGC), Ma et al. [14]		✓		✓	Based on the gravity formula, where it uses k-shell values and the shortest distance between nodes	$O(n^2)$	This method requires the calculation of the shortest distance between nodes, which is time-consuming for large graphs.
Local Gravity Model (LGM), Li et al. [9]		✓		✓	Based on the gravity formula, where it uses degree values and the shortest distance between nodes	$O(n^2)$	This method requires calculation of the shortest distance between nodes, which is time-consuming for large graphs. The free parameter may impact performance.
ProfitLeader (PL), Yu et al. [18]		✓		✓	Based on the profit capacity of nodes, where the profit capacity is computed using sharing probability and available resources	$O(n\cdot\langle k\rangle)$, where $\langle k\rangle$ is the average degree of nodes	The accuracy of the method is not adequate for finding the influential nodes in the network.
Global Importance of Node (GIN), Zhao et al. [21]		✓		✓	Considers self-importance and global importance in computing the node's influence	$O(n^2)$	This method is significantly influenced by closeness centrality, which may affect its accuracy, and it is not applicable to directed and weighted networks.
Global and Local Structure (GLS), Sheng et al. [16]		✓		✓	Considers local and global influence of the nodes in computing their importance	$O(n^2)$	The accuracy of the method is not adequate for finding the influential nodes in the network.
Generalized Mechanics Model (GMM), Liu et al. [11]		✓		✓	Uses local and global information of the node in the network to compute its importance	$O(n_k\cdot(2n^2+m))$	This method requires the computation of eigenvectors and the shortest distance between nodes, which is time-consuming.

Table 2
Important notations.

Notation	Description
$G=(V,E)$	Undirected and unweighted graph
V, E	Set of nodes and set of edges in G
n (=$\lvert V\rvert$), m (=$\lvert E\rvert$)	Number of nodes and number of edges in G
$A\in\{0,1\}^{n\times n}$	Adjacency matrix
$A_{ij}\in\{0,1\}$	Relationship between nodes v_i and v_j
$v_i\in V$	i^{th} node in the network
m_k	Total number of iterations to remove the k-degree nodes in the network
n^k	Number of iterations to remove the k^{th} degree node in the network
δ_i	Normalized iteration multiplier for node v_i
$\Gamma(v_i)$	Neighborhood of node v_i

(LR), which is based on the random walker and uses the stochastic matrix to decide the importance of a node in the network. It is a global measure that utilizes the whole network structure for node ranking. This method is a variant of the PageRank algorithm, and it emphasizes only the incoming links to the node. It does not consider the outgoing links of a node while computing its importance and is applicable also to weighted networks.

Another approach, Extended Gravity Centrality (EGC), proposed by Ma et al. [14], is based on the gravity formula and uses k-shell values and the shortest distance between nodes. This approach is based on the global network structure and can be extended to weighted networks [14]. It requires the calculation of the shortest distance between nodes, which is time-consuming for large graphs. Li et al. [9] proposed a method called the Local Gravity Model (LGM), which is based on the gravity formula and uses degree values and the shortest distance between nodes. LGM also requires the calculation of the shortest distance between nodes, which again is time-consuming for large graphs. Moreover, it has a free parameter that may impact its performance.

The approach proposed by Yu et al. [18], called ProfitLeader (PL), finds the influential nodes in real-world complex networks considering only the network structure. This method is based on the profit capacity of nodes, where the profit capacity is computed using sharing probability and available resources. Although this approach is efficient, the accuracy of the method in finding the influential nodes in the network is not adequate. Zhao et al. [21] proposed an approach based on the global network structure, the so-called Global Importance of Node (GIN). This method considers self-importance and global importance when computing the node influence. It is significantly influenced by the closeness centrality, which may affect its accuracy.

A method based on the local and global structure of the network was then proposed by Sheng et al. [16]: Global and Local Structure (GLS) which considers the local and global influence of the nodes to calculate their importance. However, the accuracy of the method is not adequate; it is also inefficient and time-consuming for large graphs. Liu et al. [11] proposed the Generalized Mechanics Model (GMM) model, which uses both local and global information in the network to compute the node's importance and can be applied to weighted networks. This method involves computing the shortest distance between nodes and requires eigenvectors to be computed, which is time-consuming. Table 1 summarizes these earlier attempts to find the influential nodes in real-world complex networks.

The present study offers a novel method for finding the set of top-N influential nodes in real-world complex networks that takes into account the limitations of earlier attempts. The proposed method uses both local and global information for the nodes in the network to compute their ranking score. It does not require domain-specific information related to nodes and edges in the network to do this, and it is parameter-free.

Therefore, the proposed method is generalizable and can be applied to real-world complex networks; it is efficient and can be applied to large graphs; and it is novel in the way it uses the immediate neighborhood of nodes to compute their local importance, thereby preventing the clustering of influential nodes and enhancing the information propagation speed and coverage in the network.

3. The proposed method

We first present the background of the study to understand the proposed method for finding the top-*N* influential nodes in real-world complex networks.

3.1. Background and preliminaries

The background, mathematical notations, and important definitions are presented in this section. Some important and frequent notations used in this paper are listed in Table 2.

Definition 1 [19] concerns unweighted and undirected networks:

Definition 1 (Unweighted and Undirected Network). An unweighted and undirected real-world complex network is depicted as a graph $G = (V,E)$ and the adjacency matrix $A \in \{0,1\}^{n\times n}$ representing relationships between nodes as follows:

$$A_{ij} = \begin{cases} 1, if\ v_i\ and\ v_j\ have\ relationship \\ 0, if\ v_i\ and\ v_j\ have\ no\ relationship \end{cases}$$

Many real-world networks are undirected and unweighted [4,19]. These networks can be represented using the adjacency matrix. In the adjacency matrix, a value of 1 is used to represent the presence of a relationship between a pair of nodes, and a value of 0 is used to represent the absence of a relationship. Since the relationship has no weightage, only 1 is used to represent its presence. In the case of undirected and unweighted networks, the adjacency matrix is symmetric and binary [19].

In this study, the problem of finding a set of top-*N* influential nodes in an unweighted and undirected real-world complex network is defined as follows [11,16]:

Definition 2 (Influential nodes). For an unweighted and undirected network, $G = (V, E)$, the problem of identification of a set of top-*N* influential nodes for information propagation is to determine the ranking scores of all nodes in the network and select the top-*N* nodes.

In the real world, networks represent the relationships between objects, and the task is to find the set of top-*N* influential nodes (i.e., to find the top-*N* nodes according to ranking scores that would propagate the information in the network very quickly).

