Computing (2022) 104:359–382
https://doi.org/10.1007/s00607-021-00960-8

REGULAR PAPER

MDER: modified degree with exclusion ratio algorithm for influence maximisation in social networks

Sanjay Kumar[1,2] · Dipti Lohia[1] · Darsh Pratap[1] · Ashutosh Krishna[1] · B. S. Panda[2]

Received: 10 May 2020 / Accepted: 13 May 2021 / Published online: 7 June 2021

Abstract

The online social network has become an integral part of our day to life and serves as an excellent platform for sharing ideas, opinions, and products. Influence maximization (IM) is a widely studied topic in the area of social network analysis. The objective of IM is to find influential nodes that can disseminate information to a larger extent in the network. Many local and global centrality measures are proposed to rank the nodes based on their spreading capability with certain limitations. Many proposed algorithms locate the spreaders sharing overlapping regions or are closely placed, which may cause interference in spreading. In this paper, based on the notion of maximum coverage of the information and minimum interference in spreading, we propose a novel semi-local algorithm named as *modified degree centrality with exclusion ratio* to identify influential nodes from diverse locations in the network. We use modified degree centrality by considering neighbours upto 2-hops and introduce the novel idea of exclusion ratio to ensure minimum overlapping between regions influenced by the chosen spreader nodes. Extensive experimental validation using classical informa-

✉ Sanjay Kumar
sanjay.kumar@dtu.ac.in

Dipti Lohia
diptilohia083@gmail.com

Darsh Pratap
darshpratap@gmail.com

Ashutosh Krishna
akrishna286@gmail.com

B. S. Panda
bspanda@maths.iitd.ac.in

[1] Department of Computer Science and Engineering, Delhi Technological University, Main Bawana Road, New Delhi 110042, India

[2] Computer Science and Application Group, Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

tion diffusion model demonstrates that the proposed method delivers better results in comparison to many popular contemporary methods of influence maximization.

Keywords Influence maximization (IM) · Independent cascade (IC) model · Node centrality · Social networks (SNs) · SIR model · Viral marketing

Mathematics Subject Classification 91D30

1 Introduction

In recent years, online social networks (OSNs) have seen tremendous growth in the number of users, which connects these numerous users to perform various online activities. In recent years, OSNs have been the epicenter of numerous research problems like influence maximization, viral marketing, link prediction, community detection, and many more where researchers from different disciplines contribute. These networks provide a great platform to reach a large number of people to promote the idea and products [1,2]. A social network can be interpreted as a graph $G = (V, E)$ where V denote the individuals in this network, while E represents the edge between the users. Here, an edge corresponds to the virtual acquaintance like friendship, follow-followee, co-authorship between the users. The idea of how one individual influences the thoughts of other individuals in the network has gained significant importance. These people having the spreading capability, are often dubbed as influential nodes, and finding such nodes to maximize the information spread in the network is known as influence maximization [3,4]. The problem of influence maximization intends to find k top spreaders from a set of n individuals ($n >> k$) in a system such that information diffusion can be maximized, where k is some constant. The selected influential nodes are often called as seed nodes, which play a significant role in creating an information diffusion cascade to influence many other nodes in the network. Effectively identifying these influential nodes in a network may lead to huge implications in applications such as viral marketing [5]. Viral marketing focuses on targeting these influential nodes for advertisements so as to maximize information dissemination. Many e-commerce companies are utilizing the idea of viral marketing to target many users for successful marketing and promoting their products. The notion behind viral marketing is the word of mouth effect. For instance, people often recommend good restaurants to their friends and family, which increases the probability of them visiting the same restaurant. The complex societal network is highly interdependent, and influence over one individual affects people in their vicinity. Apart from viral marketing, application of influence maximisation extend to rumour control [6], community detection [7], opinion leader detection [8].

The process of influence maximization (IM) usually consists of two phases—identifying influential seed nodes and then the information diffusion phase. Many node centrality based approaches have been introduced to efficiently identify influential nodes like degree centrality [9], closeness centrality [10], eigenvector centrality [11], PageRank [12]. These approaches give importance to each node based on its structural position and its connectivity to other nodes in a network. Most of the centrality based approaches find influential nodes in a biased manner by just taking into account of the strength of each node in the network. Numerous greedy-based algorithms [13,14] are proposed to solve the influence maximization problem using a greedy strategy and discrete optimization technique. Besides, many community-based influence maximization methods [15,16] are proposed with considerably improved performance. Many of these proposed measures usually come with certain limitations like high time complexity and inappropriate selection of spreader nodes.

In this paper, we propose a novel method to identify influential nodes in a complex network, named as modified degree with exclusion ratio (MDER). The proposed algorithm is a type of semi-local approach, which modifies the degree centrality of each node by considering the neighbors up to 2-hops. We argue that by considering the nodes up to 2-hop may cover the neighborhood of a node adequately for the information spreading process. We introduce the idea of exclusion ratio (ER) to ensure that selected spreaders are from diverse localities. The notion of exclusion ratio prevents the inherent biases for the selection of seed nodes belonging to highly dense regions in a network, usually, which is seen in most of the algorithms. Finally, we calculate the MDER centrality of each node by combining the value of the modified degree and exclusion ratio. The proposed method intends to identify seed nodes from a network such that the regions covered by them have a minimum intersection, i.e., in a particular locality of a network once a higher-ranked node is selected, then the chances of selection of the other higher-ranked node in the nearby locality is decreased. We adapt two popular information diffusion models, namely susceptible-infected-recovered (SIR) model [17] and independent cascade (IC) [18], to measure the spread of the information originating from source nodes. Further, we practice various evaluation criteria to judge the performance of the proposed algorithm, along with other existing methods. The results of the experimental results show that the performance of the proposed algorithm exceeds other existing classical methods. Our major contributions are summarised as follows:

1. We propose a novel approach, modified degree with exclusion ratio (MDER), to identify influential nodes in social networks through the concept of modified degree centrality and mutual exclusion.
2. We evaluate the performance of the proposed approach using ten complex real-life networks of varying applications, size, and complexity.
3. The extensive experimental validation using different performance matrices concludes that the proposed MDER algorithm outperforms many popular methods.

The remaining of the paper is organized as follows: In Sect. 2, we review the literature on various centrality measures, followed by a discussion of diffusion models used for measuring information spread. In Sect. 3, the proposed approach, “modified degree using exclusion ratio,” dubbed as MDER, is described in detail with a toy network simulation. In Sect. 4, benchmark datasets and performance metrics used for extensive experimental validation of the proposed approach are discussed. Section 5 summarise the results of the comparison of the proposed approach with existing state-of-the-art techniques. Finally, in Sect. 6, concluding remarks are made.

2 Preliminary

2.1 Influence maximization and viral marketing

Social networks are a type of complex network and are dynamic in nature, where individuals interact with each other and share their interests. This involves a large amount of information flow between the inter-connected entities. A widely studied problem in the field of social network analysis is finding influential nodes in a network to achieve influence maximization. Due to the large and complicated structure of the network, influence maximization is a challenging problem. It is a kind of optimization problem and has mathematically proved that getting an optimal answer is NP-hard using general information spreading models [3]. In social networks, hemophiliac behavior is observed among individuals [19], which means that nodes that are friends or directly connected may exhibit similar behavior and therefore indulge themselves in similar activities and share similar interests. For example, if an individual in a group buys a cellular phone, then the people connected to this individual may be more inclined to buy the same phone as compared to individuals outside this group. This is the central concept of viral marketing: the company targets those individuals who have the maximum reach in the social network and would help them extend their product horizon. Hence, viral marketing thrives on exploiting the power of the influential nodes, which play a significant role in shaping views and opinions of the masses through the word of mouth effect.

In the formulation of influence maximization (IM) algorithm, the main objective is that selected spreaders nodes can propagate the information properly in their neighborhood, leading to a maximum coverage of the information in the network at the end of the diffusion model through cascade triggering. Another crucial factor is that the selection of the spreaders is from diverse regions. Spreaders selected from close to each other may cause interference in the spreading process. Therefore, we aim at maximum coverage and minimum interference or overlapping of spreading caused between selected seed nodes.

2.2 Node centrality

Many node centrality measures have been introduced over the years to rank the nodes in order of their importance in the network. The importance of a node is evaluated based on how a node can spread a piece of information. The centrality measures exploit different topological features of the nodes like the distance between the nodes, shortest path between nodes, the number of connections a node has, or how deeply is a node is lying in the network. The following section describes the various classical centrality measures.

