Expert Systems With Applications 179 (2021) 115061

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Identifying and ranking super spreaders in real world complex networks without influence overlap

Giridhar Maji [a,*,1], Animesh Dutta [b,2], Mariana Curado Malta [c,3], Soumya Sen [d,4]

[a] Department of Electrical Engineering, Asansol Polytechnic, Asansol, India
[b] Department of Computer Science and Engineering, National Institute of Technology, Durgapur, WB, India
[c] CEOS.PP. Polytechnic of Porto, Portugal & Algoritmi Center, University of Minho, Portugal
[d] A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India

ARTICLE INFO

Keywords:
Influential spreader identification
Spreading overlap
Seed selection with minimum geodesic
SIR simulation
Monotonicity
Kendall's rank correlation

ABSTRACT

In the present-days complex networks modeled on real-world data contain millions of nodes and billions of links. Identifying super spreaders in such an extensive network is a challenging task. Super spreaders are the most important or influential nodes in the network that play the central role during an infection spreading or information diffusion process. Depending on the application, either the most influential node needs to be identified, or a set of initial seed nodes are identified that can maximize the collective influence or the total spread in the network. Many centrality measures have been proposed to rank nodes in a complex network such as 'degree', 'closeness', 'betweenness', 'coreness' or 'k-shell' centrality, among others. All have some kind of inherent limitations. Mixed degree decomposition or m-shell is an improvement over k-shell that yields better ranking. Many researchers have employed single node identification heuristics to select multiple seed nodes by considering top-k nodes from the ranked list. This approach does not results in the optimal seed nodeset due to the considerable overlap in total spreading influence. Influence overlap occurs when multiple nodes from the seed nodeset influence a specific node, and it is counted multiple times during total collective influence computation. In this paper, we exploit the 'node degree', 'closeness' and 'coreness' among the nodes and propose novel heuristic template to rank the super spreaders in a network. We employ k-shell and m-shell as a coreness measure in two variants for a comparative evaluation. We use a geodesic-based constraint (enforcing a minimum distance between seed nodes) to select an initial seed nodeset from that ranked nodes for influence maximization instead of selecting the top-k nodes naively. All models and metrics are updated to avoid overlapping influence during total spread computation. Experimental simulation with the SIR (Susceptible-Infectious-Recovered) spreading model and an evaluation with performance metrics like spreadability, monotonicity of ranking, Kendall's rank correlation on some benchmark real-world networks establish the superiority of the proposed methods and the improved seed node selection technique.

1. Introduction

A social being is 'social' due to its different kinds of relations within society. Such interactive relations among the participants/inhabitants of the society create the so-called 'social networks'. 'viral marketing', 'rumor spreading', 'word of mouth' (Ferguson, 2008) are well-known terms associated with a social network. A social network is a network of different social beings (nodes) connected through many relations (edges). Such social networks allow the formation of smaller subgroups of 'like-minded' people, known as communities (Girvan & Newman, 2002; Gleiser & Danon, 2003) based on their interests, preferences, and many other parameters. A person who belongs to a particular

* Corresponding author.
E-mail addresses: Giridhar.Maji@gmail.com (G. Maji), animesh@cse.nitdgp.ac.in (A. Dutta), mariana@iscap.ipp.pt (M. Curado Malta), iamsoumyasen@gmail.com (S. Sen).
[1] ORCID ID: https://orcid.org/0000-0003-4751-3471
[2] ORCID ID: https://orcid.org/0000-0003-4880-6903
[3] ORCID ID: https://orcid.org/0000-0002-3512-931X
[4] ORCID ID: https://orcid.org/0000-0002-9178-6410

https://doi.org/10.1016/j.eswa.2021.115061
Received 16 February 2021; Received in revised form 6 April 2021; Accepted 15 April 2021
Available online 21 April 2021

community is more densely connected to his/her own community members than to the rest of the network. Persons that overlap in different communities, i.e., those belonging to more than one community, are considered important due to their control and influence over the propagation of information between those communities (Tang et al., 2019). In earlier days, such networks have been used for many different purposes e.g. campaigning for a presidential candidate, on a market survey of a new product, purposefully spreading specific propaganda (Rehman, Jiang, Rehman, Paul, & Sadiq, 2020), among others. Information about important nodes (persons) is often required to contain or immunize an outbreak of epidemic infection, or misinformation/malign rumors (Ghoshal, Das, & Das, 2019; Kumar, Verma, & Singh, 2018; Yang, Li, & Giua, 2020). In the present COVID-19 pandemic situation, authorities need to identify the persons (nodes) that are likely to accelerate the infection spread and isolate them to contain the spread (Kabir & Tanimoto, 2020; Arefin, Masaki, Kabir, & Tanimoto, 2019) of the disease. In a similar work COVID-19 contact tracing data has been utilized with a statistical social network model to characterize the disease transmission (Nagarajan, Muniyandi, Palani, & Sellappan, 2020).

In recent times due to the pervasive and omnipresent Internet, everyone has become a global citizen without geographic boundaries (Rath, 2019). We are now part of 'online social networks' (OSN). OSN allow a person to connect to any other person worldwide. The problems of rumor spreading, new product launching, event campaigning, news spreading, among others, have become more challenging in OSN due to its enormous size and overwhelming interconnections (Franchi, Poggi, & Tomaiuolo, 2020). Not only the OSN, but many other seemingly unrelated problems can be mapped to the generic problem of identifying the most important/central/critical entities and their importance in information spreading when they are modeled as a network – e.g. (1) an authors' citation network that has the goal of understanding which are the influential researchers (Ley, 2002); (2) a collaboration network of researchers on scientific projects between large institutes with an aim to analyze the co-authorship for different fields of study (Da Silva, Malacarne, e Silva, Kirst, & De-Bortoli (2018)1469,1469); (3) an email communication network to understand the communication pattern between different groups, and identifying important entities (Guimera, Danon, Diaz-Guilera, Giralt, & Arenas, 2003); (4) powergrid networks to identify critical transmission nodes to maintain emergency supply during power outages or blackouts (Watts & Strogatz, 1998); (5) a road network to identify significant junction points that might create bottleneck (Šubelj & Bajec, 2011); (6) a network of airline flights between airports to analyze critical stops/destinations (Batagelj & Mrvar, 1998) (7) problems modeled as an optimal resource allocation with an aim to maximize the total output/yield/use (Cao, Wu, Wang, & Hu, 2011), among other examples.

Let us consider a hypothetical example to understand the problem we are interested in solving. Suppose that a startup company wants to launch a new product, and as a teaser/marketing promotion, they like to offer some free samples to selected consumers. They have a limited number of free samples to offer that can not be given to all. The marketing team aims to select persons who might help them reach their sales target by publicizing the product in their social circles. It may be considered an example of 'word of mouth' marketing strategy generally adopted for new products. But the company might be interested in finding someone with a massive influence on a larger population. Once the consumers' social network is modeled as a complex graph with persons as nodes and their interconnections as edges, it becomes a complex network analysis problem. The above problem could be analyzed by dividing it into two sub-problems. The first sub-problem estimates the individual nodes' importance and then ranks them based on their (information/spreading) capability. The second sub-problem is to find a minimal set of nodes that yields maximal spreading. The second sub-problem is known as the 'Influence Maximization' (IM) problem (Kempe, Kleinberg, & Tardos, 2003; Liu, Li, Chen, & He, 2020) that selects a fixed number of persons (nodes in the modeled network) for a maximal spreading influence which reaches the furthest corner of the network.

Due to advanced data collection techniques and the fact that storage has become cheaper in recent years (Plageras, Psannis, Stergiou, Wang, & Gupta, 2018), an unprecedented amount of data on almost everything is available. Due to the abundance of data when the above-mentioned networks are modeled as complex graph networks, these modeled networks contain very large number of nodes and links, making the analysis and processing computation heavy, and at times not-feasible (Kempe et al., 2003).

Degree centrality (k) (Bonacich, 1972) measures a node's connectivity to its neighbors in a network with the assumption that more number of neighbors yield more spreading influence for a node. Degree centrality is a local measure utilizing only a smaller part of the network while *closeness centrality* (*cc*) (Sabidussi, 1966) and *betweenness centrality* (*bc*) (Newman, 2005) use the complete network information as *cc* is the average shortest distance from all nodes in the network and *bc* is measured as the fraction of different shortest paths between any two nodes that pass through the concerned node.

Computing betweenness or closeness centrality of all nodes in a large network is computationally infeasible or at least very much resource-intensive and time-consuming. To overcome this issue, a new method called k-core or k-shell decomposition (Carmi, Havlin, Kirkpatrick, Shavitt, & Shir, 2007; Pittel, Spencer, & Wormald, 1996) emerged almost two decades ago. The k-shell is an iterative pruning based graph decomposition method that groups the vertices of a graph into various layers, and depending on the K-shell value, the innermost core–shell nodes have the highest k-shell value while the outer shell nodes have the smallest. The influentiality of a node depends on the position of the node within the shells. It is shown that a core–shell node has more influencing ability than a node from outer shells (Kitsak et al., 2010). K-shell has a drawback of putting many nodes into one shell, i.e., it allocates the same rank to many nodes. The top influential nodes identified by k-shell tend to cluster into small, densely connected parts that limit the overall spreading efficiency. Some studies have identified that not always the highest core nodes are the best spreaders (Ma, Ma, Zhang, & Wang, 2016).