3.2. Proposed method

This section discusses the proposed method for finding a set of top-*N* influential nodes in real-world complex networks. An illustration is provided to clarify how the method works.

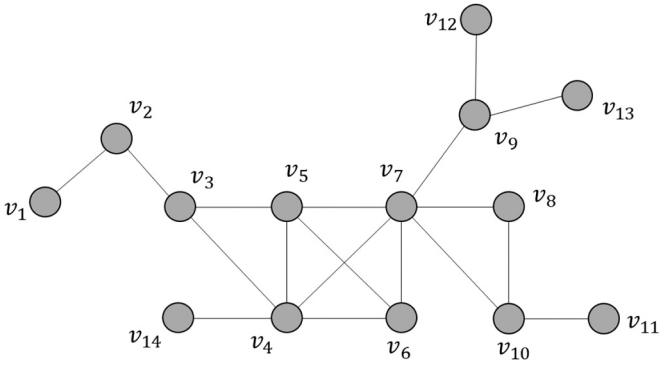

Fig. 2. Toy network for illustration.

The method is novel in terms of using the network structure. Local and global information for nodes is used simultaneously to determine the ranking scores, and the nodes are ranked in such a way that there is no clustering of influential nodes. This improves performance in terms of faster information propagation. The proposed method performs the *k*-core decomposition [10] of the nodes in the network as follows. First, remove all the nodes with degree 1 (1-core decomposition) and keep removing nodes until all the nodes have degree greater than 1. Repeat the process for the rest of the nodes to form different *k*-core values of the network. Then compute the NIM to determine the iteration number at which the node is removed from the network while performing the *k*-core decomposition. This is defined as follows:

Definition 3. Given a complex network G = (V, E), during the process of *k*-core decomposition for the *k*-degree iteration, the total number of iterations is m_k, and node $v_i \in V$ is removed in iteration number n^k. Then $1 \le n^k \le m_k$ and the normalized iteration multiplier (NIM) is defined as

$$\delta_i = \left(1 + \frac{n^k}{max(m_k)}\right) \tag{1}$$

where max(m_k) is the maximum number of total iterations for any *k*.

The NIM computes the normalized iteration number at which a node is removed from the network while performing *k*-core decomposition. By computing the NIM, we determine how important a node is and what its normalized *k*-core value in the network is. A node with a higher NIM is important in the network from a global perspective, as the NIM computes the global significance/importance of the node.

The proposed approach uses the NIM alongside the degree of node and the *k*-core decomposition value to compute the normalized global importance (NGI) as follows:

$$NGI(v_i) = \frac{deg(v_i)*KD(v_i)*\delta_i}{|V|} \tag{2}$$

The NGI computes the global importance by normalizing the NIM using the size of the network. The NGI combines the local and the global significance of the nodes to compute their global importance in the network. From the NGI, the proposed approach then computes the relative local–global importance (RLGI), taking into account the immediate neighborhood of the nodes to prevent clustering of the influential nodes. The RLGI is computed as follows:

$$RLGI(v_i) = \frac{NGI(v_i)*deg(v_i)}{\sum_{\forall j \in \Gamma(v_i)} NGI(v_i)} \tag{3}$$

Using the RLGI, the clustering of important nodes is prevented by discounting the NGI of a node using the total NGI of its neighborhood. If a node a surrounded by other important nodes according to NGI, then its importance is discounted. From this, normalized ranking scores for the nodes are computed, allowing the selection of the top-*N* influential nodes in the network.

The proposed method is effective in finding the set of top-*N* nodes by combining the local and global network structure for the nodes. It is also scalable, as the time complexity is $O(n)$. The computation of the NIM is performed in $O(n)$; since the degree computation of nodes in networks is linear and takes time in the order of $O(n)$ and the *k*-core decomposition takes $O(n)$ time, the computation of NGI will be linear in time complexity and will be $O(n)$. The RLGI requires consideration of the immediate neighborhood of nodes, which takes $O(n)$ time. Hence, the proposed method effectively takes linear time, and the time complexity is $O(n)$. This makes the method effective, and it is faster than many previously proposed methods. The algorithm for the proposed method is given below:

Algorithm

Input:
$G = (V, E)$: undirected and unweighted network
N: number of influential spreaders in the network
Output:
RS_N: top-N influential spreaders in the network
Begin

1. For $\forall v_i \in V$, compute $KD(v_i) \leftarrow K - core\ decomposition(v_i)$
2. For $\forall v_i \in V$, compute δ_i using Eq. (1)
3. Compute normalized global importance (NGI) for $\forall v_i \in V$ as follows:
$$NGI(v_i) = \frac{deg(v_i) * KD(v_i) * \delta_i}{|V|}$$
4. Compute relative local–global importance (RLGI) for $\forall v_i \in V$ as follows:
$$RLGI(v_i) = \frac{NGI(v_i) * \deg(v_i)}{\sum_{\forall j \in \Gamma(v_i)} NGI(v_i)}$$
5. Compute ranking score (RS) for $\forall v_i \in V$ as follows:
$$RS(v_i) = \frac{RLGI(v_i)}{max(RLGI)}$$
6. Return top-N influential nodes RS_N according to the ranking scores.

End

3.3. Illustration

The toy network shown in Fig. 2 clarifies the functioning of the proposed method. The network has fourteen nodes and eighteen edges. We have to find the influential nodes, or, in other words, we have to determine the ranking scores so that we can find the top-N influential nodes from the network.

For the network shown in Fig. 2, we apply the proposed algorithm in a stepwise manner as follows.

Step 1 and Step 2. On the given network, compute the k-core decomposition (KD) and NIM for all the nodes. The values are given below:

Node	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}
$deg(v_i)$	1	2	3	5	4	3	6	2	3	3	1	1	1	1
$KD(v_i)$	1	1	2	3	3	3	3	2	1	2	1	1	1	1
δ_i	1.5	2	1.5	1.5	1.5	1.5	1.5	1.5	2	1.5	1.5	1.5	1.5	1.5

Step 3. Compute the NGI for each node in the network:

Node	$NGI(v_i) = \frac{deg(v_i)*KD(v_i)*\delta_i}{\lvert V\rvert}$	Node	$NGI(v_i) = \frac{deg(v_i)*KD(v_i)*\delta_i}{\lvert V\rvert}$
v_1	$\frac{1*1*1.5}{14} = 0.107$	v_8	$\frac{2*2*1.5}{14} = 0.429$
v_2	$\frac{2*1*2}{14} = 0.286$	v_9	$\frac{3*1*2}{14} = 0.429$
v_3	$\frac{3*2*1.5}{14} = 0.643$	v_{10}	$\frac{3*2*1.5}{14} = 0.643$
v_4	$\frac{5*3*1.5}{14} = 1.607$	v_{11}	$\frac{1*1*1.5}{14} = 0.107$
v_5	$\frac{4*3*1.5}{14} = 1.286$	v_{12}	$\frac{1*1*1.5}{14} = 0.107$
v_6	$\frac{3*3*1.5}{14} = 0.964$	v_{13}	$\frac{1*1*1.5}{14} = 0.107$
v_7	$\frac{6*3*1.5}{14} = 1.929$	v_{14}	$\frac{1*1*1.5}{14} = 0.107$