Degree centrality (DC) [9] counts the number of neighbors of a node. There are two types of degree: in-degree and out-degree. The in-degree counts the number of incoming edges while the out-degree counts the number of outgoing edges. The time complexity of computing the following method is $O(m + n)$, where n is the total number of nodes, and m is the number of edges. This is a simple yet effective measure to find influential nodes. One of the major drawbacks of degree centrality is that it deals with the local structure of the network by considering only the direct edges connected to the node. Betweenness centrality [20] incorporates the global structure of the system. It calculates the number of times a node (u) lies between the shortest path of any two nodes. The idea behind this method is that the flow of information must happen from the node that is part of most of the shortest paths. Since this is a global metric, it has a high time complexity of $O(mn + n^2 logn)$. In Eq. 1, B_u is the betweenness centrality value of each node $u \in V$, $\sigma_{ij}(u)$ is the number of shortest paths that contain node u and σ_{ij} is the total number of paths between each pair of nodes i and j.

$$B_u = \sum_{u!=i!=j \in V} \frac{\sigma_{ij}(u)}{\sigma_{ij}} \tag{1}$$

Closeness centrality (CC) [10] determines how close a node u is to other nodes by calculating the distance between u and other nodes. Like betweenness centrality, this method considers the global structure of the system and therefore takes O(n^3) time in producing results. The Eq. 2 given below calculates the closeness centrality, C_u for each node u where $d(u, i)$ is the distance between nodes u and nodes i.

$$C_u = \frac{1}{\sum_{i=1}^{n} d(u, i)} \tag{2}$$

The basic idea behind Eigenvector centrality (EC) [11] is that a node considered important if it is connected to other important nodes in the system. The time complexity is $O(n^2)$. Google mainly uses PageRank [12] approach for ranking of their web pages. Here, each webpage is assigned a score calculated based on the number of links the web page contains and the weights of the links. The web page is analogous to a node, while the links are the edges. The PageRank approach works best with directed graphs. For instance, Degree centrality takes into account the number of neighbours of each node, thereby neglecting global structure. Although betweenness centrality and closeness centrality are based on global measures and provide satisfactory results, computation complexity in these methods is quite high for large networks. The k-shell centrality [21] approach assigns value to nodes based on how deeply the node is lying in the network. The core nodes or the central nodes get high value as compared to the ones on the peripheral of the network. This is done by pruning the nodes recursively until the graph is empty. The time complexity involved in finding k-shell values for all nodes is $O(m+n)$. However, this approach assigns similar core values to many nodes and does not take into account that some of the nodes having the same core value are more influential. Ma et al. [22] proposed gravity centrality inspired by the classical theory of gravity. They adapt the k-shell value of each node as its mass and the shortest path distance between the source node and all other nodes as their distance. The Eq. 3 calculates the gravity centrality(G_u) for node u. Here, ks_u represents the k-shell value of node u, ks_v represents k-shell value of the node v and d_{uv}^2 is the square of the shortest path distance between node u and node v. To compute the gravity centrality of a node u, all nodes up to k-hops from u are considered where k denotes the average degree of the nodes.

$$G_u = \sum_{i=1}^{k} \frac{ks_u * ks_v}{d_{uv}^2} \tag{3}$$

The time complexity of gravity centrality is $O((m+n)k^3)$.

Hirsch index or h-index is a popular matrix to assess the impact of a researcher based on the number of citations received. The concept can be extended in node centrality [23]. The h-index centrality of a node is the maximum h-value such that it has at least h neighbors, and each of these neighbors must have a degree greater or equal to the value of h. Sheikhamadi et al. [24] proposed a DegreeDistance approach which improves in two phases, FIDD and SIDD. They recognised the common drawback of many centrality measures that the selected set of seed nodes have a high number of common neighbours and they overcome this by applying the distance measure to select appropriate nodes. Berahmand et al. [25] proposed a local centrality based spreading method named DCL, which ranks the nodes based on its degree, the degree of its neighbors, common connections between the neighbors of the node, and inverse local cluster coefficient. Authors in [26] introduced a local centrality approach based on the notion of edge ratio and neighborhood diversity measure with a focus on identifying global bridge for the maximum spread of the information. They argued that a node with a high edge ratio and a more diverse neighborhood could contribute significantly to information diffusion. Rui et al. [27] suggested a type of local method of IM named Fixed Neighbor Scale (FNS), which considers neighbors of a node up to multiple levels to determine its spreading influence. They determine the spreading capability of a node by adding multi-level neighbors' weights in the form of their distances from the source node.

2.3 Information diffusion

The process of circulation of information to various users originating from some source nodes in a network is called information dissemination or information diffusion. The modeling of information propagation requires a suitable mathematical formulation [28]. Researchers use the various information diffusion models to evaluate how an influence maximization approach is useful in spreading the information in a real-life scenario. One of the popular categories of information diffusion models is based on epidemic spreading [29]. This is based on how a contagious disease spreads over a network. Initially, few people are infected with the disease, and the rest of them are prone to it (i.e., susceptible). The disease can flow from the infected to the susceptible person. In this manner, a piece of information can spread in a network. The susceptible-infected-recovered (SIR) model is one of the widely used information diffusion model [17]. In this model, nodes are classified into three categories: (i) S-susceptible nodes are the ones that are vulnerable to the disease or information. (ii) I-infected nodes are the ones who know the information, and they can spread it to the nodes that are in its vicinity, and (iii) R-recovered nodes are the ones that were infected and now have been cured. The recovered nodes can not spread the information further.

Other popular categories of information diffusion models are predictive models such as the independent cascade (IC) [18] and the linear threshold (LT) model. In IC model, each inactive node v can be activated by an active adjacent node u with a probability $p(u, v)$. In other words, if at the beginning few of the nodes are active, let's say u is one such node, while the rest are inactive, let's v is one such node, then at time t, u that was activated at time $t - 1$ can activate its neighbor v with probability $p(u, v)$. This probability is called the activation ratio. Once the node is activated, it remains activated. LT model [30], on the other hand, is similar to the IC model but is based on the cumulative effect of the active nodes. Each node decides on a threshold value ranging in [0, 1]. At every step, node v check the weighted sum of the active neighbors. If the weighted sum is greater than the threshold, then v gets activated. Once the node is activated, it remains activated. For the evaluation of influence spread, in this manuscript, we use the SIR and IC model.

3 Proposed algorithm: MDER

In this section, we present our proposed algorithm based on the notion of maximum coverage and minimum interference or overlapping of spreading caused between selected seed nodes. Here, overlapping means if a high-rank node (u) can diffuse information to an area independently, then the other influential node (v) in the same locality may cause overlapping in the diffusion process and therefore is kept relatively lower in the rank list. We introduce the idea of exclusion ratio (ER) in our approach to take care of minimum interference in information dissemination. Therefore, our main objective in case of influence maximization is to choose the seed nodes from the different parts or locality of the network rather than choosing multiple high-rank nodes that are in lying in the vicinity of each other. This is especially useful for an e-commerce company to target influential customers of different regions to promote the product through word-of-mouth strategy and to optimize the overall advertisement. The brief overview of the proposed algorithm and its working is depicted in Fig. 1. The proposed algorithm is outlined as follows:

The brief descriptions of Algorithm 1 are: *Step 1* Firstly, we assign a score to each node (u) by slightly modifying degree centrality. For this we consider i-hop neighbours of node u. These neighbours are used to compute modified degree score of u as follows:

$$MD_i(u) = \frac{\text{No. of } i\text{-hop neighbours}}{\text{total no. of nodes in the network}} \tag{4}$$

Next, we rank the nodes in the network based on the descending order modified degree calculated above and store in list l_i.