Many improvements to the classical k-shell method have been proposed in recent years. There are two types of improvement done on k-shell. One category of improvement directly modify the decomposition technique or upgraded the attributes based on which the shell decomposition is performed (Zeng & Zhang, 2013; Wei, Liu, Wei, Gao, & Deng, 2015; Zareie & Sheikhahmadi, 2018; Liu, Tang, Zhou, & Do, 2015). Zeng and Zhang (2013) used *mixed degree* (see Section 3.6) of nodes instead of *degree* while pruning the network and termed the derived shells as 'M-shell' while the method named as 'mixed degree decomposition'. On the other category of improvement, k-shell is used in conjunction with some other parameters such as degree, closeness, shortest distance, among other, to arrive at some hybrid node ranking methods (Chen, Lü, Shang, Zhang, & Zhou, 2012; Bae & Kim, 2014; Li et al., 2019; Pei, Muchnik, Andrade, Zheng, & Makse, 2014; Maji, Namtirtha, Dutta, & Malta, 2020; Liu, Wang, & Deng, 2020). Recently, Wang, Li, Guo, Peng, and Li (2020) have computed the entropy of nodes using node degree and the total degree of a network to improve the ranking monotonicity (uniqueness of the ranked elements). A similar node information entropy based approach is used in the *EnRenew* method (Guo et al., 2020). Yang, Benko, Cavaliere, Huang, and Perc (2019) proposed a weighted degree decomposition method to rank and identify cheaters in a dynamic evolving cooperation network.

Many researchers model the problem of choosing the minimal starting nodeset to maximize the spreading reaching to a certain fraction of the network as an optimization problem. It is also termed as an *Influence Maximization* (IM) problem (Kempe et al., 2003; Liu et al., 2020). A complete solution of IM problems is not feasible due to NP-completeness (Kempe, Kleinberg, & Tardos, 2015). To solve this problem, there are many heuristic measures based on different network

attributes, that estimate the importance of an interest node in any network.

As we need an epidemic model as a benchmark to compare performances of different node ranking techniques. There are some benchmark simulation models to proxy the actual spreading dynamics on a network. The most referred in the existing literature are *Independent Cascade* (IC) (Kleinberg, 2007), *Linear Threshold* (LT) (Kempe et al., 2003), and epidemic models like *Susceptible-Infected-Recovered* (SIR) (Pastor-Satorras & Vespignani, 2001), *Susceptible-Infected* (SI) (Pastor-Satorras, Castellano, Van Mieghem, & Vespignani, 2015), *Susceptible-Infected-Susceptible* (SIS) (Pastor-Satorras et al., 2015), *Susceptible-Infected-Recovered-Susceptible* (SIRS) (Pastor-Satorras et al., 2015) or *Susceptible-Exposed-Infected-Recovered* (SEIR) (He, Peng, & Sun, 2020). Wang et al. (2016) presents a detailed review of the various epidemic models. In the last year, because of the COVID-19 pandemic situation, many new epidemic models has been proposed to estimate the various future outcomes, and new parameters that could be used to slowdown or contain the spreading (Arefin et al., 2019; Kabir & Tanimoto, 2020). Most of the models are complex in nature (extensions of the SIR or SEIR models with much more parameters) and they are in their nascent stage of development. They are evolving with every new study in the epidemiology domain, in fact we think that they still need more time to show some maturity.

We choose to use SIR since it is a mature model widely tested and used among the OSN research community. It is mostly used to approximate the actual spreading behavior of an infection or information diffusion. Individual node's spreadability is estimated by simulating an infection spreading process using benchmark SIR model simulation for a large number of times to account for its probabilistic nature. SIR model could not be directly used for ranking nodes as the SIR simulations for large and complex networks are computation heavy and time-consuming, so different node ranking heuristics are proposed that can estimate the actual spreading using some network topology and attributes in almost linear time. These heuristics are evaluated by comparing them with the node ranking generated by SIR simulations (Moreno, Pastor-Satorras, & Vespignani, 2002; Newman, 2002). In the present study, the SIR benchmark simulations are used to compare the performance of the proposed technique to the state-of-the-art mainly due to its simplicity and wide use among OSN research community. It is important to note that the computation-intensive SIR model is used in our study only to establish that the proposed method could effectively shadow the benchmark with less complexity. Also, the SIR model generated node rankings are considered as the benchmark standard for comparing other state-of-the-art. A vital evaluation metric that compares items' relative position in two lists, known as Kendall's Tau rank correlation (Kendall, 1945), will be employed along with a monotonicity metric (Bae & Kim, 2014) that determines the items' uniqueness in a ranked list.

We observed that *mixed degree decomposition* (mdd) or M-shell method is better in ranking the influential nodes than K-shell and also generates more unique ranks but rarely authors have exploited its full potential in designing hybrid heuristic measures. K-shell is widely used in all kinds of hybrid methods as a coreness component. In the present study we aim to exploit mdd method as a measure of coreness along with k-shell for a comparative evaluation. When different node-ranking methods are used in IM problems to select the initial multiple seed nodes, the most trivial way is to select the top-k nodes from the ranked list. In many IM studies (Namtirtha, Dutta, & Dutta, 2018; Wang et al., 2020) spreading overlap is not considered, i.e., when the total final spreading influence for a seed nodeset is computed, many nodes get counted more than once as they might have infected from more than one seed node. In such cases, the spreading will be lower than the computed metric.

The goal of this article is to present a node ranking heuristic template that balances between local (node degree) and global (k-shell/m-shell and closeness) measures for selecting the most powerful node in a large network with a critical analysis on influence overlap. Two variants of the proposed template are evaluated; one with k-shell as coreness component and another with m-shell. The present study considers the influence overlap problem (i.e., the same node reached by multiple seed nodes is considered only once) and alleviates it with a modification of the existing models and metrics during experimentation. This heuristic further extends the single node identification heuristics to a seed nodeset selection problem by employing an improved seed node selection criteria to improve the total spread in comparison to the state-of-the-art methods.

The rest of this paper is organized as follows. Theoretical definitions are presented in Section 2. State-of-the-art is presented in Section 3. Our proposed methods are detailed in Section 4, and our proposed improved seed nodeset selection criteria in Section 5. Section 6 presents the methodology used in the evaluation of the proposed method and of the proposed improved seed nodeset selection technique; it includes the experimental setup, the description of the experiments, the description of the network datasets and the simulation model used in the experiments; it finishes with the metrics used in the evaluation. The experimental results are presented and discussed in Section 7. Finally, in Section 8 we conclude the paper and present future work.

2. Theoretical background

A network modeled as a graph is represented as $G < V, E, W >$ where V is the set of nodes, $|V| = n$, the number of nodes, E the set of all edges connecting any two nodes, and W a weighted adjacency matrix with $w_{ii} = 0$. Mathematically,

$$V = \{u, v, \cdots\}$$

$$E = \{e_{uv} = e_{vu} = \{u, v\} | u, v \in V\}$$

$W = [w_{uv}]$, the weight or connection strength of edge e_{uv} between nodes u and v, also $w_{uv} = 0$. In the above definition of a graph, we have $e_{uv} = e_{vu}$ for an undirected network, also the edge weights $w_{uv} = w_{vu}$. Again, the weighted adjacency matrix elements contain the normalized edge weights i.e., $w_{uv} \in [0, 1]$. In the case of a unweighted graph, all the edge weights are the same i.e., $w_{uv} = 1$. The present study focus on the undirected and unweighted models of complex networks hence, our graph representation can be simplified as $G < V, E >$ and defined as a binary adjacency matrix $A_{n \times n}$ where

$$a_{uv} = \begin{cases} 0, & \text{if an edge is connecting node } u \text{ and node } v \\ 1, & \text{otherwise} \end{cases}$$

Nodes in the networks modeled as graphs are ranked based on their spreadability by employing the SIR epidemic model. A spreading process started from any arbitrary node does not yield the same final spread. Based on the topological location, some nodes become more efficient in spreading information or an infectious disease. Such nodes are also known as 'super spreaders', and they are the most influential ones. Graph models of networks are utilized in ranking nodes depending on

spreading ability. Different heuristics formulate their measures using various graph attributes (e.g., shortest path, neighborhood, number of edges, to name a few) to rank the nodes for identifying the influential ones. In the following section, we discuss such state-of-the-art techniques in detail.

3. State-of-the-art

3.1. Introduction

Ranking the nodes of a complex network based on their influence-ability has become mainstream research recently. It is generally achieved with some proxy attributes of the nodes in terms of network topology and node's unique location within the network. Many heuristics are being proposed frequently for many different variations of the problem of identifying the most important node or the super spreader in a network, and many other are proposed to maximize the total influence (Kempe et al., 2003; Kempe et al., 2015) with a minimum number of initial seed nodes. Many different approaches have been taken for these two similar but different problems.

Initially, many centrality measures based on network topology were proposed and applied for the single super spreader identification such as degree centrality (Bonacich, 1972), betweenness (Newman, 2005), closeness (Sabidussi, 1966), k-shell centrality (Carmi et al., 2007), among others. After these first proposals, many other types of measures were used for node ranking such as PageRank (Page, Brin, Motwani, & Winograd, 1999), LeaderRank (Lü, Zhang, Yeung, & Zhou, 2011), VoteRank (Zhang, Chen, Dong, & Zhao, 2016), HITS (Kleinberg, 1999), HybridRank (Ahajjam & Badir, 2018), improved HybridRank (Bhat, Aggarwal, & Kumar, 2020), h-index (Lü, Zhou, Zhang, & Stanley, 2016), extended h-index (Zareie & Sheikhahmadi, 2019), among others. Many of the above methods were extended to weighted networks such as the weighted k-shell (Al-garadi, Varathan, & Ravana, 2017), the weighted LeaderRank (Li, Zhou, Lü, & Chen, 2014) and the weighted h-index (Gao, Yu, Li, Shen, & Gao, 2019).