Step 4. Compute the RLGI:

Node	$RLGI(v_i) = \frac{NGI(v_i)*deg(v_i)}{\sum_{\forall j \in \Gamma(v_i)} NGI(v_i)}$	Node	$RLGI(v_i) = \frac{NGI(v_i)*deg(v_i)}{\sum_{\forall j \in \Gamma(v_i)} NGI(v_i)}$
v_1	$\frac{0.107*1}{0.286} = 0.375$	v_8	$\frac{0.429*2}{2.571} = 0.333$
v_2	$\frac{0.286*2}{0.75} = 0.762$	v_9	$\frac{0.429*3}{2.143} = 0.6$
v_3	$\frac{0.643*3}{3.179} = 0.607$	v_{10}	$\frac{0.643*3}{2.464} = 0.783$
v_4	$\frac{1.607*5}{4.929} = 1.63$	v_{11}	$\frac{0.107*1}{0.643} = 0.167$
v_5	$\frac{1.286*4}{5.143} = 1$	v_{12}	$\frac{0.107*1}{0.429} = 0.25$
v_6	$\frac{0.964*3}{4.821} = 0.6$	v_{13}	$\frac{0.107*1}{0.429} = 0.25$
v_7	$\frac{1.929*6}{5.357} = 2.16$	v_{14}	$\frac{0.107*1}{1.607} = 0.067$

Step 5. From the RLGI, compute the RS of the nodes:

Node	$RS(v_i) = \frac{RLGI(v_i)}{max(RLGI)}$	Node	$RS(v_i) = \frac{RLGI(v_i)}{max(RLGI)}$
v_1	$\frac{0.375}{2.16} = 0.174$	v_8	$\frac{0.333}{2.16} = 0.154$
v_2	$\frac{0.762}{2.16} = 0.353$	v_9	$\frac{0.6}{2.16} = 0.278$
v_3	$\frac{0.607}{2.16} = 0.281$	v_{10}	$\frac{0.783}{2.16} = 0.363$
v_4	$\frac{1.63}{2.16} = 0.755$	v_{11}	$\frac{0.167}{2.16} = 0.077$
v_5	$\frac{1}{2.16} = 0.463$	v_{12}	$\frac{0.25}{2.16} = 0.116$
v_6	$\frac{0.6}{2.16} = 0.278$	v_{13}	$\frac{0.25}{2.16} = 0.116$
v_7	$\frac{2.16}{2.16} = 1$	v_{14}	$\frac{0.067}{2.16} = 0.031$

Step 6. From the RS of the nodes, find the top-N influential nodes in

Table 3
Real-world networks used for experiments.

Dataset	Description	$n(\|V\|)$	$m(\|E\|)$	$\langle k \rangle$	k_{max}	CC	Size of LCC	
							$n(\|V\|)$	$m(\|E\|)$
Zachary-Karate-Club[a]	Karate club network of a university collected by Zachary	34	78	4.5882	17	0.2556	34	78
Jazz[a]	Collaboration network between jazz musicians	198	2742	27.697	100	0.5202	198	2742
USAir97[b]	US Air Flights infrastructure network	332	2126	12.8072	139	0.3963	332	2126
Network-Science[a]	Network of co-authorships in network science	1461	2742	3.7536	34	0.6934	379	914
Wikipedia-Chameleon[c]	Wikipedia page–page network on the topic Chameleon	2277	31,371	27.5547	732	0.3136	2277	31,371
Hamsterster-full[a]	Links between users of the hamsterster.com	2426	16,631	13.7106	273	0.2314	2000	16,098
Oregon-1[c]	Autonomous systems peering information inferred from Oregon route-views from May 26, 2001	11,174	23,409	4.1899	2389	0.0096	11,174	23,409
Bio-CE-CX[b]	Gene functional associations network	15,229	2,45,952	32.6445	375	0.2874	15,063	2,45,862

Note: The table gives the number of nodes $n(=|V|)$, edges $m(=|E|)$, average degree $\langle k \rangle$, maximum degree k_{max}, clustering coefficient (CC), and size (number of nodes and edges) of the largest connected component (LCC).

[a] http://konect.cc/networks/

[b] http://networkrepository.com/

[c] https://snap.stanford.edu/data/

the network by sorting the nodes according to the RS:

Node	$RS(v_i)$	Rank	Node	$RS(v_i)$	Rank
v_7	1	1	v_6, v_9	0.278	7
v_4	0.755	2	v_1	0.174	8
v_5	0.463	3	v_8	0.154	9
v_{10}	0.363	4	v_{12}, v_{13}	0.116	10
v_2	0.353	5	v_{11}	0.077	11
v_3	0.281	6	v_{14}	0.031	12

From the above table, we see that v_7 is the most influential node in the network. The spreading ability of v_7 is the highest in the network as per the structural information. Nodes v_6 and v_9 and nodes v_{12} and v_{13} are ranked as having equal importance according to the structure of the network.

4. Experimental setup

In this section, we discuss the experimental setup applied to eight real-world complex network datasets from various domains ranging from social networks to biological networks. All the experiments in this study to find the influential nodes in the networks were performed on a system using R version 4.0.2.

4.1. Datasets

The experiments use real-world datasets that are undirected and unweighted networks from different domains. For these networks, the number of nodes, number of edges, average degree, maximum degree of nodes, and clustering coefficients are listed in Table 3. In this study, we consider the largest connected components of the networks, as the problem under study is how to find the influential nodes for information propagation. The size of the largest connected component (i.e., the number of edges and the number of nodes) is given in Table 3. The log–log degree distribution of the eight datasets is shown in Fig. 3.

The Zachary-Karate-Club is a very famous and highly utilized dataset for various mining tasks. That contains the Karate club network of a university collected by Zachary. In this network, the 34 nodes represent the club members, and the 78 edges represent their interactions outside the club. The Jazz network is a collaboration network between Jazz musicians that represents the musicians as nodes and the collaborations between them as edges. A collaboration occurs if two musicians play together in a band. This network has 198 nodes representing the musicians and 2742 edges representing the collaborations between them. USAir97 is an airport transport network dataset with 332 nodes and 2126 edges representing the paths between the source and destination. The Network-Science dataset is a network of co-authorships in the network science domain with 1461 nodes representing researchers and 2742 edges representing co-authorship.