Step 2 In this step, we try to ensure mutual exclusion. For doing this, we calculate the exclusion ratio (ER) associated with each node (v). We sequentially pick nodes from the list (l_i) generated in step 1. For each node (v) we introduce two parameters,

Algorithm 1 : MDER

```
Input: $G = (V, E)$, where $n = \|V\|, m = \|E\|$
Output: $\mathcal{R}$
1: for $i \in \{1, 2\}$ do
2:     for $v \in \{1, 2, ..., n\}$ do
3:         $k$ = number of $i$-hop neighbours of node $u$
4:         $MD_i(v) = \frac{k}{n}$
5:     end for
6:     $l_i \leftarrow$ sorted list of $V$ in descending order of $MD_i(u)$
7:     $MDER_i \leftarrow \{\}$
8:     for each vertex $v \in l$ do
9:         $old_v$ = number of $i$-hop neighbours of $v$ that are also $i$-hop neighbours of the nodes higher than
           $v$ in the list $l_i$
10:        $new_v$ = number of $i$-hop neighbours of $v$ that are not included in $old_v$
11:        Calculate the value of $ER(v)$ for each node $v$ using Eq. No. 5.
12:        $MDER_i(v) = MD_i(v) \times ER(v)$
13:    end for
14: end for
15: for $v \in \{1, 2, ..., n\}$ do
16:     $MDER(v) = MDER_1(v) \times MDER_2(v)$
17: end for
18: $\mathcal{R}$ = top $c$ nodes based on the final $MDER$ score.
19: return $\mathcal{R}$
```

namely old and new, where, v_{new} is i-hop neighbors of v that have not been previously covered by any higher-ranked node other than v, while v_{old} counts i-hop neighbors of v that have already been covered by a higher-ranked node present in list l_i. Now, we compute the exclusion ratio (ER) as follows:

The exclusion ratio ensures a minimum overlapping between regions influenced by the nodes. Hence, the notion of exclusion ratio ensures that selected influential nodes lead to maximize information dissemination and prevents the inherent biases for nodes belonging to highly dense regions in a network, which tend to have a higher probability of selection.

Step 3 Now, we compute $MDER_i$ score for each node v by considering its i-hop neighbors. The initially computed centrality value MD in step 1, is multiplied by the exclusion ratio ER for that node, computed in step 2. Hence, the final centrality value of each node v is defined as:

$$MDER_i(v) = MD_i(v) \times ER_i(v) \tag{5}$$

Step number 1, 2 and 3 are going to execute two times for computing MDER score of each node (v), as mentioned in the line no. 1 of the proposed Algorithm 1. First time for $i = 1$ that considers 1-hop neighbors and second time for $i = 2$ that considers 2-hop neighbors.

Step 4 Finally, we combine MDER results computed for 1-hop neighbours as well as 2-hop neighbours i.e $i = 1$ and $i = 2$ to get the final MDER value for each node v by using the mathematical relation:

$$MDER(v) = MDER_{i=1}(v) \times MDER_{i=2}(v) \tag{6}$$

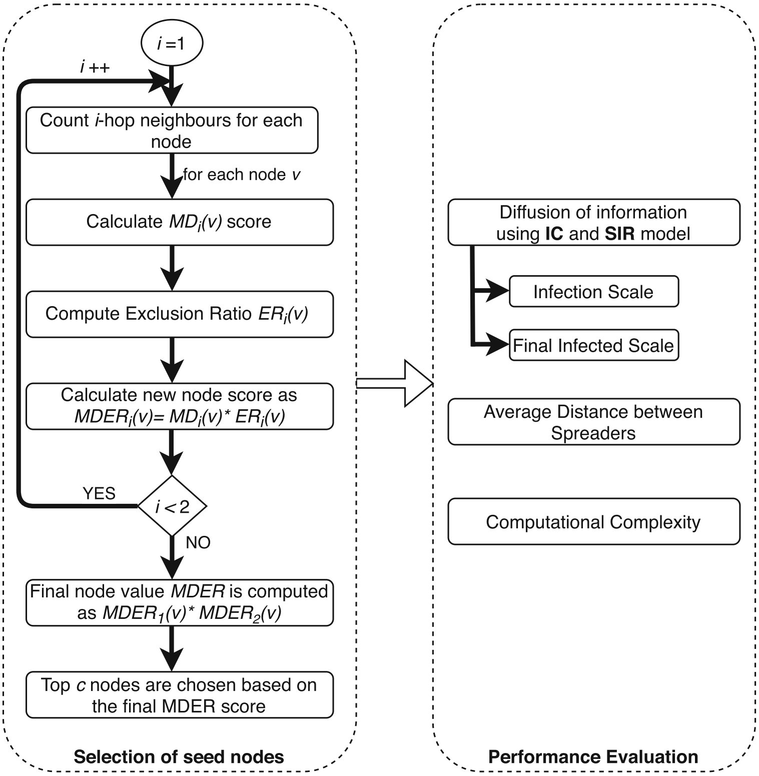
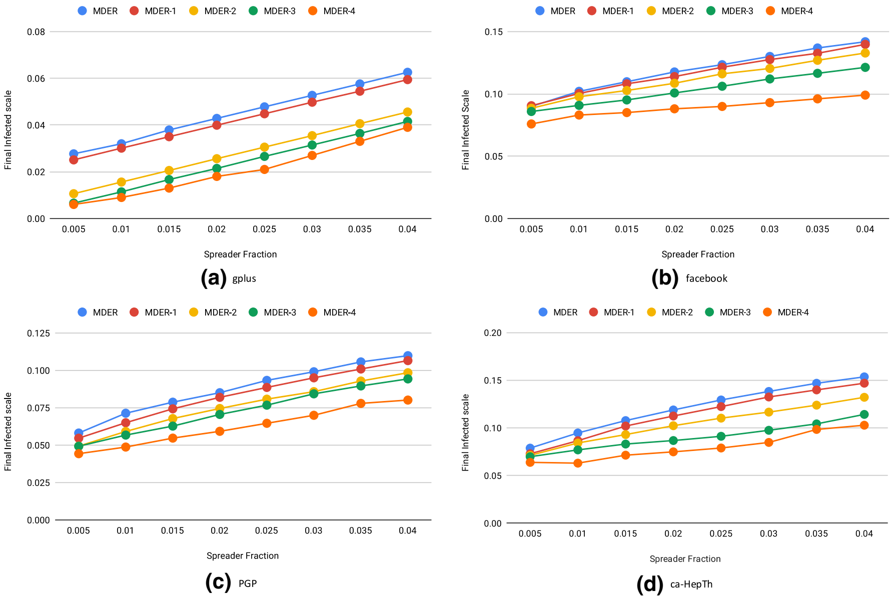

i =1
i ++
Count i-hop neighbours for each node
for each node v
Calculate MD_i(v) score
Compute Exclusion Ratio ER_i(v)
Calculate new node score as MDER_i(v)= MD_i(v)* ER_i(v)
YES
i < 2
NO
Final node value MDER is computed as MDER_1(v)* MDER_2(v)
Top c nodes are chosen based on the final MDER score
Selection of seed nodes
Diffusion of information using IC and SIR model
Infection Scale
Final Infected Scale
Average Distance between Spreaders
Computational Complexity
Performance Evaluation

Fig. 1 Brief overview and working of the proposed algorithm

The values in the $MDER_i$ list are normalized before multiplying so that one of the factors does not overpower the other and does not reduce the effect of each other completely.

We argue that in case of influence maximization, mostly a node can influence its neighbors up to 2-hops. This implies that by considering all the nodes up to 2-hops from a source node u, its neighborhood can be adequately covered in terms of spreading influence. For this, we consider the various combination of $i = 1, 2, 3, 4$-hop neighbors, and practically we achieved the best performance in the influence spread by combining 1-hop and 2-hop neighbors. The comparison of the MDER value for different values of i ($i \in 1, 2, 3, 4$) has been made as shown in the Fig. 2. The results in the form of the final infected scale versus spreaders fractions using the SIR information diffusion model on gplus, facebook, PGP and ca-Hepth. Here, Ca-hepth is a collaboration network dataset, gplus, and facebook belong to the social network dataset, and PGP is a communication network dataset. The obtained results show that the proposed MDER (in blue color), which is the hybrid of the values of $MDER_{i=1}$ and $MDER_{i=2}$ outperforms the others where $MDER-1$, $MDER-2$, $MDER-3$ and $MDER-4$ represents MDER score of each node by considering only its 1-hop, 2-hop, 3-hop, and 4-hop neighbors respectively.