A new trend of hybrid centrality measures is emerging where more than one classic centrality measures are combined using some weight parameters. With proper tuning of the parameters, these methods generally provide competitive performance. Some of the recent methods in this category are the local gravity model (Li et al., 2019), weighted gravity model (Liu et al., 2020) that combines node's interaction-ability, the degree and distance, generalized gravity method (Li, Shang, & Deng, 2021) that uses local clustering coefficients along with node degree, local centrality method (Chen et al., 2012), the neighbors' degree method (Pei et al., 2014), coreness centrality method (Bae & Kim, 2014), multi-local dimension method (Wen, Pelusi, & Deng, 2020) based on fractal property, K-shell hybrid method (*ksh*), and its improved version (*kshi*) (Namtirtha et al., 2018; Maji et al., 2020). A few potential edge weight-based methods are also proposed where each edge of an unweighted network is assumed to have potential weights. Different methods are proposed depending on how these edge weights are designed (Wang, Hou, Li, & Ding, 2017; Shao, Liu, Zhao, & Liu, 2019; Namtirtha, Dutta, & Dutta, 2020; Maji, 2020). Very recently, Shang, Zhang, Li, and Deng (2021) have proposed another hybrid method that employs iterative edge weight updating based on network efficiency. Yang, Cavaliere, Zhu, and Perc (2020) employed a multiple ranking strategy to identify the influential cheaters in a structure cooperative network where relationships evolve dynamically.

As already discussed in the Introduction, an exact solution to the IM problem (Kempe et al., 2003) is not feasible for even moderately large networks. So to overcome this problem, researchers have used evolutionary (Weskida & Michalski, 2019), bio-inspired (Gao, Lan, Zhang, & Deng, 2013), and many other meta-heuristic techniques such as gray-wolf optimization (Zareie, Sheikhahmadi, & Jalili, 2020), discrete particle swarm optimization (Han, Zhou, Tang, Yang, & Huang, 2021), artificial bee colony optimization (Sheikhahmadi & Zareie, 2020) to rank nodes. In some application domains, it is also termed as a constrained-budget problem (Banerjee, Jenamani, & Pratihar, 2019), where using a fixed number of initial nodes [5], a maximum spreading is expected [6].

Morone and Makse (2015) have employed the 'optimal percolation theory' to identify the top seed nodes for a maximum spread or collective influence. Zareie et al. (2020) defined the best seed nodeset as those nodes in a network upon removal of whom the whole network breaks into many parts. Recently, Hong, Qian, and Tang (2020) proposed a greedy heuristic using reverse influence estimation to competitive influence maximization problems with a theoretical approximation guarantee. Maji, Mandal, and Sen (2020) provide an informative survey of K-shell hybrid methods with a detailed step-by-step computation guide. It helped to implement and verify existing methods.

In the following subsections relevant methods[7] are detailed with their formulation and computation measures.

3.2. Degree centrality

Degree centrality (Bonacich, 1972) is measured as the number of edges incident on a node. The degree (k) of a vertex v in an undirected graph is defined as presented in Eq. (1).

$$k(v) = \sum_{u \in \phi(v)} a_{uv} \tag{1}$$

where $\phi(v)$ is the set neighboring vertices of v. A normalized measure of degree centrality is also defined as $= \frac{k(v)}{N-1}$ where N is the network size (Freeman, 1978).

This is the most simple and easily computable measure for the local influence (Bonacich, 1972). But this measure becomes ineffective when complete topology is present as it is not utilizing any global network information. It is mostly used in conjunction with other centrality measures to counter the limitations.

3.3. Betweenness centrality

Betweenness centrality (Newman, 2005) is one of the global measures using the shortest path when one want to reach a node in a network. It is computed as the ratio of two numbers. The numerator contains the count of shortest paths passing through the concerned node and the denominator is the total number of shortest paths between any two nodes as formulated in Eq. (2).

[5] Assume as the fixed marketing budget to support only a limited number of promotional activity or free sample products

[6] Assume the positive news of the product reached to a certain percentage of the target consumers

[7] These methods are used by the work of Bae and Kim (2014), Namtirtha et al. (2018), Namtirtha et al. (2020), Wang et al. (2020), Ma et al. (2016), Li et al. (2014), Liu, Ren, and Guo (2013), Maji et al. (2020), Chen et al. (2012), Zeng and Zhang (2013), Zhang et al. (2016), Pei et al. (2014), Li et al. (2019), Li et al. (2018), Maji et al. (2020), Maji (2020), Ahajjam and Badir (2018), Al-garadi et al. (2017), Zareie and Sheikhahmadi (2019), Zareie and Sheikhahmadi (2018), and many other recent papers of the field, this is the reason why we consider them thee relevant ones.

$$bc(v) = \sum_{v \neq s \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

where σ_{st} represents the total number of shortest paths between nodes s and t, while $\sigma_{st}(v)$ represents the count of shortest paths between s and t passing through v.

Due to the involvement of all pairs of shortest paths, computation becomes infeasible with large number of vertices and links. Brandes (2001) developed the fastest algorithm to compute bc in $O(N * E)$ time for unweighted networks.

3.4. Closeness centrality

It is also a global centrality metric that uses geodesic during computation (Sabidussi, 1966). It measures the closeness of the concerned node with the rest of the network nodes. It is computed as the reciprocal of the average shortest distance from the concerned node to all others in the network. It is shown in Eq. (3).

$$cc(v) = \frac{1}{\sum_{u \in V \setminus v} d(u, v)} \tag{3}$$

where d(u, v) is the shortest distance between u and v.

3.5. K-Shell Centrality

K-Shell Centrality (Carmi et al., 2007) ranks different nodes by assigning them a non-negative integer value known as the k-shell index (k_s). This process of assigning k_s values to the nodes is done by pruning the network iteratively based on the node degree. Due to the iterative nature of the process, it becomes a global measure. Through the iterative decomposition of the network, nodes are clustered into different shells identified by the k_s value. All nodes in a shell (K-shell) have the same k_s value (see Algorithm 1):

1. The process starts by pruning the nodes with a degree of one in the network and assigning $k_s = 1$ to the nodes.
2. Next, the algorithm checks recursively if any new nodes become degree-1 and removes them by assigning $k_s = 1$.
3. The process follows by repeating the same action (step 1 and 2) but now with degree-2 nodes and removes all such nodes by assigning $k_s = 2$ to all.
4. Again, due to the removal of degree-2 nodes, many new nodes' degree become less or equal to 2, those nodes are recursively pruned and allotted $k_s = 2$.
5. The same process of iterative removal of nodes and assigning them (the nodes) to a k-shell, is continued till there remain only nodes with a degree of 3 or higher.
6. The process repeats until all remaining nodes are tagged with a k_s value.

Batagelj and Zaveršnik (2011) have implemented the above process with nearly linear time complexity. The shell containing the largest k-shell-indexed nodes is termed as the core–shell, and (Kitsak et al., 2010) proved that the nodes that belong to the innermost core are more influential than other nodes in the distant periphery.

An important drawback of the k-shell decomposition is the assignment of lot of nodes to the same k-shell. On Application the K-shell method to the toy network shown in Fig. 2 with 20 vertices and 29 edges, all nodes get divided into three shells. The innermost core–shell consists of nodes e, f, g, h, and the middle shell consists of nodes $a, b, c, d, j, k, m, p, o$ with all other belonging to the outermost shell with $k_s = 3$.

Algorithm 1. K-shell centrality

input : An unweighted network graph $G < V, E >$
parameters: node degree k
output : $Rank[v, k_s]$
where $v \in V$ and k_s is the k-shell index of node v
The list *Rank* contains the nodes with corresponding k_s value
Initially *Rank* contains all nodes of G with k_s value set to 1
$k = 1$
while *!isEmpty(V)* **do**
  repeat
    Find all nodes in G with degree = k
    foreach *such node v with degree k* **do**
      assign $k_s = k$
      update k_s value in the nodelist *Rank*
      Remove node v from G
  until *All remaining nodes in G have node degree* $> k$
  $k = k + 1$

3.6. Mixed degree decomposition

Zeng and Zhang (2013) propose to use the contribution of the removed nodes in the K-shell decomposition process to overcome the monotonicity problem. They termed the number of links to the removed nodes during every iteration as 'exhausted degree' (k^r) and the number of links to the remaining nodes as 'residual degree' (k^e). They have modified the K-shell decomposition process to consider the effect of both residual and exhausted degree, and named the method mixed degree decomposition (mdd). Authors have designed the mixed degree (k^m) of a node as presented in Eq. (4).

$$k^m(v) = k^r(v) + \lambda * k^e(v) \tag{4}$$

The *mixed degree decomposition* steps are presented below:

1. In the beginning of the process there are no removed nodes so, for all nodes in the network: $k^e = 0$ and mixed degree $k^m = k^r$.
2. Next, all nodes with the smallest mixed degree value are removed, k^m ($= M$, say) and they are assigned to M-shell.
3. For all remaining nodes, the mixed degree (k^m) values are recomputed following Eq. (4). Then all nodes with $k^m \leqslant M$ are pruned recursively, and allotted to M-shell until all remaining nodes have a k^m value not less than or equal to M.
4. Repeat Step 2 and Step 3 with the next higher M value until all nodes have been assigned to some M-shell.