The Wikipedia-Chameleon dataset has 2277 and 31,371 edges. The nodes in the network represent Wikipedia pages and the edges represent links between pages. The Hamsterster-full network dataset has 2426 nodes representing users of the Hamsterster.com website, with 16,631 edges representing the links between them. The Oregon-1 network dataset is a network of Autonomous Systems peering information inferred from Oregon route-views from May 26, 2001. This dataset has 11,174 nodes and 23,409 edges. From the domain of biology, the Bio-CE-CX has 15,229 nodes and 2,45,952 edges; the nodes represent genes, and the edges represent the associations of genes in the network.

In the present study, to find the set of top-N influential nodes in each network, the connected graphs were utilized. Of the eight datasets, the Network-Science, Hamsterster-full, and Bio-CE-CX datasets are not connected. Accordingly, we pre-processed these datasets for experiments, taking the largest connected component (listed in Table 3), and performing a comparison of the proposed method with other algorithms.

4.2. Performance evaluation

The performance of the proposed method was then compared with that of several existing techniques for finding the set of top-N influential nodes in a network. These techniques are applicable to unweighted and undirected complex networks and therefore highly relevant to the problem under study. They are also domain-independent and do not require any domain-specific information/assumption for nodes and edges.

- Leader Rank (LR): The leader rank technique is based on a random walk model for finding the influential nodes in real-world complex networks. It is parameter-free and utilizes the stochastic matrix to decide the importance of a node in the network [13].
- Extended Gravity Centrality (EGC): This method is based on the gravity formula where it uses k-shell values and the shortest distance between nodes [14]. The k-shell values are taken as the mass of the nodes, and the distances of the shortest paths are taken as the distance between the nodes when computing the gravity.
- Local Gravity Model (LGM): This is based on the gravity formula, using degree values and the shortest distance between nodes [9]. The

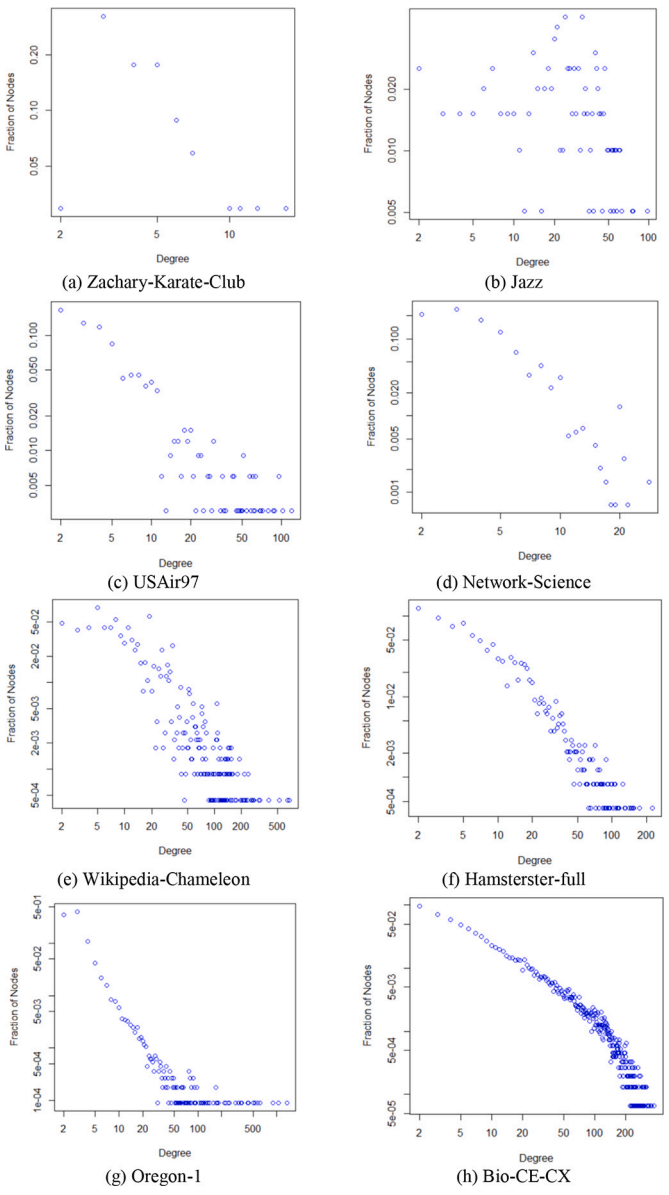

Fig. 3. Log–log degree distribution of the datasets.

LGM uses both neighborhood and path information to measure the importance of nodes in the network.

- ProfitLeader (PL): This method is based on the profit capacity of nodes, where profit capacity is computed using sharing probability and available resources. [18]. This approach is based on the concept of profit given by a node to other nodes in the network. The more profit a node offers to other nodes in the network, the more important it is in the network.
- Global Importance of Node (GIN): This method considers self-importance and global importance while computing the node influence. It is significantly influenced by closeness centrality [21] and can identify the seemingly unimportant nodes that are influential in the network.
- Global and Local Structure (GLS): This method considers the local as well as the global structure of the network. The closeness of nodes to all other nodes in the network is used as a measure of global influence, while closeness to nearest neighbors is used as a measure of local influence [16].
- Generalized Mechanics Model (GMM): The proposed model uses both local and global information for the node in the network to compute its importance in the network. It requires eigenvectors and the shortest distance between nodes to be computed [11].

In addition to the above-mentioned methods, we also considered degree centrality (DC) [1], EC [2], and k-shell centrality (KSC) [7] as the baseline methods for the comparison with the proposed approach.

To evaluate the different algorithms, we used the monotonicity index $M(X)$ for a ranking list X of nodes in the network [14], as given below:

$$M(X) = \left[1 - \frac{\sum_{c \in V} N_c(N_c - 1)}{N(N-1)}\right]^2 \tag{4}$$

where the size of the network is N and the number of nodes with the same rank c is N_c. If $M(X) = 1$, the ranking algorithm has given a distinct rank to all the nodes in the network (high resolution of the algorithm); if $M(X) = 0$, the ranking algorithm has given the same rank to all the nodes in the network (zero resolution).