(a) gplus **(b)** facebook

(c) PGP **(d)** ca-HepTh

Fig. 2 Final infected scale versus spraeder fractions using SIR model for **a** gplus, **b** facebook, **c** PGP, and **d** ca-Hepth dataset. The result shows the comparision of results for different values of $MDER_i$ (color figure online)

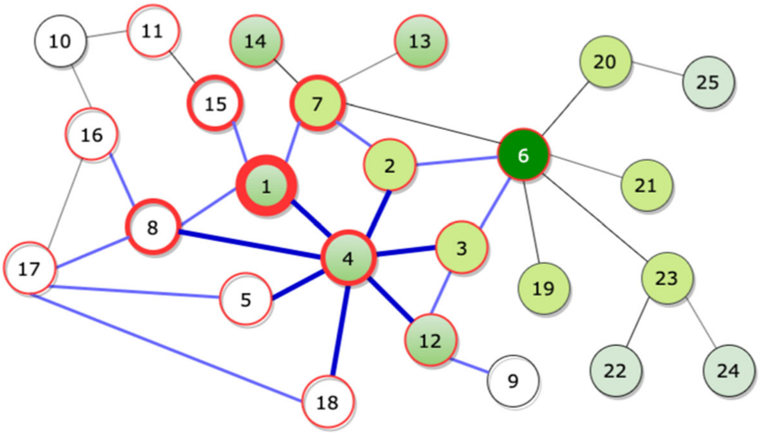


Fig. 3 Toy Network having 25 nodes. The neighbours of node number 6 upto 2nd level is represented in green color. The neighbours of node number 1 upto 2^{nd} level is outlined with red color and the neighbours of node number 4 upto 2nd level is shown having edges of the color blue (color figure online)

3.1 Explanation on how our algorithm achieves mutual exclusion through a toy network

The proposed algorithm is explained with the help of a toy network consisting of 25 nodes in Fig. 3. The node numbered 6 (marked in green color) has got the highest

MDER score, followed by node numbered 1 (marked in red color). The neighbors of these nodes are marked with a lighter version of their respective colors.

The idea of exclusion ratio is illustrated in the toy network is given as follows: Node number 6 ranks the highest in list, this means that the value of parameter old for node 6 is zero, and the value of parameter new for node 6 is 15, which is total number of neighbours up to 2-hops. Hence, the value of $E(6) = 15$. Second in the rank list is node numbered 1. For node 1, the value of parameter old is the number of common 2-hop neighbours of nodes 6 and 1, since node 6 is higher in the list. The list of common neighbours up to 2-hop of both nodes includes node number 2, 3, 4, 6, 7, 12, 13, and 14. Therefore, the value of parameter old for node 1 is 8 and the value of parameter new for node 1 is the number of its remaining neighbours up to 2-hops, which is equal to 7 and includes node numbered 5, 8, 11, 15, 16, 17, and 18. Hence the value of exclusion ratio for node 1, i.e. $E(1)= \frac{7}{8}$.

According to degree centrality, node 4 and node 6 should be in the same position in the rank list since the number of neighbors for them is the same, i.e., 7. But, the MDER method ranks node number 4 lower in the rank list. According to our method, nodes in the neighborhood of nodes 6 and 1 completely cover the nodes covered by 4. This means that the exclusion ratio, for node 4, is very low since it fails to cover multiple new nodes. Therefore, node 4 is not given a higher position in the rank list. This ensures that there is minimal overlap between the regions covered by higher-ranked nodes hence maximizing the diffusion of information.

3.2 Computational complexity

The time complexity of the proposed approach is determined from the pseudo-code presented in Algorithm 1. The outer for loop in line number 1 runs only two times, as we consider neighbors up to 2-hops. In lines from 2 to 5, where the list $MD(v)$ is calculated, the time taken is $O(n + m)$. In line number 6, to rearrange the list in descending order of values can be obtained using a general sorting algorithm like heap sort or quick sort in time $O(nlog(n))$. To calculate the exclusion ratio (ER) and MDER score for each node (v) in lines 8 to 13, requires $O(m + n)$ time, as for each node, we have to look up its neighbors up to 2-hops. In line number 18, we need to find top c nodes, as in influence maximization (IM) objective is to find constant seed nodes, where $c << n$. In the worst case, it can be done in $O(c * n)$ time. Therefore, the final time complexity of the proposed algorithm is $O(n + m) + O(nlog(n)) + O(n * c) = O(nlog(n))$ in the worst case. Here, we considered the time involved in sorting activity in line number 6 as $O(nlog(n))$. But in the actual scenario, we can ignore those nodes in the sorting activity whose number of neighbors up to 2-hops are very less, as in case of the real-life networks preferential attachment and power-law is observed in the degree distribution, which implies many nodes in the network has very few degrees and very few nodes have relatively high degree [31]. By doing some prepossessing on the dataset, we can discard those nodes in the ranking whose degree is relatively very less. Thus the actual running time of MDER is going to be far less than $O(nlogn)$, which is further evident from the results obtained in Sect. 5.4.

4 Dataset and performance metrics

In order to perform the detailed performance analysis of the proposed algorithm, along with popular existing algorithms of IM, we use ten real-life datasets of various sizes and nature and adopt four different evaluation criteria. The datasets and evaluation metrics used are presented below. Most of the practiced real-life datasets are referred from SNAP Stanford [32], which is a publicly available repository maintained by the researchers of Stanford.

4.1 Dataset

In this paper, we utilize a total of ten datasets of varying application, size, nature. Out of that, there are six social network datasets, namely twitter, gplus, twitter-ego, Brightkite, musae-facebook, and musae-twitch. There are three collaboration networks: Enron-email, ca-Condmat, and ca-Hepth and one communication network, Pretty Good Privacy (PGP). Table 1 lists the statistical properties of each dataset where d_{avg} and d_{max} represent the average degree and maximum degree of nodes, respectively. The β_{th} denotes the rate of infection for the information spread. We need to suitably choose the rate of infection (β_{th}) in SIR and activation probability (p_{uv}) in the IC model, which decides the rate at which an infected or active node infects or influences its neighbors. In the literature, most of the IM algorithms have considered the rate of infection as an epidemic threshold which is calculated as

$$\beta_{th} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} (d_i)^2}, \tag{7}$$

where d_i denote the degree of node v_i. Hence, the value β_{th} for a dataset is equal to the average degree of the nodes divided by the square of the average degree of the nodes.

Enron-email consists of communication of half a million emails among the employees of the Enron corporation [33]. If an email is sent from an employee to another, an undirected link is made between the two. The CA-Condmat dataset consists of nodes as authors, and an edge between the nodes exists if a physics author has co-authored with another author on a paper [34]. Musae-facebook network contains page-page networks of verified Facebook sites. Nodes represent official Facebook pages, and the links are the mutual likes between them [35]. PGP, i.e., Pretty Good Privacy, is an encrypted communication network [36]. Ca-Hepth is a collaboration network of authors who collaborated on papers that were submitted to the High Energy Physics—Theory category [34]. Musae-twitch is a social network dataset, where the nodes are the twitch users, and the link is between them if they are friends online [35]. Brightkite is a location based social network that logs your location and shares it with your connections. This dataset is obtained from Brightkite social network containing friendship networks [37]. Gplus is a social network dataset consists of ’circles’ from Google+. The data was obtained from users who had manually shared their circles using the ’share circle’ feature in Google+ [38]. Twitter is a widely used social networking ser-

Table 1 Brief statistical descriptions of the different datasets

S.no.	Dataset	No. of nodes (n)	No. of edges (m)	d_{max}	d_{avg}	β_{th}
1	Twitter	81,306	1,342,296	3383	33.018	0.00579
2	Brightkite	58,228	214,078	1134	7.353	0.01569
3	Enron-email	36,697	183,835	1383	10.020	0.00713
4	Gplus	23,628	39,194	2761	3.317	0.00265
5	Twitter-ego	23,370	32,831	238	2.809	0.02597
6	Ca-Condmat	23,133	93,439	279	8.078	0.04533
7	Musae-facebook	22,470	171,002	709	15.204	0.0163
8	PGP	10,638	24,301	205	4.568	0.05299
9	Ca-hepth	9882	25,977	65	5.257	0.07984
10	Musae-twitch	7126	35,324	720	9.914	0.01678

vice where individuals interact by sharing short messages or "tweets". A link is formed if they reply, retweet or follow another individual. The Twitter dataset consists of user lists crawled from Twitter. The dataset includes node features, circles, and ego networks [38]. Twitter-ego is a directed networks containing Twitter user-user following information. It is an ego network that has a focal point called an ego and each ego is connected to its social links called as alters. This data is publicly available at http://konect.cc/networks/ego-twitter/.

4.2 Performance matrices

(i) *Infection scale* The infection scale is the measure that indicates the scale at which nodes get infected or become active and then recovered or inactive due to infection originated from selected seed nodes with respect to time. The calculation of the number of nodes that are affected by the information in the spreading process is determined using an information diffusion model. In the SIR model, we select initial seed nodes that start the diffusion of information in the network. These seed nodes pass the information to their adjacent nodes with an infection probability of β and then recover with a probability of δ. The recovered nodes do not further act as seed nodes and therefore stop infecting their neighbors. The Infection scale is the summation of infected nodes and recovered nodes at a particular timestamp. The infection scale initially rises significantly since the number of recovered nodes is very less, and later the process slows down due to the high number of recovered nodes.