The tunable weight parameter λ balances between the K-shell decomposition and the degree centrality methods. When $\lambda = 0$ it becomes the same as K-shell. Again with $\lambda = 1$, it becomes same as the degree centrality method. Also, the allotted indices in K-shell are always

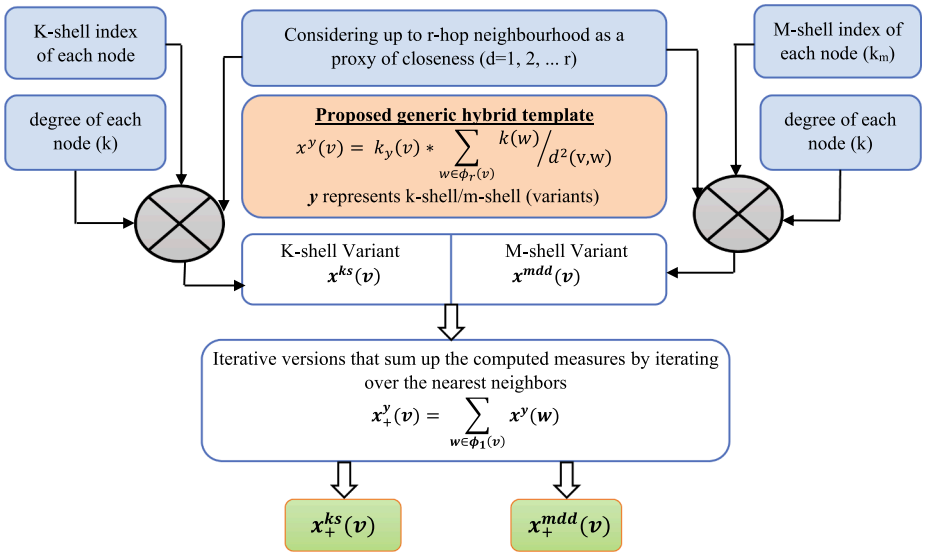


Fig. 1. Block diagram of the proposed method. One variant uses k-shell, degree and closeness in designing the heuristic while the other replaces k-shell by m-shell. From there, we propose iterative versions of both variants. They sum up computed measures of all nearest neighbors for a node v.

whole numbers but the same is not true for M-shell values.

3.7. The improved method (θ)

Liu et al. (2013) have used the shortest distance of a node from the core shell nodes to rank them with the assumption that a node near to a lot of core nodes is more important. Before applying their method, the network is decomposed using the K-shell method, and all nodes are assigned to different k-shells. Nodes having the maximum k_s value belong to the core shell. The improved measure for nodes belonging to the same K-shell is computed using Eq. (5).

$$\theta(v|k_s) = (k_s^{Max} - k_s + 1) \sum_{u \in \Gamma_{k_s^{Max}}} d_{uv}, \quad v \in S_{k_s} \tag{5}$$

where k_s^{Max} is the largest k_s value in the network, Γ_{k_s} is the set of nodes that belongs to the core i.e., with $k_s = k_s^{Max}$. d_{uv} represents the shortest distance (geodesic) between node u and v. S_{k_s} is the set of nodes with k-core index equal to k_s. This method yields competitive performance though it highly depends on k-shell for core nodeset selection and k-shell core shell does not always contain all influential nodes.

3.8. Neighborhood coreness method

The neighborhood coreness heuristic Bae and Kim, 2014 aims at a balanced combination of node-degree as well as the coreness of the nearest neighbors while computing a node's influence. This method overcomes the monotonicity problem associated with the K-shell rankings. Authors have assumed that any node that has more connections to the core-nodes becomes more important than others in terms of their spreadability. Considering the above assumption, Bae and Kim, 2014 have defined the neighborhood coreness shown in Eq. (6).

$$nc(v) = \sum_{u \in \Gamma_v} k_s(u) \tag{6}$$

where Γ_v represents the neighboring nodes of v, and the k-shell index of vertex u is represented by $k_s(u)$. Hence the extended iterative version nc_+ is depicted in Eq. (7).

$$nc_+(v) = \sum_{w \in \Gamma_v} nc(w) \tag{7}$$

This method ranks the nodes well but computing the shortest distance between node-pairs is time consuming and this fact limits the use of the method with very large networks.

3.9. The gravity formula method

Ma et al., 2016 have exploited the law of gravity in formulating their heuristic to compute the influencing ability of a vertex in an spreading phenomena. They have considered the k-shell index of a node as the *mass* and the shortest distance between any two nodes as the *distance*. The gravity formula method is presented in Eq. (8).

$$g(v) = \sum_{u \in \Psi(v)} \frac{k_s(v) * k_s(u)}{d_{uv}^2} \tag{8}$$

where d_{uv} represent the shortest distance between u and v. $\Psi(v)$ represent all nodes up to r-level neighborhood of v.

Ma et al. (2016) have chosen $r = 3$ during their experiments, so, $\Psi(v)$ consists of all vertices in 3 levels of neighborhood of v.

Ma et al. (2016) have also proposed an extended iterative version of the gravity formula method known as 'gravity plus' method (see Eq. (9)).

$$g_+(v) = \sum_{u \in \phi(v)} g(u) \tag{9}$$

where $\phi(v)$ represents all the nearest neighbors. Selecting the level of neighborhood optimally makes a lot of difference in terms of computation. With dense networks and larger values of r, the number of nodes increases, and hence, more shortest paths are required to be computed, making the method computation–heavy.

3.10. Neighbors degree method

Pei et al. (2014) propose a neighbors degree based measure to estimate the spreading influence of the nodes and rank them. They

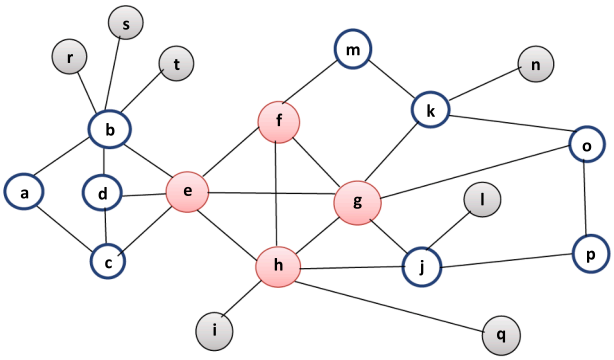


Fig. 2. Example network with 20 nodes, 29 edges, and epidemic threshold $\beta_{th} = 0.322$. Different K-shell nodes are depicted using different colors: nodes $\{e, f, g, h\}$ (in orange) belong to $k_s = 3$, nodes $\{a, b, c, d, m, k, j, o, p\}$ (in blue) belong to $k_s = 2$ and the gray nodes belong to the outermost shell $k_s = 1$.

formulate their heuristic using the sum of the neighborś node degree value of a node as shown in Eq. (10). This method does not use any global measures like K-shell, or computation heavy shortest distances or closeness and instead uses the most simple attribute of a node i.e., degree value. It also works well with the networks that don't have complete network structure information and quick computation is feasible for very large networks.

$$k_{sum}(v) = \sum_{u \in \phi(v)} k(u) \tag{10}$$

3.11. Final considerations

The discussion in this section on different benchmark node ranking methods and their formulation of the heuristics measure to estimate the spreadability suggests that no single measure such as node degree, K-shell, betweenness, or closeness is able to capture the node importance for all kinds of networks. Also, some measures, like closeness or betweenness, are computation heavy for large scale networks and may not be useful. Depending on the network's attributes like dense or sparse, like the level of knowledge of the topological structure of the network, or like the availability of global properties, among other issues, affect deeply the performance of these different benchmark methods (Namtirtha et al., 2020). Many state-of-the-art methods employ some free parameters while designing their heuristics, and tuning those parameters for various networks to achieve optimal results limits the application to unknown new networks (Maji et al., 2020).

We have seen that the *mdd* method uses a free tunable parameter λ, and "somehow"[8] authors have identified a value of "0.7" for the experiments. Many other node ranking methods exist with one or more tunable parameters, and setting their optimal values is entirely empirical (Namtirtha et al., 2018; Namtirtha et al., 2020).

The next section presents a new method that overcomes this problem by proposing a heuristic template that utilizes degree centrality, k-shell/m-shell and an easier mechanism to avoid shortest distance computation. It also avoids the use of any tunable free parameters.

4. The proposed method

The proposed heuristic uses degree centrality as a local proxy of a node's influence and its k-shell index/M-shell index as a global contribution. The proposed method employs a three-hop neighborhood concept to incorporate the benefits of closeness metric without incurring associated high computational complexity. We have proposed two different variants of the heuristic, one using the K-shell index and other using the M-shell index. The K-shell variant, termed as x^{ks}, is formulated in Eq. (11), and the M-shell variant, termed as x^{mdd}, is formulated in Eq. (12). The *mdd* is a more refined and improved measure of K-shell, and nodes have more uniqueness in the ranking than K-shell. Also, most of the literature uses K-shell while devising any hybrid centrality measure, we could not find any hybrid method that employs the M-shell index. This absence of literature made us pursue the idea o exploring the use of the M-shell index, and hence the M-shell variant of the proposed heuristic is proposed. The outline of the proposed method is shown in Fig. 1. The underlying assumptions behind the proposed formulation are based on the following facts:

- It is established that within a small locality node degree is a close proxy of a node's influentiality. In other terms, the number of direct neighbors is analogous to a node's spreading capability in a localized context. A node with more direct connections has more chances of spreading a disease or rumor.
- Nodes with the highest K-shell index values are globally the most densely connected ones. In other terms, a node with a high K-shell index generally belongs to the core community nodes with a large number of interconnections among themselves. All such nodes are important in a spreading phenomenon as any such infected node could practically reach to all within a short time. M-shell is just an improved K-shell, as discussed earlier.
- The importance or influence of a node has an impact on its neighbors, and this extends to very distant neighbors with a diminishing weight. In other terms, a node's importance increase if it is well-connected to many other important nodes, i.e., neighbors' influentiality has contributions on a node's spreading capability. However, considering all distant neighbors like global centrality measures, will make the scheme computationally infeasible. To overcome this, we consider three levels of neighbors while computing the spreadability of a node: $x^{ks}(v)$ (see Eq. (11)) and $x^{mdd}(v)$ (see Eq. (12)).

$$x^{ks}(v) = k_s(v) * \sum_{w \in \phi_r(v)} \frac{k(w)}{d^2(v,w)} \tag{11}$$

where, $\phi_r(v)$ is the set of r level neighbors of node v.

$$x^{mdd}(v) = k_m(v) * \sum_{w \in \phi_r(v)} \frac{k(w)}{d^2(v,w)} \tag{12}$$

where, $k_m(v)$ is the M-shell index of node v, and $\phi_r(v)$ is the set of r level neighbors of node v.