In addition to monotonicity, the complementary cumulative distribution function (CCDF) was utilized for better evaluation of the differentiation of the ranking list of nodes in the network [6]. The CCDF function captures the distribution of the nodes in the network to different ranks. Methods that rank fewer nodes in the same rank are considered good from the point of view of differentiating the nodes. This is helpful for selecting the top-N nodes from a network based on their rank; if a larger number of nodes have the same rank, then it will not be possible to select a node for that rank value. If more nodes are cumulated in the same rank, then the CCDF function will approach zero quickly, and the slope will be steep [6]. In contrast, uniformly distributed nodes for different ranks will cause the CCDF to have a gradual slope [6].

Differentiation of nodes based on their rank is not sufficient to show the success of a ranking algorithm. Therefore, in addition to monotonicity and the CCDF, we evaluated the methods based on their ability to spread the information in the networks. A node located at an important position in the complex network has a strong infectious ability and is considered influential [21]. In this study, to evaluate the

Table 4
Monotonicity values for eight real-world datasets.

Method	Dataset							
	Zachary-Karate-Club	**Jzz**	**USAir97**	**Network-Science**	**Wikipedia-Chameleon**	**Hamsterster-full**	**Oregon-1**	**Bio-CE-CX**
Proposed	0.976961	0.999282	0.998363	0.997112	0.999465	0.999673	0.999877	0.999988
LR	0.950712	0.990537	0.994439	0.994799	0.999639	0.999186	0.999822	0.999868
EGC	0.948975	0.999641	0.99789	0.998437	0.999532	0.999909	0.999878	0.999994
LGM	1	0.999179	0.998417	0.997405	0.999472	0.999656	0.999894	0.999998
PL	0.980488	0.999282	0.998308	0.997042	0.999466	0.999622	0.999871	0.999982
GIN	1	0.999282	0.998381	0.997405	0.999471	0.999656	0.999893	0.999998
GLS	0.980488	0.999282	0.998417	0.997419	0.999472	0.999652	0.999894	0.999995
GMM	1	0.999282	0.998690	0.997488	0.999551	0.999795	0.999896	0.999998
KSC	0.942041	0.989924	0.994003	0.994716	0.999125	0.998999	0.999821	0.999867
EVC	0.980488	0.999384	0.998817	0.999037	0.999815	0.999792	0.999904	0.999999
DC	0.950711	0.990536	0.994438	0.994799	0.99914	0.999013	0.999821	0.999867

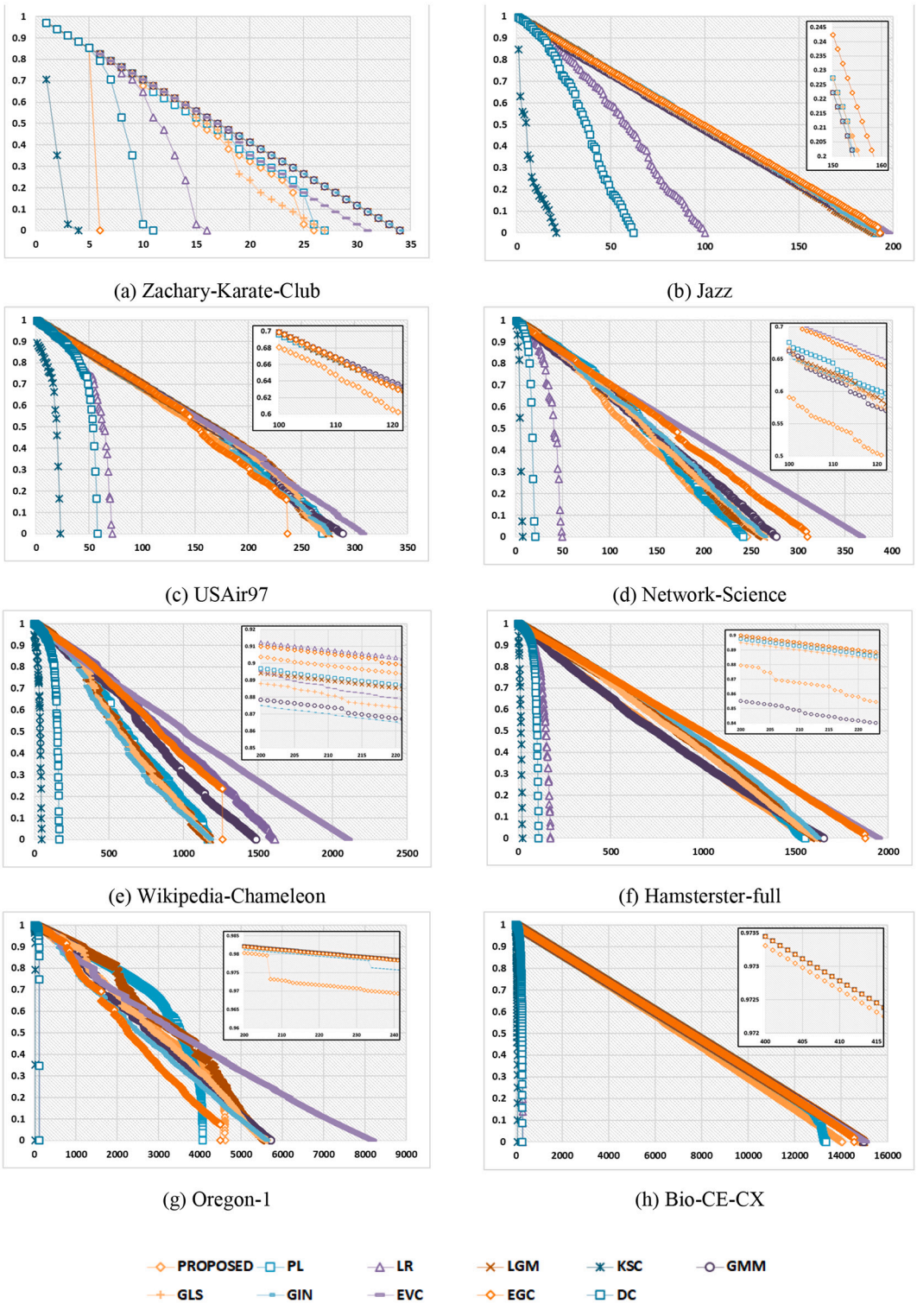

Fig. 4. Complementary cumulative distribution function (CCDF) plots for eight real-world datasets.

infectious ability of the set of top-*N* nodes in the complex networks, the SI (susceptible–infectious) model was adopted [11,21]. There is a positive correlation between the infectious ability of a node in the network and its importance in the network [11]. This means that if a node is important, its ability to spread the infection in the network is high [11,21].