(ii) *Influence spread of total number of infected nodes* We calculate the final infection scale, which counts the total number of infected nodes after the end of the information dissemination process originating from selected influential nodes. As discussed SIR model passes information with the rate β and recovers with a rate of δ, whereas in the case of the IC model, the nodes that seed nodes are initially activated with information disperse information to its neighbors and do not recover and become inactive, instead they continue to diffuse information. The rate at which the node is activated is determined using activation probability p_{uv}. Active nodes or the final infected scale includes the total number of nodes that received information at the end of the spreading process.

(iii) *Average distance between spreaders* This metric calculates the average shortest path distance between the selected seed nodes. A large value of the average shortest distance between spreaders implies that the seed nodes are located diversely and not in the vicinity of each other. The selection from diverse regions means that information dissemination is maximized. Assume shortest path distance between two spreader nodes v_i and v_j is d_{ij} and number of chosen spreaders are c. Then the following equation provides the average distance between spreaders (Ls):

$$Ls = \frac{\sum d_{ij}}{\frac{c(c-1)}{2}} \tag{8}$$

where $\sum d_{ij}$ denotes the sum of shortest path between each pair of seed nodes.

(iv) *Time* We also assess the absolute time required to produce the ranking of nodes in terms of their spreading capability by the proposed model and other methods in comparison. The time complexity to produce a ranking list of nodes in the case of degree centrality, k-shell, and h-index is $O(m+n)$. In the case of PageRank and eigenvector centrality, it is equal to $O(n^2)$, and for gravity centrality time complexity is $O((m+n)k^3)$, where m, n, and k represent the number of edges, number of nodes, and the average degree of the nodes in the network respectively. However, in the worst case and without any data prepossessing, the time complexity of the proposed algorithm is $O(nlogn)$. The actual running time on real-life networks can give a better perspective on the effectiveness of the algorithm.

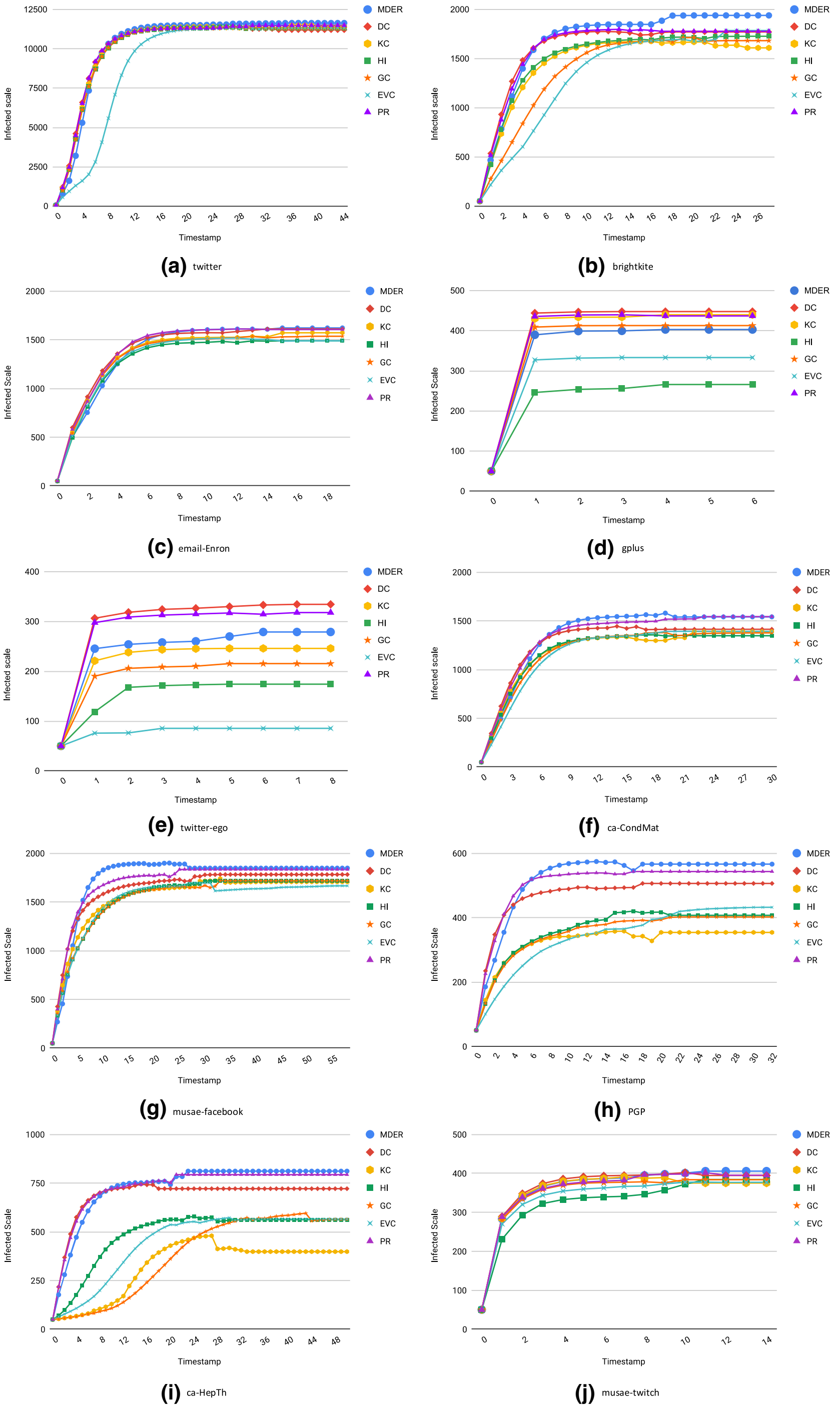
5 Experimental results and analysis

In this section, we present the exclusive experimental results and analysis to judge the performance of the proposed approach. The performance matrices described in Sect. 4.2 are used to obtain comparisons between the proposed algorithm and several baseline measures for influence maximization. These measures include degree centrality (DC) [9], h-index (HI) [23], Eigenvector (EVC) [11], Pagerank (PR) [12], k-shell (KC) [21], and gravity centrality (GC) [22]. To capture the utility of the introduced algorithm, MDER, simulations are performed on ten diverse real-life networks, as depicted in Table 1. To spread the information originating from seed nodes, the value of infection rate or activation probability for each dataset is taken as β_{th}, which is mentioned in Table 1.

The selection of the node to be infected is chosen randomly from the set of neighbors of the seed node. For example, in the case of the PGP dataset, $\beta_{th} = 0.05299$, which implies an active or infected node at time t can randomly infect 5.29% of its neighbors at time $t+1$. We performed the execution to obtain the results on a personal computer with a configuration of 1.8 GHz Intel Core i5 processor and 8 GB of primary memory.

5.1 Infection scale

Figure 4 depicts the result of the infection scale at increasing timestamps of the proposed algorithm, along with other existing methods. Infection scale at any time t is the sum of infected and recovered nodes. For the experiment, we initially activate 100 top spreaders as selected by each algorithm to commence the spreading process. We utilize the SIR model for the information spreading, and the results are averaged over 100 independent simulations. For each dataset, the value of the infection rate is taken as β_{th}. The number of infected nodes is initially increasing, and after a while, when no further infected nodes are present, then the infected scale becomes constant. In the case of ca-CondMat network, we observe that till timestamp 22, the infection is increasing, and after that, the number of infected nodes became constant with a value of 1540. In contrast, other approaches, not only reach recovery state early but also infect less number of nodes. For example, degree centrality infects a total of 1413 nodes and reaches recovery state at timestamp 18. From the results, it is clear that



(a) twitter

(b) brightkite

(c) email-Enron

(d) gplus

(e) twitter-ego

(f) ca-CondMat

(g) musae-facebook

(h) PGP

(i) ca-HepTh

(j) musae-twitch

Fig. 4 Infection scale at different timestamps for each dataset by taking top-100 influential nodes from each method. The results are obtained over 100 independent SIR simulations with rate of infection as β_{th}

the proposed MDER algorithm performs better than most of the other approaches. However, in the case of email-Enron and musae-twitch networks, the performance of MDER and Pagerank are very close to each other.