Like many other hybrid measures we also propose the iterative versions of the two variants. These are termed as x_+^{ks} and x_+^{mdd}, respectively, and the formulation is shown in Eq. (13) and Eq. (14). Both of the above methods (x^{ks} and x^{mdd}) iterate over the neighborhood to assign the iterative measure of a node.

$$x_+^{ks}(v) = \sum_{u \in \phi(v)} x^{ks}(u) \tag{13}$$

where, $\phi(v)$ denotes all the neighboring nodes of v.

[8] Meaning that authors use the "trial error" technique to find the best parameter.

$$x_{+}^{mdd}(v) = \sum_{u \in \phi(v)} x^{mdd}(u) \tag{14}$$

where, $\phi(v)$ denotes all the neighboring nodes of v.

To illustrate the computational steps, we compute the x^{ks} measure of a concerned node (e) on the toy network in Fig. 2. We take ($r = 3$) hop/level of neighbors for computation. The 3-hop neighboring vertices of e are as shown next: first level $\{b,c,d,f,g,h\}$, second level $\{a,r,s,t,m,k,i,j,q,o\}$, and third level $\{l,p,n\}$.

$$x^{ks}(e) = k_s(e) * [\{\frac{k(b)}{d^2(e,b)} + \frac{k(c)}{d^2(e,c)} + \frac{k(d)}{d^2(e,d)} + \frac{k(f)}{d^2(e,f)} + \frac{k(g)}{d^2(e,g)} + \frac{k(h)}{d^2(e,h)}\} +$$
$$\{\frac{k(a)}{d^2(e,a)} + \frac{k(r)}{d^2(e,r)} + \frac{k(s)}{d^2(e,s)} + \frac{k(t)}{d^2(e,t)} + \frac{k(m)}{d^2(e,m)} + \frac{k(k)}{d^2(e,k)} + \frac{k(i)}{d^2(e,i)} + \frac{k(j)}{d^2(e,j)} + \frac{k(q)}{d^2(e,q)} + \frac{k(o)}{d^2(e,o)}\} +$$
$$\{\frac{k(l)}{d^2(e,l)} + \frac{k(p)}{d^2(e,p)} + \frac{k(n)}{d^2(e,n)}\}] = 3 * [\{\frac{6}{1} + \frac{3}{1^2} + \frac{3}{1^2} + \frac{4}{1^2} + \frac{6}{1^2} + \frac{6}{1^2}\} +$$
$$\{\frac{2}{2^2} + \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^2} + \frac{2}{2^2} + \frac{4}{2^2} + \frac{1}{2^2} + \frac{4}{2^2} + \frac{1}{2^2} + \frac{3}{2^2}\} +$$
$$\{\frac{1}{3^2} + \frac{2}{3^2} + \frac{1}{3^2}\}]$$
$$= 100.33$$

detailed in Algorithm 2 for a minimum distance (md) of 2 hops.

This implementation allows the selection of seed nodes to maintain a minimum threshold distance of 2 hops without using computation heavy classical shortest path algorithms. It only explores the 2-hop neighborhood of the seed nodes, and if the next node does not belong to them, it is added to the seed nodeset. The time complexity of the method with a minimum distance threshold of md becomes $O(N* < k >^{md-1})$.

The process is as follows: we first rank the nodes as usual, and then while selecting seed nodes, instead of selecting top-k nodes, we select them by maintaining a minimum distance between the seed nodes. We first select the top-ranked node as a seed and then check its minimum

Algorithm 2. Improved seed nodeset selection maintaining a minimum distance of 2 hop.

input : An unweighted network graph $G < V, E >$, Ranked list (R) of nodes based on any heuristics/centrality metric
parameters: Proportion of initially infected nodes, p
output : Set of initial seed nodes Φ for maximal total final spread
begin
  Let, $N = |V|$ = size of the network graph
  $seed\_count = p * N$
  Initially, seed nodeset Φ ={ }
  nodeset Neighbors Φ_N ={ }, the set of first hop neighbors of all nodes in Φ
  $node\_count$ = 0;
  Rank index r = 1;
  while ($node\_count \leq seed\_count$) **do**
    $U = \{u\} = R[r]$ = set of nodes with rank r
    foreach (*node* $u \in U$) **do**
      if $u \notin \Phi_N$ **then**
        $\Phi = \Phi \cup u$ * Add u to Φ */
        $\Phi_N = \Phi_N \cup G.neighbors(u)$ * Add all neighbors of node u to Φ_N */
        $node\_count$++;
    r++;
  Return Φ
end

5. Improved seed node selection technique

A trivial approach to select k-seed nodes is to take top-k nodes sequentially from the ranked list. The K-shell generated ranking exhibits a "rich-club" phenomenon where most crucial nodes are tightly connected into a small core area. With such a phenomenon, selecting top-k nodes as the optimal set does not yield good collective spreading. Instead, selecting seed nodes in a distributed fashion may allow a wide-spreading (Liu, Jing, Zhao, Wang, & Song, 2017). Hu, Liu, Yang, and Ren (2014) have observed that the distance among the multiple initial spreaders greatly influences total spreading.

We extend the proposed heuristics to select a set of initial nodes by employing a minimum threshold distance criteria. The process is

distance from the second node. If the distance is less than the threshold, we skip that node and continue the process with the next node in the list. As soon as we find another node that maintains the minimum distance, we add it to the seed nodeset. The process continues with the next node and computes the distance from all nodes in the seed nodeset. It continues until k nodes are chosen as seed nodeset.

We assume that for a fixed number of seed nodes, the more the lower rank we reach, the more distant and loosely connected clusters are reached. Also, selecting seed nodes that are not very close to each other, it provides a wider spread/diffusion throughout the network. To corroborate the above fact, we compute the last seed node's rank for all the cases. We compare the last rank up to which the search reaches while selecting seed nodes, maintaining a minimum distance between them.

Let us now investigate the effects of placing a minimum distance constraint while choosing the seed nodeset from the top of the ranked list generated by different heuristics. The reasoning behind such constraint is that it is found that most of the cases, top ranked nodes form densely connected clusters or cliques, and selecting any one node from each such cluster is expected to be more beneficial than considering all such nodes from a few clusters. For example, let us consider the toy network and k-shell ranking with a budget of 15%, i.e. (20 $*$ 15%=) 3 seed nodes. As per the simple rule of taking nodes from the top, we would select any three nodes among *e*,*f*,*g*, and *h* as all are having a rank of 1 (with highest k-shell value = 3). With a minimum distance constraint ($minDist \geqslant 1$), the selected three seed nodes will be *e*, *a*, and *k*. Let us now consider the same with degree centrality (*k*). In this case, the top three nodes are *b*,*e*, and *g*, whereas with keeping a minimum distance of two hops between selected seed nodes, we have three nodes as *b*, *h*, and *k*.

6. Methodology for evaluation of the proposed methods

6.1. Introduction

In order to evaluate our proposed methods we have conducted a series of experiments. The goal is to compare the performance of our proposed methods with the state-of-the-art methods (degree (*k*), k-shell (*ks*), *mdd*, improved method (θ), neighborhood coreness (nc_+), gravity plus (g_+), and neighbors' degree (k_{sum})) when:

1. finding the super spreader on a network;
2. ranking the nodes based on their spreadability, and;
3. searching for a set of initial seed nodes for maximal collective influence, using the improved seed node selection technique presented in Section 5.

We perform the following experiments to achieve the preceding three objectives:

- Experiment 1 and 2 evaluate the proposed methods, and compare them with the recent methods for single super spreader identification on a small toy network as a controlled experiment (Experiment 1), and then, on real-world networks (Experiment 2).
- Experiment 3 explores the effect of various minimum distances on the k-shell method used in selecting multiple initial seed nodes to identify an empirical optimal value that yields a maximal final spread without overlap.
- Experiment 4 applies the improved seed node selection using a minimum distance criteria on the proposed methods as well as on state-of-the-art's with the goal of comparing their performance improvement for a maximal final spreading influence in the network.

The next sections present the experimental setup, the details of the experiments, the details of the network datasets and the simulation model used in the experiments, and the evaluation metrics of the experiments.

6.2. Experimental setup

Experimental simulations were conducted on a macOS Catalina v10.15 system with Intel core i5 dual-core 1.8 GHz, 8 MB of RAM and 128 GB Hard Disk. The algorithms were implemented in Java 1.8.

6.3. Experiments

EXPERIMENT 1

Goal: To rank the vertices of our toy network depending on the measures computed by the proposed methods and state-of-the-art as a controlled experiment.

Description: The proposed methods and state-of-the-art methods rank the nodes of the toy network, and compute the heuristic measure values assigned to the nodes. The used small toy network helps in understanding the computational details and also allows for manual verification of the results without any software implementation.

Results: Results are presented and discussed in Section 7.1;

EXPERIMENT 2

Goal: To compare the performance of the proposed methods and the state-of-the-art methods in the selection of the most important node, and in the ranking of the nodes in real-world networks.

Description: The methods are employed on real-world benchmark datasets (see Sub-Section 6.4 for details about this network datasets). We use the average Kendall's correlation coefficient, the average spreading influence (σ_{avg}) and the monotonicity values of the different methods for the comparison (see Sub-Section 6.6 for details about this metrics). During the computation of σ_{avg}, we have considered various proportions (p) of initial seed nodes from 0.01 to 0.1, i.e., top 1% nodes to top 10% nodes to compare the effectiveness of different methods without any bias. Like (Guo et al., 2020), the β values are taken as 1.5 $*$ β_{th} during the simulation.

Results: Results are presented and discussed in Section 7.2.

EXPERIMENT 3

Goal: To explore the effect of influence overlap on the total spreading and show that a minimum distance criteria in the seed node selection improves the total spread on K-shell without overlap.