Accordingly, to evaluate the methods, the average infection ability of top-*N* nodes was considered. The top-10 nodes ($N = 10$) were selected as initially infected nodes, and the remaining nodes were considered as susceptible nodes in the SI model. At each step, the infectious nodes spread infection to their neighbors according to the spreading probability (β) value. As we used the top-*N* nodes for information dissemination in the network, the focus was on the top-*N* nodes in the network, where the value of *N* depends on the application and domain. The value of *N* is usually not very large, as in real-world scenarios the section and utilization of these influential nodes for information dissemination is

Table 5
Spread of infection, Zachary-Karate-Club dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	**43.16**	**54.21**	**63.02**	**69.96**	**75.73**	**80.43**	**84.19**	**87.22**	**89.36**	**91.28**
LR	40.81	50.46	58.45	65.22	71.08	75.83	79.95	83.41	86.15	88.63
EGC	41.3	51.37	59.74	66.66	72.75	77.75	81.67	85.04	87.78	90.04
LGM	40.56	50.14	58.08	64.97	70.85	75.66	79.8	83.26	86.21	88.65
PL	40.81	50.46	58.45	65.22	71.08	75.83	79.95	83.41	86.15	88.63
GIN	40.56	50.19	58.33	65.26	71.21	76.24	80.29	83.82	86.41	88.68
GLS	40.81	50.46	58.45	65.22	71.08	75.83	79.95	83.41	86.15	88.63
GMM	39.97	49.13	57.12	63.91	69.8	74.95	79.24	82.86	85.74	88.21
KSC	38.96	47.6	55.22	62.06	68.07	73.39	77.99	81.74	84.61	87.26
EVC	39.81	48.89	56.86	63.7	69.78	74.91	79.26	82.84	85.75	88.03
DC	40.81	50.46	58.45	65.22	71.08	75.83	79.95	83.41	86.15	88.63

Table 6
Spread of infection, Wikipedia-Chameleon dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	24.48	49.19	64.24	**71.98**	**76.77**	**80.4**	**83.19**	**85.56**	**87.56**	**89.17**
LR	32.77	54.44	**64.68**	70.87	75.33	78.81	81.76	83.94	85.99	87.55
EGC	14.12	25.4	38.27	50.78	60.36	67.1	71.93	75.59	78.44	80.66
LGM	32.67	**54.65**	64.5	70.76	75.35	78.9	81.79	84.02	86.02	87.71
PL	14.76	26.09	39.47	52.26	61.71	67.72	72.44	75.99	78.77	80.95
GIN	31.78	54.17	64.23	70.6	75.26	78.88	81.63	83.98	85.83	87.54
GLS	**32.95**	54.57	64.5	70.73	75.19	78.76	81.72	83.86	85.77	87.36
GMM	15.04	26.35	39.38	52.9	61.73	68.08	72.87	76.3	79.12	81.33
KSC	12.46	22.14	33.31	45.46	55.79	63.43	68.68	72.83	76.13	78.69
EVC	14.79	26.41	39.77	52.42	61.56	67.87	72.45	76.05	78.81	81.11
DC	32.77	54.44	**64.68**	70.87	75.33	78.81	81.76	83.94	85.99	87.55

Table 7
Spread of infection, USAir97 dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	28.73	**45.27**	**55.69**	**63.26**	**69.12**	**73.68**	**77.39**	**80.44**	**83.08**	**85.21**
LR	**28.85**	44.91	54.89	62.14	67.71	72.17	75.78	78.74	81.31	83.44
EGC	25.76	41.96	52.19	59.74	65.63	70.24	74.06	77.27	79.8	82.11
LGM	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
PL	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
GIN	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
GLS	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
GMM	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
KSC	25.56	41.9	52.28	59.93	65.66	70.29	74.07	77.28	79.86	82.19
EVC	28.64	44.55	54.47	61.69	67.36	71.69	75.31	78.39	80.96	83.13
DC	**28.85**	44.91	54.89	62.14	67.71	72.17	75.78	78.74	81.31	83.44

Table 8
Spread of infection, Hamsterster-full dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	16.89	**39.49**	**54.63**	**64.58**	**71.27**	**76.16**	**79.81**	**82.71**	**85.07**	**86.98**
LR	**17.36**	38.84	53.78	63.89	70.75	75.72	79.52	82.46	84.87	86.8
EGC	13.6	34.11	49.69	60.58	68.14	73.69	77.84	80.99	83.63	85.76
LGM	**17.36**	38.84	53.78	63.89	70.75	75.72	79.52	82.46	84.87	86.8
PL	16.47	37.26	52.38	62.7	69.92	75.09	78.94	82.05	84.47	86.48
GIN	16.84	37.99	52.9	63.11	70.21	75.3	79.16	82.16	84.64	86.6
GLS	16.31	36.9	51.93	62.26	69.55	74.74	78.69	81.79	84.3	86.28
GMM	0.55	0.61	0.69	0.8	1.01	1.29	1.86	2.47	3.52	5.24
KSC	3.84	15.56	33.69	48.46	58.82	66.18	71.73	75.9	79.22	81.94
EVC	16	36.5	51.57	62.02	69.31	74.58	78.59	81.7	84.19	86.24
DC	**17.36**	38.84	53.78	63.89	70.75	75.72	79.52	82.46	84.87	86.8

Table 9
Spread of infection, Oregon-1 dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	**8.53**	**17.91**	**26.69**	**34.69**	**41.89**	**48.32**	**54.05**	**59.13**	**63.63**	**67.63**
LR	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5
EGC	6.85	15.95	24.54	32.46	39.65	46.08	51.85	57.04	61.65	65.77
LGM	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5
PL	7.95	17.23	25.96	33.9	41.09	47.5	53.23	58.33	62.86	66.91
GIN	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5
GLS	8.43	17.79	26.55	34.53	41.72	48.14	53.86	58.92	63.43	67.45
GMM	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5
KSC	7.27	16.28	24.9	32.81	39.97	46.42	52.12	57.3	61.9	66
EVC	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5
DC	8.49	17.81	26.61	34.58	41.77	48.2	53.91	59	63.48	67.5

Table 10
Spread of infection, Network-Science dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	**7.75**	**13.48**	**19.5**	**25.6**	**31.48**	**37.15**	**42.53**	**47.58**	**52.11**	**56.36**
LR	7.49	12.89	18.48	24.12	29.77	35.27	40.4	45.41	50.08	54.22
EGC	3.77	5.09	6.51	8.02	9.53	11.15	12.82	14.55	16.14	17.85
LGM	7.33	12.47	17.78	23.13	28.38	33.59	38.52	43.3	47.85	52
PL	7.33	12.47	17.78	23.13	28.38	33.59	38.52	43.3	47.85	52
GIN	6.83	11.28	15.96	20.62	25.33	30.03	34.74	39.08	43.21	47.29
GLS	7.33	12.47	17.78	23.13	28.38	33.59	38.52	43.3	47.85	52
GMM	4.4	6.24	8.16	10.11	12.12	14.19	16.22	18.19	20.13	22.04
KSC	4.84	7.25	9.88	12.62	15.36	18.21	21.11	23.81	26.63	29.29
EVC	4.28	6	7.76	9.54	11.4	13.26	15.19	16.93	18.79	20.71
DC	7.49	12.89	18.48	24.12	29.77	35.27	40.4	45.41	50.08	54.22