5.2 Final infected scale or active nodes

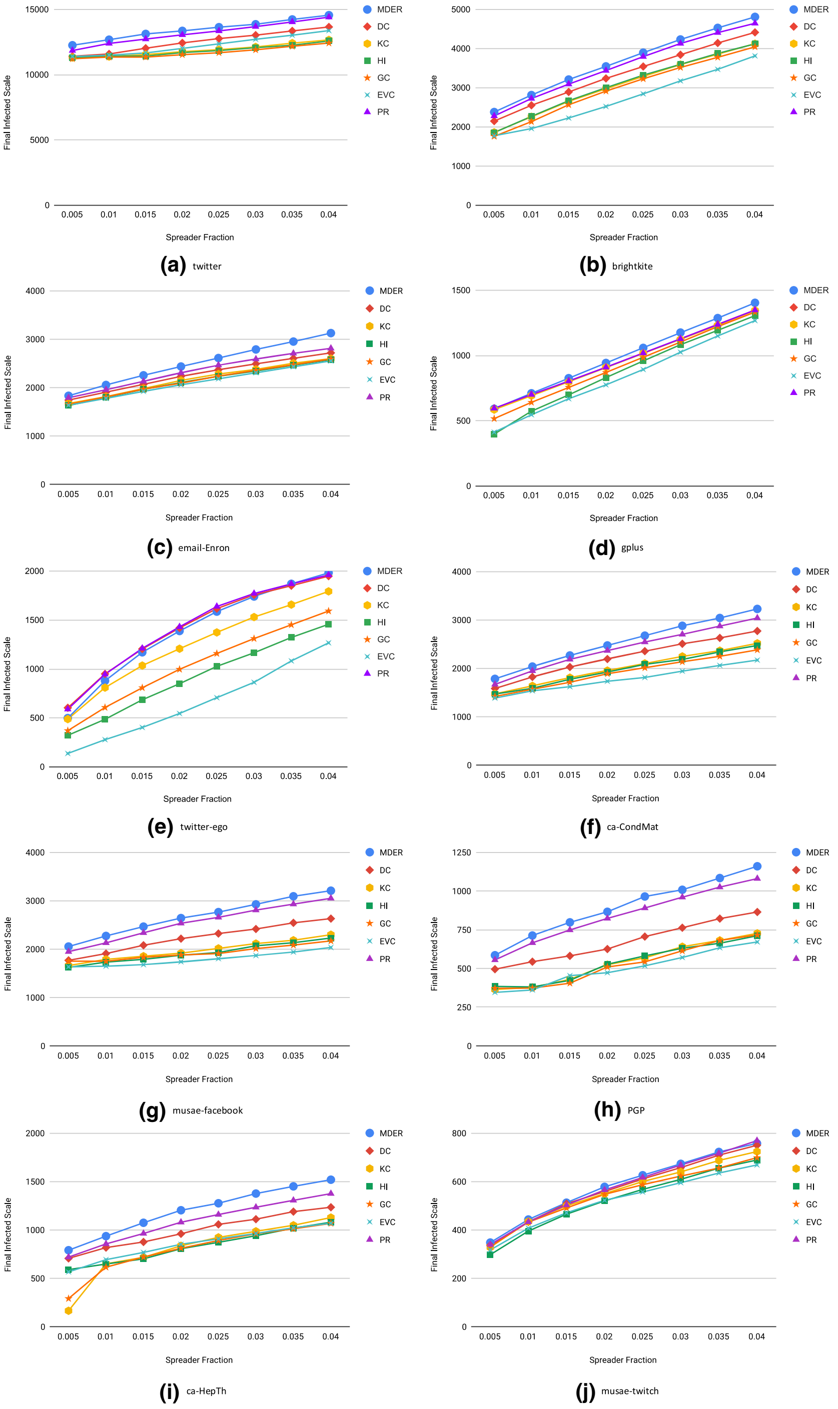
The final infected scale or active nodes is one of the primary criteria to asses the influence spread produced by triggering from chosen influential nodes. It counts the total number of nodes who got infected or influenced after the end of the spreading process. We utilize the SIR and IC model as the information diffusion model to find the final infected scale with respect to spreader fractions, where spreaders are chosen by MDER and other methods in comparison. To obtain the results for each dataset considered, we keep the spreader fractions in the range of 0.005, 0.01, 0.015, 0.020, 0.025, 0.03, 0.035, and 0.040, and the value of the final infected scale is computed against each fraction of spreader. For example, in the case of the PGP dataset containing 10, 638 nodes, an absolute number of influential spreaders are taken in the range of 53, 106, 159, 212, 265, 319, 372, and 425 to plot the final infected scale.

Figure 5 shows the results of the final infected scale or active nodes versus fraction of influential nodes chosen by MDER and other methods in comparison using the SIR information diffusion model. The obtained results are averaged over 100 simulations of the SIR model with infection rate as β_{th}, which is mentioned in Table 1. In Fig. 5f, for the ca-Condmat network, there are a total of 23, 133 nodes out of which we select 0.005 percent of nodes (which is approximately 115 nodes). We observe that the final infection scale is 1781 nodes. From the results in Fig. 5, it is clear that MDER produces consistent and best results as compared to all other methods in comparison for each dataset. Similarly, Fig. 6 shows the results of total active nodes produced at the end of the spreading process due to influential seed nodes under the IC model. The results are sampled over 100 independent simulations of the IC model. In Fig. 6f for the ca-Condmat dataset, we note that when a fraction of 0.005 nodes are activated initially, then there are a total of 8602 active nodes at the end of the process. From the results in Fig. 6, it is evident that the proposed algorithm beats all other methods comfortably in each dataset.

The performance obtained by MDER using both diffusion models, namely SIR and IC, exceed the performance of other popular measures, including k-shell, PageRank, and gravity centrality. Hence, the experimental results achieved in this matrix echoes the utility and superiority of the quality of seed selection by the proposed algorithm, MDER.

5.3 Average distance between spreaders

To achieve maximum coverage of information with minimum interference cause in spreading, the average shortest distance between the spreaders (L_s) is an important matrix. The value of L_s is computed using Eq. 8 to demonstrate that the selection of nodes is made from diverse regions. The results of this experiment are shown in Fig. 7.