Description: In experiment 2 we were interested in ranking individual nodes, so overlap was not an issue. The effect of influence overlap was not considered during the computation of σ_{avg}. This experiment aims to select a set of starting seed nodeset that yield a maximum total spread, but in such an experiment, the 'influence overlap' produces inaccurate results. So, to compensate for the above, we compute and compare the total spreading influence without overlap (σ_{total}). It gives the total influence in terms of the total number of infected nodes after the infection dies out in the SIR model. A trivial approach to select k-seed nodes is to take the top-k nodes from the ranked node-list. But we have used a minimum geodesic based seed nodeset selection technique. We test our hypothesis that using a minimum distance criteria between seed nodes yields improved total spreading influence on the K-shell method. Instead of selecting the seed nodes consecutively from the ranked list, we choose them, maintaining a minimum distance between them (see Algorithm 2). We compare the total spreading without overlap due to different minimum distance (1-hop, 2-hop and 3-hop), and with different fraction of initial seed nodes (5%, 10%, and 15% of total nodes) on k-shell method and empirically obtained the optimal value of minimum distance.

Results: Results are presented and discussed in Section 7.3.

EXPERIMENT 4

Goal: To establish that our improved seed nodeset selection technique performs better than the top-k nodeset selection technique.

Description: We choose seed nodes by employing the improved seed node selection technique defined in Section 5, maintaining a minimum of 2 hop distance between the seed nodes in our proposed methods. In contrast, all other methods follow the common top-k nodes as seed nodes. As before, we have considered p values as 0.01 to 0.1 i.e., 1% to 10% of total nodes as initial seed nodeset. For each such nodeset, the modified SIR model is run for 1000 times, and the average number of infected nodes is taken as the total influence (σ_{total}) of that seed nodeset. We compare the total spreading due to multiple initial seed nodes with the state-of-the-art methods. We enforce our improved seed selection technique with a minimum distance of 2 hop and calculate the percentage improvement in total spreading on the state-of-the-art methods. We have computed the improvement in total spreading influence (without overlap) with different fractions of seed nodes maintaining a minimum distance of 2 hops as

$$\%\sigma_I = \frac{\sigma_{total}^{md=2} - \sigma_{total}}{\sigma_{total}} * 100\%$$

.

Results: Results are presented and discussed in Section 7.4.

6.4. Network datasets

We have used the following network datasets:

- Net Science (Newman, 2003): research collaboration network of scientists working on network theory based on co-authorship relation;
- C. elegans (Duch & Arenas, 2005): an interconnected graph of neurons in C. elegans worms;
- Blogs (Adamic & Glance, 2005): a network of hyperlinks between different blog posts on US politics during 2004 US elections;
- Advogato (Massa, Salvetti, & Tomasoni, 2009): a user-to-user trust relationship network on a OSN community platform called Advogato;
- Zachary (Zachary, 1977): a friendship network between 34 members of a karate club at a US university;
- Powergrid (Watts & Strogatz, 1998): a network representation of the Western States Power Grid of the United States;
- Hamsterster (Konect, 2017): a social network of friendship between users of the website urlhttp://www.hamsterster.com;
- USAir97 (Batagelj & Mrvar, 1998): graph of the US air transportation network;
- USAirports (Opsahl, 2013): a network of 500 US airports connected by direct flight paths;
- Jazz (Gleiser & Danon, 2003): a collaboration network of jazz musicians;
- DBLP-cite (Ley, 2002): a citation network of DBLP, a scientific publication database;
- Euroroads (Šubelj & Bajec, 2011): infrastructure network of European roads;
- Macaques (Takahata, 1991): a network of dominance relationship among Japanese female monkeys;
- Condmat (Leskovec, Kleinberg, & Faloutsos, 2007): a co-authorship network between researchers posting pre-prints on the Condensed Matter Archive;
- PGP (Boguná, Pastor-Satorras, Díaz-Guilera, & Arenas, 2004): a network of users of the secure data exchange algorithm named Pretty-Good-Privacy;
- ODLIS (Reitz, 2002): a network of different terms in Online Dictionary of Library and Information Science, where two terms are connected if one term is used to express the meaning of the other term;
- AstroPh (Leskovec et al., 2007): a co-authorship network between researchers posting pre-prints on the Astrophysics Archive;
- CA-HepTh (Leskovec et al., 2007): a co-authorship network between researchers posting pre-prints on the High Energy Particle Theory Archive;
- CA-GrQc (Leskovec et al., 2007): a scientific collaboration network between authors on pre-prints submitted to General Relativity and Quantum Cosmology category of Arxiv;
- CA-CSphd (Leskovec et al., 2007): a network of PhDs and their supervisors in theoritical computer science;
- Email-EU-core (Yin, Benson, Leskovec, & Gleich, 2017): a network of email exchanges between the members of a large European research institute;
- Wiki-Vote (Leskovec, Huttenlocher, & Kleinberg, 2010): wikipedia who-vote-whom for adminship network.
- Facebook (McAuley & Leskovec, 2012): A directed user-user friendship ego network in facebook.
- Brightkite (Cho, Myers, & Leskovec, 2011): A shared location based friendship network where users are the nodes and edges represent the friendship among them.

We present the following standard network attributes in Table 1 computed using Gephi (Bastian, Heymann, & Jacomy, 2009) for the network datasets used in the study:

- vertex count (V); Also referred as network size.

Table 1
Network parameters of the selected benchmark datasets used in the experiments. V: vertex count; V': number of vertex in the largest connected component; E edge count; β_{th}: SIR model infection threshold (Newman, 2002); D: network diameter; $< k >$: average node-degree; p<mean>: average shortest path length; C: average clustering coefficient; ρ: graph density; β: infection probability used during SIR simulation. All measures are calculated using Gephi (Bastian et al., 2009) with network representation as undirected, unweighted, without self-loop and duplicate edges.

Networks	V	V'	E	β_{th}	D	<k>	p<mean>	C	ρ	β
Net Science	1589	1461	2742	0.144	17	3. 415	5. 823	0. 878	0. 002	0.216
C. elegans	306	297	2148	0.038	5	14.039	2. 455	0. 308	0. 046	0.057
Poll-Blogs	1490	1224	16718	0.012	8	22. 44	2. 738	0. 361	0. 015	0.018
Advogato	6541	5155	51127	0.012	9	7. 819	3. 275	0. 408	0. 002	0.018
Zachary	34	34	78	0.129	5	2. 294	2. 408	0. 588	0. 139	0.194
Powergrid	4941	4941	6594	0.258	46	2. 669	18. 989	0. 107	0. 001	0.387
Hamsterster	1858	1858	12534	0.022	15	8. 104	5. 104	0. 136	0. 007	0.033
USAir97	332	332	2126	0.023	6	6. 4	2. 738	0. 749	0. 039	0.035
USAirports	1858	1574	17215	0.009	8	21.874	3.115	0.637	0.014	0.014
Jazz	198	198	2742	0.026	6	13. 848	2. 235	0. 633	0. 141	0.039
DBLP-cite	12591	12591	49743	0.023	10	3. 951	4. 423	0. 192	0. 001	0.035
Euroroads	1174	1174	1417	0.333	62	2. 414	18. 371	0. 02	0. 002	0.5
Macaques	62	62	1187	0.026	2	19. 145	1. 383	0. 667	0. 617	0.039
Condmat	16726	16264	47594	0.084	0	5.691	6.628	0.737	0	0.126
PGP	10680	10680	24316	0.053	24	2.276	7.487	0.44	0	0.08
ODLIS	2909	2900	18241	0.014	8	11.263	3.17	0.351	0.004	0.021
AstroPh	18821	18771	198050	0.016	14	21.05	4.194	0.677	0.001	0.024
CA-HepTh	9879	9875	25973	0.087	18	5.263	5.945	0.6	0.001	0.131
CA-GrQc	5244	5241	14484	0.063	17	5.529	6.049	0.687	0.001	0.095
CA-CSphd	1891	1882	1740	0.263	34	1.85	11.856	0.016	0.001	0.394
Email-EU-core	1005	986	16064	0.0136	7	31.968	2.587	0.45	0.032	0.02
Wiki-Vote	7117	7115	100762	0.00693	7	28.316	3.248	0.209	0.004	0.01
Facebook-ego	4039	4039	88234	0.0095	8	43.69	3.693	0.617	0.011	0.01
Brightkite	58228	58228	214078	0.016	18	7.35	7.37	0.271	0	0.05

- vertex count in the largest connected component (V');
- edge count (E);
- SIR epidemic model infection probability threshold (Newman, 2002) (β_{th});
- network diameter (D): largest value of the shortest path distance between any two nodes;
- average node-degree ($< k >$);
- average shortest path length (p<mean>): the length of the shortest paths averaged over all the node-pairs of the largest connected component of a network;
- average clustering coefficient (C): the average local clustering coefficient over all the nodes;
- graph density (ρ): the ratio of the number of edges present in the network to the number of possible edges. Very small value of ρ of a network indicates minimally connected sparse network, and;
- infection probability used during the SIR simulation (β).

6.5. Simulation model

We use the SIR epidemic model (Pastor-Satorras & Vespignani, 2001; Newman, 2002) to emulate the real spreading phenomena on a network. There are many different models to emulate real spreading phenomena such as Threshold models (Kempe et al., 2003; Tong, Li, Wu, & Du, 2016), cascade models (Kleinberg, 2007; Song, Zhou, Wang, & Xie, 2014), and compartmental epidemic models such as SIR, SIS, SIRS, SEIR among others (Wang et al., 2016). A large number of literature uses the SIR model as a benchmark while comparing performance of different node ranking heuristics. The present study uses the SIR epidemic model (Pastor-Satorras & Vespignani, 2001; Newman, 2002)[9] to measure the spreading influence of the nodes in an unweighted network.