Table 11
Spread of infection, Jazz dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	**50.83**	**79.85**	**88.67**	**92.02**	**93.77**	**94.84**	**95.68**	**96.33**	**96.8**	**97.26**
LR	48.12	76.14	86.53	90.74	92.83	94.12	95.01	95.73	96.33	96.77
EGC	39.18	70.04	83.68	89.4	91.92	93.48	94.49	95.25	95.96	96.45
LGM	48.13	76.31	86.72	90.84	92.91	94.15	95.08	95.83	96.36	96.78
PL	47.75	75.86	86.49	90.76	92.8	94.07	95.01	95.72	96.33	96.74
GIN	48.13	76.31	86.72	90.84	92.91	94.15	95.08	95.83	96.36	96.78
GLS	47.58	76.04	86.52	90.7	92.81	94.1	95.02	95.75	96.38	96.79
GMM	6.38	11.96	24.54	41.53	58.61	72.69	82.62	89.82	93.73	96.26
KSC	40.94	70.26	83.37	89.13	91.83	93.28	94.39	95.19	95.8	96.29
EVC	47.78	76.1	86.62	90.76	92.85	94.11	95.07	95.78	96.33	96.82
DC	48.12	76.14	86.53	90.74	92.83	94.12	95.01	95.73	96.33	96.77

Table 12
Spread of infection, Bio-CE-CX dataset.

Method	Probability of spreading (β)									
	0.01 (%)	0.02 (%)	0.03 (%)	0.04 (%)	0.05 (%)	0.06 (%)	0.07 (%)	0.08 (%)	0.09 (%)	0.1 (%)
Proposed	**22.64**	**47.84**	**59.86**	**67.21**	**72.42**	**76.35**	**79.46**	**81.98**	**84.09**	**85.84**
LR	22.04	47.41	59.62	67.04	72.26	76.22	79.36	81.9	84	85.76
EGC	18.24	44.41	57.6	65.47	70.94	75.07	78.32	80.96	83.14	84.99
LGM	22.07	47.43	59.61	67.05	72.28	76.23	79.35	81.9	84.01	85.76
PL	19.76	45.61	58.43	66.16	71.54	75.6	78.79	81.4	83.56	85.36
GIN	21.91	47.32	59.53	66.98	72.2	76.16	79.3	81.84	83.95	85.71
GLS	19.3	45.27	58.2	65.96	71.38	75.47	78.68	81.29	83.46	85.29
GMM	19.73	45.72	58.48	66.17	71.56	75.63	78.83	81.42	83.58	85.39
KSC	8.03	34.7	51.2	60.69	67.06	71.77	75.45	78.4	80.85	82.92
EVC	19.12	45.1	58.05	65.83	71.26	75.35	78.57	81.18	83.36	85.19
DC	22.04	47.41	59.62	67.04	72.26	76.22	79.36	81.9	84	85.76

costly in terms of time and effort required) [16,18].

5. Results

In this section, the results of experiments performed on eight datasets (Zachary-Karate-Club, Jazz, USAir97, Network-Science, Wikipedia-Chameleon, Hamsterster-full, Oregon-1, and Bio-CE-CX) are presented for monotonicity, CCDF, and SI infectious model.

5.1. Monotonicity

From Table 4, we see the monotonicity values of various methods for different datasets. The monotonicity index values are very close to 1 (higher the value of M better it is in terms of resolution), which indicates that all the methods were able to give distinct ranks to a significant number of nodes in the network.

From these results, it is clear that the proposed method is effective for resolution of a significant number of nodes in these networks; therefore, using the proposed method, we can select top-*N* nodes without any problem for even a large value of *N*. However, it is not possible to compare the methods using the monotonicity index alone, as the differences between the monotonicity values are not significant. We therefore include CCDF and SI model in the comparison.

5.2. CCDF

Fig. 4 shows the CCDF plots for the eight datasets. It is clear that the proposed method is able to differentiate the nodes on the basis of the network structure, and that the ranking scores given to nodes in the network will be different in a large number of cases. Profit Leader (PL) was able to give different ranking scores to a larger number of nodes in the network than the other methods. From this point of view, degree centrality (DC) and *k*-shell centrality (KSC) were the worst-performing methods.

We see that the proposed method was largely able to differentiate the nodes in the network by giving different ranking scores to a large number of nodes. In the real world, for top-*N* influential nodes selection, the value of *N* is typically taken as between 10 and 200, depending on the size of the network and the associated cost of selecting the nodes as initial spreaders [11,21]. Considering the practical aspects of the influential node selection problem, we can conclude that the proposed method has performed satisfactorily.

5.3. SI infectious model

The results for spreading ability using the SI infectious model are shown in Table 5, 6, 7, 8, 9, 10, 11 and 12. The percentage coverage of each network for different values of spreading probability (β) is given in the tables. Higher coverage (spread) values indicate that the method propagates the information in the network efficiently. For the experiments, the value of β was varied from 0.01 to 0.1 for different spreading phenomena. A higher value of β indicates that the neighboring nodes have a higher chance of being infected from the currently infected nodes in the network. For SI infectious model, we considered time steps *t* from 1 to 10, and each experiment for a given value of β was repeated 100 times. The average values are reported in the results [11,21]. To simulate the information-spreading abilities of the nodes and determine which method best finds the influential nodes in the network, ten time steps were taken, as in previous work [11]. By taking an appropriate value for the number of iterations, we were able to compare the results of using each method.

It is clear that the proposed method outperformed all the comparison methods. For the Zachary-Karate-Club, Oregon-1, Network-Science, Jazz, and Bio-CE-CX datasets, for all values of spreading probabilities, the spread percentage from the top 10 nodes from the proposed method was higher than for all the other methods. For the USAir97 and Hamsterster-full datasets, the proposed method outperformed all the other methods when the value of spreading probability was greater than 0.01. For the Wikipedia-Chameleon dataset, the proposed method outperformed all the other methods when the value of the spreading probability was greater than 0.03. For all the datasets, if the spreading condition was appropriate, the proposed method outperformed all the other methods. It is also clear that the proposed method can spread information in a real-world complex network for a wide range of spreading conditions simulated through the different values of spreading probabilities.