(a) twitter

(b) brightkite

(c) email-Enron

(d) gplus

(e) twitter-ego

(f) ca-CondMat

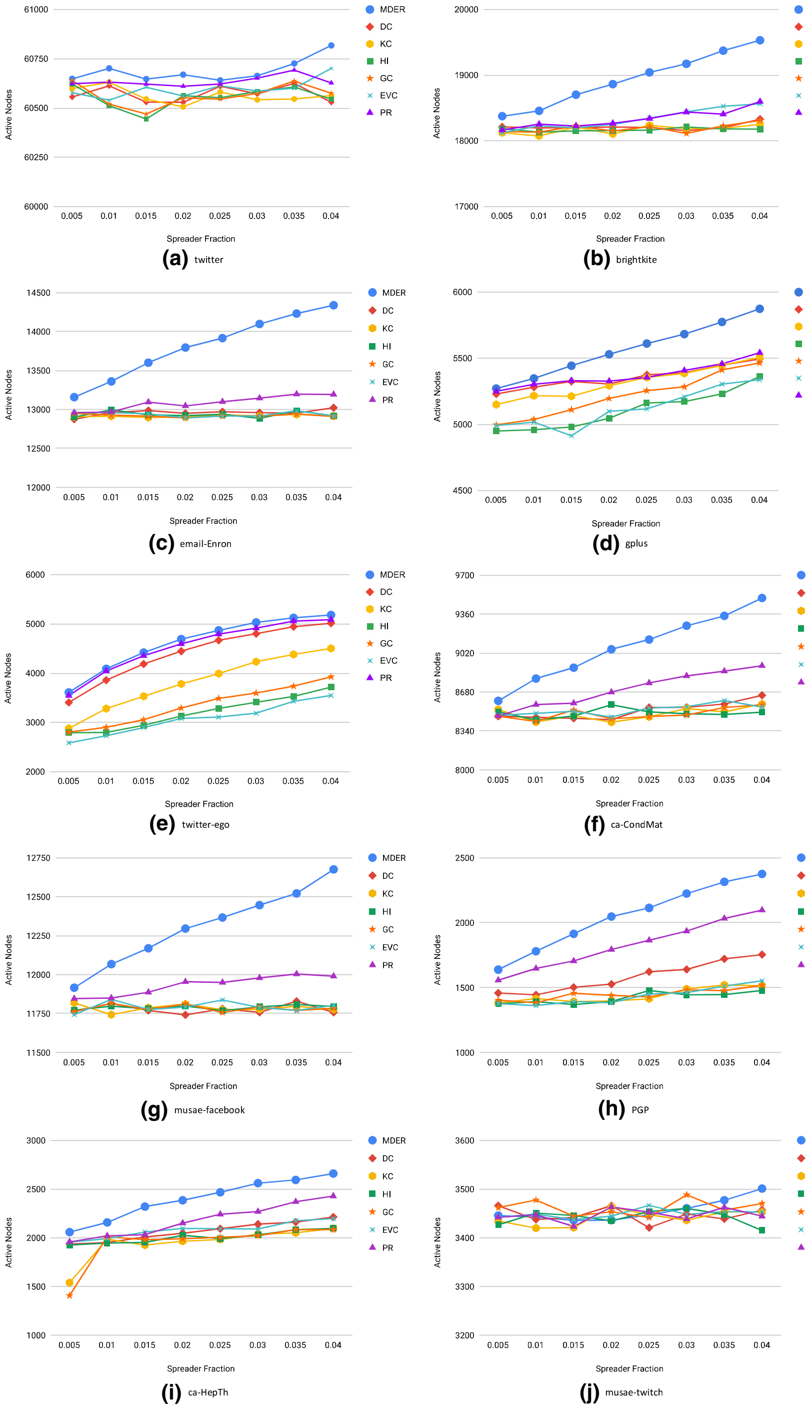
(g) musae-facebook

(h) PGP

(i) ca-HepTh

(j) musae-twitch

Fig. 5 Final infected scale with respect to spreader fractions using SIR model on ten different real-life networks: **a** twitter, **b** brightkite, **c** email-Enron, **d** gplus, **e** twitter-ego, **f** ca-CondMat, **g** musae-facebook, **h** PGP, **i** ca-HepTh, and **j** musae-twitch. The results are obtained using 100 independent simulations



(a) twitter

(b) brightkite

(c) email-Enron

(d) gplus

(e) twitter-ego

(f) ca-CondMat

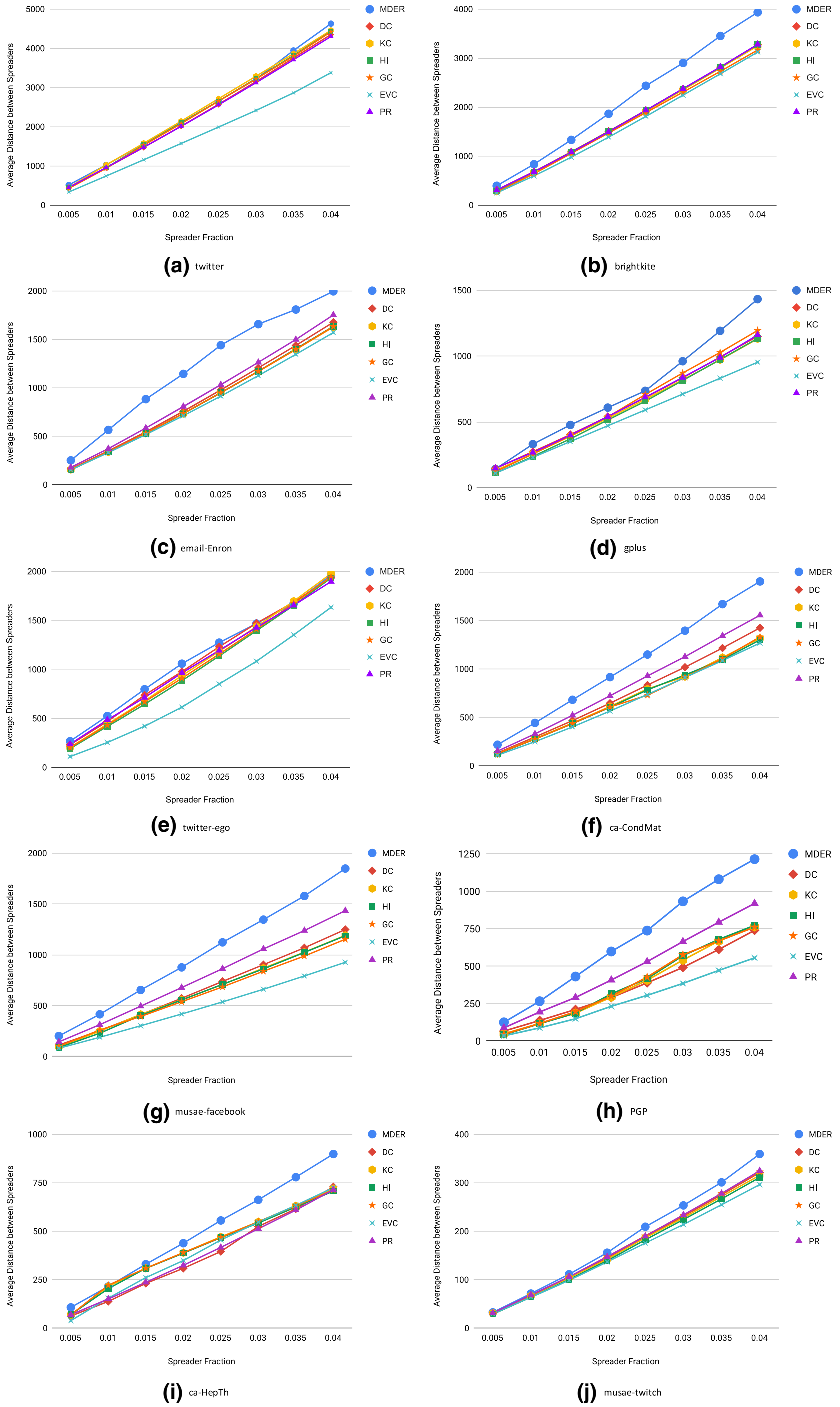
(g) musae-facebook

(h) PGP

(i) ca-HepTh

(j) musae-twitch

Fig. 6 Number of active nodes with respect to spreader fraction using IC model on six different real networks—**a** twitter, **b** brightkite, **c** email-Enron, **d** gplus, **e** twitter-ego, **f** ca-CondMat, **g** musae-facebook, **h** PGP, **i** ca-HepTh, and **j** musae-twitch. The results are sampled over 100 independent simulations



(a) twitter

(b) brightkite

(c) email-Enron

(d) gplus

(e) twitter-ego

(f) ca-CondMat

(g) musae-facebook

(h) PGP

(i) ca-HepTh

(j) musae-twitch

Fig. 7 **a–j** The average shortest path distance between the selected spreaders (L_s) versus spreader fraction for each dataset