In this propagation model, all nodes are initially considered to be in *Susceptible* (S) state except the seed node which is in the *Infected* (I) state. The goal is to simulate the infection spread in the network due to the infected seed node over time. At every time interval, an infected seed node propagates the contagion to its surrounding "S" nodes with a propagation probability β. It changes its state as *Recovered* (R) with a probability β_r. A "R" node is immunized and cannot be re-infected. There are many variations of the model. Some studies consider recovered and removed nodes alike i.e., not all nodes survive the infection and die. In another variation, the population size remains unchanged, so no one dies. However, in both variations, the common fact remains the same. Once a node is in the "R" state, then it can not get infected again. This epidemic propagation runs till there remains no more infected nodes in the network. We assume complete recovery/immunity of the infected nodes in subsequent time-steps, i.e., recovery probability $\beta_r = 1$with a fixed population i.e., no node is removed/died. With a larger β value, the contagion quickly reaches the whole network; again with a smaller value, spreading is contained within a minuscule portion of the network. The epidemic threshold is estimated as $\beta_{th} \approx \frac{<k>}{<k^2>}$, where $<k>$ denotes the average degree and $<k^2>$ denotes the second-order average of node-degree Moreno et al., 2002. An epidemic spread with an infection probability value less than the threshold dies out quickly without spreading beyond the local neighborhood. To observe the real epidemic spreading or information diffusion process β values are considered above the epidemic threshold value, β_{th}, to cause a larger outbreak. The portion of the "R" nodes measures the spreading efficiency of the seed node. We have modified the standard model to incorporate the spread-overlap effect while computing total spread for a seed nodeset. The modified algorithm used to simulate with an initial seed nodeset instead of a single node is presented in Algorithm 3. During our experiments, simulations were run 1000 times for small networks with less than 10,000 nodes and 100 times with larger networks.

Algorithm 3. The SIR epidemic model for total influence without overlap.

input : An unweighted network graph $G < V, E >$, top p fraction of initially infected nodes
parameters: Infection probability threshold β_{th}
output : total spread σ_{total} in terms of total number of recovered nodes after infection dies out
begin
  Mark/Label top p fraction of nodes as "I" and rest as "S"
  $R\_node\_count = 0$
  $isInfected = true$
  while *(isInfected)* **do**
    foreach *node* $v \in V$ **do**
      if *(v.label == Infected)* **then**
        $isInfected$ = false
        foreach *node* $u \in v.neighbors$ **do**
          if *(u.label==susceptible AND Math.random* $> \beta_{th}$*)* **then**
            u.label = *Infected*
            $isInfected = true$
        v.label = *Recovered*
        R_node_count++
  Return $\sigma_{total} = R\_node\_count$
  end

6.6. Evaluation metrics

To evaluate the performance of our ranking method we consider undirected, un-weighted, no self-loop and no parallel links representation of the network datasets presented in Table 1.

The performances of the proposed methods are compared as follows:

- we compare the methods according to their ability in identifying the top super spreader i.e., ranking the vertices depending on their spreading influence - we use the proposed methods;
- we compare the methods according to their ability to select a certain fraction of total nodes as "seed nodeset" with maximized total

[9] There also exists a weighted SIR model targeted for weighted networks (Sun, Liu, Zhang, & Zhang, 2014) and many recently developed complex epidemic models with large parameter settings that mimic the real spreading of various diseases (Arefin et al., 2019; Kabir & Tanimoto, 2020)

Table 2
Node measures allotted to the toy network by various ranking methods such as degree centrality (k), betweenness centrality (bc), closeness centrality (cc), k-shell (ks), mixed degree decomposition (mdd), improved method (θ), iterative neighborhood coreness (nc_+), iterative gravity method (g_+), neighbors degree method (k_{sum}) and four variants of the proposed method (one that uses k-shell (x^{ks}) and other using M-shell (x^{mdd})) along with their iterative versions (x^{ks}_+, x^{mdd}_+) respectively. The spreading score obtained with the SIR model simulations for each node are shown in the second column (denoted SIR).

Node	SIR	k	bc	cc	ks	mdd	θ	nc_+	g_+	k_{sum}	x^{ks}	x^{ks}_+	x^{mdd}	x^{mdd}_+
a	3.894	2	1.3	0.317	2	2	22	17	50.1	9	27.6	77.83	27.56	141.82
b	5.192	6	58	0.432	2	4.5	14	32	111.2	14	40.8	205.81	91.87	281.35
c	4.873	3	7	0.38	2	2.7	14	26	94.8	11	37	171.72	49.95	247.26
d	5.129	3	1.3	0.404	2	2.7	14	32	107.7	15	43.8	178.17	59.17	302.36
e	6.838	6	85.3	0.543	3	4.8	3	63	229.3	28	100.33	387.08	160.53	595.69
f	5.69	4	12.8	0.463	3	3.7	3	48	185.7	20	81.8	315.44	100.82	485.84
g	6.738	6	56.3	0.528	3	4.8	3	63	229.8	27	97.7	388.08	156.27	583.62
h	6.497	6	47.6	0.514	3	4.8	3	56	202.6	22	86	350.42	137.6	518.28
i	2.826	1	0	0.345	1	1	21	13	52.3	6	13.9	86	13.92	137.6
j	5.393	4	27.4	0.404	2	3.4	12	34	129.7	15	42.8	219.11	72.82	329.31
k	5.062	4	21.1	0.396	2	3.1	14	29	104	12	38.3	177.14	59.42	249.33
l	2.604	1	0	0.292	1	1	30	9	26	4	9.6	42.83	9.61	72.82
m	4.137	2	1	0.358	2	2	14	19	71	8	31.4	120.08	31.44	160.24
n	2.564	1	0	0.288	1	1	33	8	24.5	4	9.2	38.33	9.19	59.42
o	4.72	3	6.5	0.388	2	2.7	14	27	96.6	12	38.8	161.83	52.43	241.52
p	3.748	2	1	0.317	2	2	20	16	49	7	25.8	81.67	25.83	125.24
q	2.963	1	0	0.345	1	1	21	13	52.3	6	13.9	86	13.92	137.6
r	2.761	1	0	0.306	1	1	33	10	27.7	6	11.4	40.83	11.36	91.87
s	2.825	1	0	0.306	1	1	33	10	27.7	6	11.4	40.83	11.36	91.87
t	2.806	1	0	0.306	1	1	33	10	27.7	6	11.4	40.83	11.36	91.87

spreading influence - we use the improved seed node selection technique using a minimum distance threshold.

The Kendall's τ value is a standard approach to compare two ranked lists (Knight, 1966). To compare the ranking generated by various node ranking methods, we compute the τ value between the benchmark SIR model generated ranking list and the heuristic generated ranking list. Kendall (1945) proposed this metric to compute the similarity between two ranked list in terms of rank correlation based on the relative positions of ranked items, as shown in Eq. (15).

$$\tau = \frac{2 * (N_1 - N_2)}{N * (N-1)} \tag{15}$$

where N_1 and N_2 are the number of concordant and discordant pairs respectively and, N is the total number of items in the list. Suppose, (a_1, b_1) and (a_2, b_2) be a pair of joint rank where the ranks a_1, b_1 belong to the first list and ranks a_2, b_2 belong to the second list. If $a_1 > a_2$ and $b_1 > b_2$ or $a_1 < a_2$ and $b_1 < b_2$, then rank (a_1, b_1) and (a_2, b_2) are concordant. If $a_1 > a_2$ and $b_1 < b_2$ or $a_1 < a_2$ and $b_1 > b_2$, they are discordant. If $a_1 = a_2$ or $b_1 = b_2$, the rank is neither concordant nor discordant. An ideal value of $\tau = 1$signifies highest similarity i.e., both the ranked lists are the same while any lower value points to a dissimilarity. If one list is reversed and Kendall's τ is computed with the unaltered list, then τ value becomes -1.

In the case of the single super spreader identification[10], as we compute the spreadability of a selected node with the SIR epidemic model, the process becomes dependent on the infection probability parameter (β) used in the SIR model settings. Nodes are ranked based on their estimated spreadability by the SIR model. The spreadability of a node v is $\sigma(v)$, defined as the number of recovered nodes once the infection dies.

Due to the infection propagation's stochastic nature, the spreadability varies a little every time, even if we start the infection with the same node. Hence, to enhance the accuracy and minimize any statistical error, we repeat the simulation with the same initial node for T times to obtain $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\sigma(v^t)$. Here, $\sigma(v^t)$ is the spreading influence due to node v at the t^{th} run of the simulation. We have set $T = 1000$ in the present study. A similar approach to minimize the statistical error has been used in many other studies (Wang et al., 2020; Liu et al., 2013).

To counter the varying effect of infection probability on different ranking methods, we compute the average Kendall's rank correlation (see Eq. (16)) while varying the infection probability parameter values from β_{th} to twice of that with a 10% increase on every run.

$$\tau_{avg} = \frac{1}{B}\sum_{b=1}^{B}\tau(\beta = \beta_{th} + \delta * b) \tag{16}$$

Another concern raised in many studies is that some methods may select the top node "luckily" with the maximum SIR spreading influence but fail to rank subsequent nodes. In other words, a single node's spreading influence has large fluctuations. To overcome this possibility, we compare the average spreading influence per node (σ_{avg}) for some selected number (fraction of total network size, N) of nodes (see Pei et al., 2014 & Basaras et al., Basaras, Katsaros, & Tassiulas, 2013).

The σ_{avg} metric gives the average spreading influence per node in terms of the total number of infected nodes after the infection dies out in the SIR model. It starts with an initial seed nodeset that contains the top p fraction of nodes generated by different ranking heuristics. It then calculates the total number of infected nodes by adding the individual node's influences (i.e., the number of infected nodes when that node has started the infection). The seed nodes are considered as a fraction/percentage of the network size. We compute the average spreading influence as follows.

1. we first identify the seed nodeset using different ranking heuristics.
2. Then, for each node, simulate the SIR model 1000 times to get its spreading influence in terms of the average number of infected nodes.
3. Finally, all such individual influences of seed nodeset are summed up and averaged to find the average influence per node (σ_{avg}) (see Eq. (17)) of some ranking method.

$$\sigma_{avg} = \frac{1}{p * N}\sum_{u \in N_p}\overline{\sigma}(u) \tag{17}$$

where $\overline{\sigma}$ is the SIR simulated spreading influence of a single node, p is the top fraction of the total network considered during the above

[10] The other case is the multiple seed nodeset identification.