6. Discussion

In this section, we discuss the results of the experiments. As we see from the monotonicity index values and the CCDF plots, the proposed method is able to differentiate the nodes in the network by giving them distinct ranking scores. If a method gives distinct ranking scores to the nodes in a network, then it is possible to select the top-*N* influential nodes in that network; otherwise, it is not possible to select the top-*N* nodes, as there will be many nodes for any particular position. From the plots, we see that although the proposed method is not the best of all the methods, it is able to select distinct nodes for a sufficiently large value of *N* respective to a network. This demonstrates the usability of the proposed method in real-world applications.

However, the monotonicity index and CCDF are not enough to establish the effectiveness of a given method of information propagation, hence we have also used the SI infectious model. The experimental results show that the proposed method outperforms all the other methods. It is effective for a wide range of probabilities, which demonstrates its effectiveness for the selection of top-*N* nodes in the network for information propagation.

We also see that the proposed method is effective in finding the influential nodes in the networks. It computes the local as well as the global importance, and on that basis, assigns ranking scores to nodes and thus this work adds to the existing body of literature.

It is parameter-free and independent of domain-specific assumptions, which makes it useful in practical terms for different real-world complex networks, making it very useful for the implementation purpose. The proposed method is also efficient, can be applied to large graphs, and is novel in the way it uses the immediate neighborhood of nodes to compute their local importance, thereby preventing the clustering of influential nodes in the network.

7. Conclusions and future research

This study develops a novel method for finding the top-*N* influential nodes in unweighted and undirected real-world complex networks. The proposed method uses only the network structure and does not require domain-specific information related to nodes or networks. It is effective in finding the influential nodes and can propagate information in the network rapidly. It is also scalable, as it has linear time complexity and can be applied to large real-world complex networks. It uses the local and global network structure to find the ranking score of nodes and thereby identify the set of top-*N* influential nodes in the network. This avoids clustering of the influential nodes in the network and is therefore more effective for information propagation than the other comparison methods. The experiments confirm the effectiveness of the method using several real-world complex networks.

A limitation of the proposed method is that it is applicable only to unweighted and undirected real-world networks; real-world networks may be weighted or directed, or both weighted and directed. Therefore, future research could develop the model by applying it to weighted and directed networks to find the top-*N* influential nodes. Another interesting extension of the proposed approach would be to find the seed nodes in social networks for discovering polarized communities.

References

[1] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, J. Math. Sociol. 2 (1) (1972) 113–120.
[2] P. Bonacich, P. Lloyd, Eigenvector-like measures of centrality for asymmetric relations, Soc. Networks 23 (3) (2001) 191–201.
[3] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Networks 1 (3) (1978) 215–239.
[4] M. Gupta, P. Kumar, Recommendation generation using personalized weight of meta-paths in heterogeneous information networks, Eur. J. Oper. Res. 284 (2) (2020) 660–674.
[5] M. Gupta, R. Mishra, Network projection-based edge classification framework for signed networks, Decis. Support. Syst. (2020) 113321.
[6] J.C. Helton, Probability, conditional probability and complementary cumulative distribution functions in performance assessment for radioactive waste disposal, Reliab. Eng. Syst. Saf. 54 (2–3) (1996) 145–163.
[7] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[8] P. Kumar, S. Gupta, B. Bhasker, An upper approximation based community detection algorithm for complex networks, Decis. Support. Syst. 96 (2017) 103–118.
[9] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, T. Zhou, Identifying influential spreaders by gravity model, Sci. Rep. 9 (1) (2019) 1–7.
[10] J.H. Lin, Q. Guo, W.Z. Dong, L.Y. Tang, J.G. Liu, Identifying the node spreading influence with largest k-core values, Phys. Lett. A 378 (45) (2014) 3279–3284.
[11] F. Liu, Z. Wang, Y. Deng, GMM: a generalized mechanics model for identifying the importance of nodes in complex networks, Knowl.-Based Syst. 193 (2020) 105464.
[12] D. Lu, Q. Li, S.S. Liao, A graph-based action network framework to identify prestigious members through member's prestige evolution, Decis. Support. Syst. 53 (1) (2012) 44–54.
[13] L. Lü, Y.C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS One 6 (6) (2011), e21202.
[14] L.L. Ma, C. Ma, H.F. Zhang, B.H. Wang, Identifying influential spreaders in complex networks based on gravity formula, Physica A: Statistical Mechanics and its Applications 451 (2016) 205–212.
[15] M.E. Newman, A measure of betweenness centrality based on random walks, Soc. Networks 27 (1) (2005) 39–54.
[16] J. Sheng, J. Dai, B. Wang, G. Duan, J. Long, J. Zhang, K. Guan, S. Hu, L. Chen, W. Guan, Identifying influential nodes in complex networks based on global and local structure, Physica A: Statistical Mechanics and its Applications 541 (2020) 123262.
[17] P. Wang, J. Lü, X. Yu, Identification of important nodes in directed biological networks: a network motif approach, PLoS One 9 (8) (2014), e106132.
[18] Z. Yu, J. Shao, Q. Yang, Z. Sun, ProfitLeader: identifying leaders in networks with profit capacity, World Wide Web 22 (2) (2019) 533–553.
[19] R. Zafarani, M.A. Abbasi, H. Liu, Social Media Mining: An Introduction, Cambridge University Press, 2014.
[20] A. Zareie, A. Sheikhahmadi, K. Khamforoosh, Influence maximization in social networks based on TOPSIS, Expert Syst. Appl. 108 (2018) 96–107.
[21] J. Zhao, Y. Wang, Y. Deng, Identifying influential nodes in complex networks from global perspective, Chaos, Solitons Fractals 133 (2020) 109637.

Mukul Gupta is currently working as an Assistant Professor in the Information Systems area at the Indian Institute of Management Indore, India. He received his Ph.D. in Information Technology and Systems area from the Indian Institute of Management Lucknow, India. He did his M.Tech from Dayalbagh Educational Institute, India, in Computer Science and B.Tech in Computer Science and Engineering. His current research interest includes e-Commerce, Recommendation Systems, Information Networks, Machine Learning, Social Media Analytics, Web and Data Mining.

Rajhans Mishra is an Associate Professor in the Information Systems Area at the Indian Institute of Management Indore (India). He has also served as a visiting faculty at the Indian Institute of Management Ahmedabad and Indian Institute of Management Lucknow. His research interest includes recommendation systems, web mining, data mining, text mining, e-Governance and business analytics.