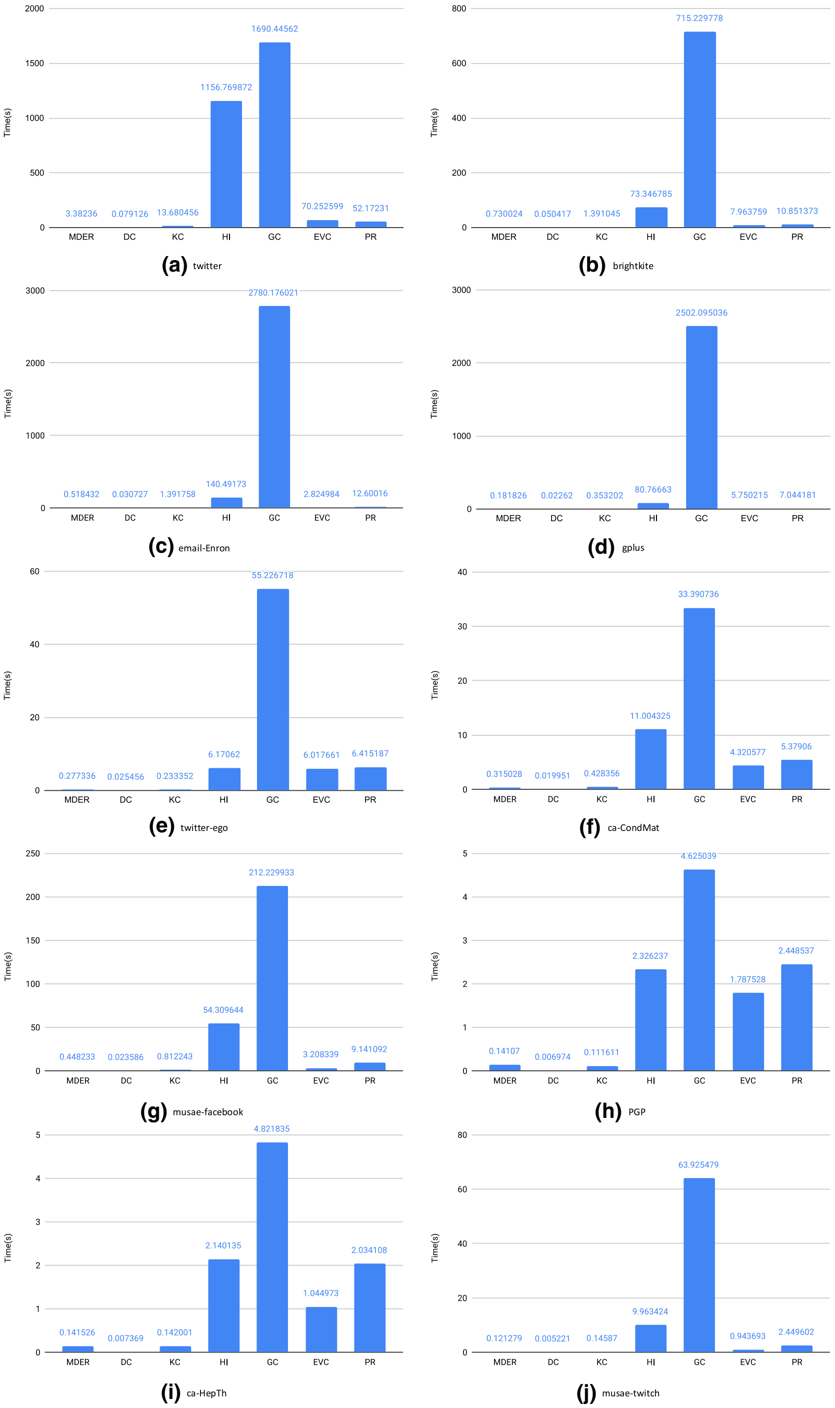


Fig. 8 **a–j** The actual running time to generate the ranking list of nodes under different methods for **a** twitter, **b** brightkite, **c** email-Enron, **d** gplus, **e** twitter-ego, **f** ca-CondMat, **g** musae-facebook, **h** PGP, **i** ca-HepTh, and **j** musae-twitch datasets

The graphs show that the distance between the spreaders chosen through the MDER approach is consistently higher than the spreaders elected by other approaches in every dataset. For instance, In the email-Enron network, when 0.04 percent spreaders are selected, then the value of L_s for MDER is 1993, whereas others lie in the range of 1568 and 1749. From the obtained results, it is evident that the proposed algorithm outperforms all other methods.

5.4 Time

In Fig. 8, we compare the actual running time required to rank the nodes based on the spreading capability using the proposed algorithm, MDER, and all other algorithms in the comparison. We performed this examination on ten data sets, namely twitter, brightkite, email-Enron, gplus, twitter-ego, ca-Condmat, musae-facebook, PGP, and ca-HepTh and musae-twitch dataset. We observed that the time taken by MDER for the generation of the ranked list is close to that of degree and k-shell centrality. To process 23, 133 nodes in ca-CondMat, MDER only takes only 0.31 s, which is significantly faster than h-index, gravity centrality, Eigenvector centrality, and PageRank.

6 Conclusion

Finding influential speeders is of utmost importance to achieve optimal advertisement of the products and ideas in a complex network like social networks. In this paper, we introduced an influence maximization algorithm named Modified Degree with Exclusion Ratio (MDER). The method aims to select nodes that have the highest connections based on modified degree centrality up to 2-hops, and the regions covered by the chosen nodes do not overlap with each other. The introduction of mutual exclusion in the proposed approach prevents the inherent biases for the selection of many nodes belonging to highly dense regions in a network. It ensures that influential nodes are selected from diverse regions, which eventually leads to a maximum coverage of the network. The simulation of the proposed algorithm using classical information diffusion models on ten datasets of various sizes and applications echoes the efficiency and utility of the algorithm in real-life scenarios. The method outperforms the widely used centrality measures, including k-shell, PageRank, and gravity centrality.

References

1. Heidemann J, Klier M, Probst F (2012) Online social networks: a survey of a global phenomenon. Comput Netw 56(18):3866–3878
2. Krasnova H, Spiekermann S, Koroleva K, Hildebrand T (2010) Online social networks: why we disclose. J Inf Technol 25(2):109–125
3. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 137–146
4. Li Y, Fan J, Wang Y, Tan KL (2018) Influence maximization on social graphs: a survey. IEEE Trans Knowl Data Eng 30(10):1852–72

5. Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 57–66
6. Vega-Oliveros DA, da Fontoura CL, Rodrigues FA (2020) Influence maximization by rumor spreading on correlated networks through community identification. Commun Nonlinear Sci Numer Simul 83:105094
7. Kumar S, Panda BS, Aggarwal D (2020) Community detection in complex networks using network embedding and gravitational search algorithm. J Intell Inf Syst. https://doi.org/10.1007/s10844-020-00625-6
8. Oueslati W, Arrami S, Dhouioui Z, Massaabi M (2021) Opinion leaders' detection in dynamic social networks. Concurr Comput Pract Exp 33(1):e5692. https://doi.org/10.1002/cpe.5692
9. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–39
10. Okamoto K, Chen W, Li XY (2008) Ranking of closeness centrality for large-scale social networks. In: International workshop on frontiers in algorithmics. Springer, Berlin, Heidelberg, pp 186–195
11. Bonacich P (2007) Some unique properties of eigenvector centrality. Soc Netw 29(4):555–64
12. Brin S, Page L (2012) Reprint of: The anatomy of a large-scale hypertextual web search engine. Comput Netw 56(18):3825–33
13. Cheng S, Shen H, Huang J, Zhang G, Cheng X (2013) Staticgreedy: solving the scalability–accuracy dilemma in influence maximization. In: Proceedings of the 22nd ACM international conference on information & knowledge management, pp 509–518
14. Goyal A, Lu W, Lakshmanan LV (2011) Celf++ optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th international conference companion on World wide web, pp 47–48
15. Huang H, Shen H, Meng Z (2020) Community-based influence maximization in attributed networks. Appl Intell 50(2):354–64
16. Kumar S, Singhla L, Jindal K, Grover K, Panda BS (2021) IM-ELPR: Influence maximization in social networks using label propagation based community structure. Appl Intell. https://doi.org/10.1007/s10489-021-02266-w
17. Satsuma J, Willox R, Ramani A, Grammaticos B, Carstea AS (2004) Extending the SIR epidemic model. Phys A Stat Mech Appl 336(3–4):369–75
18. Goldenberg J, Libai B, Muller E (2001) Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata. Acad Mark Sci Rev 9(3):1–8
19. Murase Y, Jo HH, Török J, Kertész J, Kaski K (2019) Structural transition in social networks: the role of homophily. Sci Rep 9(1):1–8
20. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 35–41. https://doi.org/10.2307/3033543
21. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888–93
22. Ma LL, Ma C, Zhang HF, Wang BH (2016) Identifying influential spreaders in complex networks based on gravity formula. Phys A Stat Mech Appl 451:205–12
23. Lü L, Zhou T, Zhang QM, Stanley HE (2016) The H-index of a network node and its relation to degree and coreness. Nat Commun 7:10168
24. Sheikhahmadi A, Nematbakhsh MA, Shokrollahi A (2015) Improving detection of influential nodes in complex networks. Phys A Stat Mech Appl 436:833–845
25. Berahmand K, Bouyer A, Samadi N (2019) A new local and multidimensional ranking measure to detect spreaders in social networks. Computing 101(11):1711–33
26. Samadi N, Bouyer A (2019) Identifying influential spreaders based on edge ratio and neighborhood diversity measures in complex networks. Computing 101(8):1147–75
27. Rui X, Yang X, Fan J, Wang Z (2020) A neighbour scale fixed approach for influence maximization in social networks. Computing. 102(2):427–449. https://doi.org/10.1007/s00607-019-00778-5
28. Kumar S, Saini M, Goel M, Panda BS (2021) Modeling information diffusion in online social networks using a modified forest-fire model. J Intell Inf Syst 56(2):355–377
29. Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42:599–653
30. Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci 99(9):5766–71

31. Yamasaki K, Matia K, Buldyrev SV, Fu D, Pammolli F, Riccaboni M, Stanley HE (2006) Preferential attachment and growth dynamics in complex systems. Phys Rev E 74(3):035103
32. Leskovec J, Krevl A, SNAP Datasets (2014) Stanford large network dataset collection, vol 2016, p 49. http://snap.stanford.edu/data
33. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123
34. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov Data (TKDD) 1(1):2-es
35. Rozemberczki B, Allen C, Sarkar R (2019) Multi-scale attributed node embedding. arXiv preprint arXiv:1909.13021
36. Boguná M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. Phys Rev E 70(5):056122
37. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: friendship and mobility: user movement in location-based social networks. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD)
38. McAuley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. In: NIPS, vol 2012, pp 548–556