Table 3
Ranked nodes of the toy network by different methods such as degree centrality (k), betweenness centrality (bc), closeness centrality (cc), k-shell (ks), mixed degree decomposition (mdd), improved method (θ), iterative neighborhood coreness (nc_+), iterative gravity method (g_+), neighbors degree method (k_{sum}) and four variants of the proposed method (one that uses k-shell (x^{ks}) and other using M-shell (x^{mdd})) along with their iterative versions (x^{ks}_+, x^{mdd}_+) respectively. Ranks obtained with SIR model simulations are shown in the second column.

Rank	SIR	k	bc	cc	ks	mdd	θ	nc_+	g_+	k_{sum}	x^{ks}	x^{ks}_+	x^{mdd}	x^{mdd}_+
1	e	b,e,g,h	e	e	e,f,g,h	e,g,h	e,f,g,h	e,g	g	e	e	g	e	e
2	g	j,k,f	b	g	a,j,k,m,o,p,b,c,d	b	j	h	e	g	g	e	g	g
3	h	o,c,d	g	h	l,n,p,q,r,s,i	f	k,m,o,b,c,d	f	h	h	h	h	h	h
4	f	a,m,p	h	f	–	j	p	j	f	f	f	f	f	f
5	b	l,n,q,r,s,t,i	j	b	–	k	q,i	b,d	j	d,j	d	j	b	j
6	j	–	k	j,d	–	o,c,d	a	k	b	b	j	b	j	d
7	k	–	f	k	–	a,m,p	l	o	d	k,o	b	d	k	b
8	d	–	c	o	–	l,n,q,r,s,t,i	n,r,s,t	c	k	c	o	k	d	k
9	o	–	o	c	–	–	–	m	o	a	k	c	o	c
10	c	–	a,d	m	–	–	–	a	c	m	c	o	c	o
11	m	–	m,p	q,i	–	–	–	p	m	p	m	m	m	m
12	p	–	l,n,q,r,s,t,i	a,p	–	–	–	q,i	q,i	i,q,r,s,t	a	i,q	a	a
13	a	–	–	r,s,t	–	–	–	r,s,t	a	l,n	p	p	p	i,q
14	q	–	–	l	–	–	–	l	p	–	i,q	a	i,q	p
15	i	–	–	n	–	–	–	n	r,s,t	–	r,s,t	l	r,s,t	r,s,t
16	t	–	–	–	–	–	–	–	l	–	l	r,s,t	l	l
17	s	–	–	–	–	–	–	–	n	–	n	n	n	n
18	r	–	–	–	–	–	–	–	–	–	–	–	–	–
19	l	–	–	–	–	–	–	–	–	–	–	–	–	–
20	n	–	–	–	–	–	–	–	–	–	–	–	–	–

computation, and N_p is the set of seed nodes that consists of top $p * N$ nodes.

Finally, to understand the effectiveness and quality of the ranking list, we compute and compare the monotonicity (see Bae & Kim, 2014). Monotonicity is a measure of the uniqueness of ranks assigned to each element. If some ranking list has many nodes with the same rank, then its monotonicity is less. In contrast, if all the elements in a ranking list have unique ranks, then the ranking list is a perfectly monotonous with a monotonicity value of 1. This metric is formally defined as shown in Eq. (18).

$$M(R) = \left[1 - \frac{\sum_{r \in R} n_r(n_r - 1)}{N(N-1)}\right]^2 \tag{18}$$

where n_r is the total number of ties, i.e., elements with the same rank (r), and N denotes the size of the ranking list R.

7. Results of the evaluation and discussion

7.1. Experiment 1

Here we attempt to compute the measures of importance by different node ranking methods on our toy network and then based on those measures we shall rank the nodes. The results of Experiment 1 are presented in Tables 2 and 3. On the toy network $\beta_{th} = 0.322$ and the infection probability $\beta = 0.35$. $\lambda = 0.7$ for the mdd method.

The spreading influence of the individual nodes measured with SIR simulation (see 'SIR') is shown in the second column of Table 2. We assume these node values as the benchmark values (proxy of actual spreading measure), and the ranking of nodes based on this SIR assigned node values as the standard. The individual values assigned to the nodes by the different methods are not important, but the relative ordering of values are. For example, the SIR model assigned the highest value to node 'e'(=6.838) and the x^{mdd} method also assigned the highest value to the same node 'e'(=160.53) despite the actual values, the x^{mdd} method has correctly ranked node 'e'. We notice that degree centrality and k-shell methods have assigned the same measure values to many nodes, and these nodes with the same measure value are placed in the same rank. For example, there are only three distinct k-shell measures, namely, 1, 2, and 3; degree centrality assigns five different measures to the nodes, namely, 1, 2, 3, 4, and 6. However, most of the recent hybrid methods assign unique measure values to most of the nodes. That allows them to rank the nodes more uniquely. If we observe the last ranks assigned to the toy network nodes, then it is clear that the recent hybrid methods and our proposed methods reach lower ranks while classical methods have only a few ranks and many nodes in each rank.

7.2. Experiment 2

Here we apply our proposed methods along with other state-of-the-art methods on real networks to compare their relative performance in ranking the nodes based on their influentiality. The results of Experiment 2 are presented in Table 4. We observe that the average rank correlation values are almost 90% for most of the networks with the proposed x^{ks} and x^{mdd} methods. The iterative x^{ks}_+ method also performs well in most cases, but x^{mdd}_+ fails in some networks. If a direct comparison of x^{ks} with k-shell is performed, we notice that our proposed method improves over the classical k-shell. Similarly, a comparison between x^{mdd} with the classical mdd reveals an improvement in the correlation metric. Among other state-of-the-art, k_{sum} is also performing well, while the improved method (θ) yields competitive results on some networks.

Fig. 3 presents the average spreading influence per node for different fractions (p) of the initially infected seed nodes on various networks.

It is clear from the plots that:

- all the proposed methods yield a consistent higher average in the Netscience, Euroroad, Powergrid and the USAirline network: x^{mdd}_+ has the highest average spread with any fraction of seed nodes beyond 1%.
- In the C.elegans and Pollblogs networks almost all methods except k-shell generate similar average spread while x^{ks}_+ has a little edge; the k-shell has the lowest spread.
- For the ODLIS, Advogato and Hamsterster networks, the proposed methods perform at par with the others.
- We also observe from the plots that k_{sum}, and g_+ generate a similar spreading influence when the influence overlap is considered.

Table 4
Average Kendall's rank correlation τ of various ranking schemes such as degree centrality (k), k-shell (ks), mixed degree decomposition (mdd), improved method (θ), iterative neighborhood coreness (nc_+), iterative gravity method (g_+), neighbors degree method (k_{sum}) and four variants of the proposed method (one that uses k-shell (x^{ks}) and other using M-shell (x^{mdd})) along with their iterative versions (x_+^{ks}, x_+^{mdd}) respectively. The $\lambda = 0.7$ is taken while simulating mdd method. β_{exp} is changed from β_{th} to 2* β_{th} with 10% increase on every execution. Results are averaged over 10 execution for every β_{exp}. A higher and closer to 1 τ value signify better performance of the heuristics in terms of ranking nodes

Networks	$\tau(k)$	$\tau(ks)$	$\tau(mdd)$	$\tau(\theta)$	$\tau(nc_+)$	$\tau(g_+)$	$\tau(k_{sum})$	$\tau(x^{ks})$	$\tau(x_+^{ks})$	$\tau(x^{mdd})$	$\tau(x_+^{mdd})$
NetScience	0.6809	0.6509	0.6819	0.6514	0.8333	0.8695	0.8539	0.8780	0.8927	0.8869	0.9060
C.elegans	0.7659	0.7806	0.7859	0.8256	0.8795	0.9029	0.8515	0.8671	0.9076	0.8659	0.7870
Pollblog	0.8787	0.8953	0.8846	0.85475	0.9279	0.9398	0.9048	0.9121	0.9389	0.9035	0.8992
Advogato	0.7752	0.7898	0.7775	0.8657	0.8750	0.8822	0.8612	0.8424	0.8817	0.8333	0.8366
Zachary	0.7082	0.6483	0.7224	0.6785	0.9074	0.9090	0.8971	0.7625	0.9213	0.8102	0.8744
Powergrid	0.4317	0.3991	0.4530	0.3502	0.6472	0.7164	0.6277	0.6714	0.7604	0.6526	0.7475
Hamsterster	0.7345	0.7866	0.7459	0.8831	0.9095	0.9272	0.8787	0.8696	0.9273	0.8445	0.8786
US-Airport	0.6040	0.6357	0.6104	0.8712	0.8864	0.8909	0.8139	0.7427	0.8912	0.7204	0.8733
US-Airline-97	0.7632	0.7969	0.7722	0.9282	0.9445	0.9525	0.9336	0.8841	0.9587	0.8676	0.9143
Jazz	0.8506	0.8150	0.8762	0.7934	0.9455	0.9265	0.9389	0.8922	0.9389	0.9124	0.9497
EURO-Road	0.4761	0.5591	0.5521	0.5121	0.7026	0.7589	0.6608	0.7484	0.7879	0.7094	0.7836
Macaques	0.9461	0.4588	0.9335	0.8945	0.9435	0.9480	0.9437	0.9437	0.9490	0.9481	0.9414
PGP	0.4540	0.4675	0.4571	0.6855	0.7420	0.8175	0.6457	0.7876	0.8197	0.7819	0.7864
DBLP-cite	0.6927	0.7194	0.6977	0.8095	0.8349	0.8530	0.7907	0.8290	0.8512	0.8185	0.8009
Odlis	0.6776	0.7216	0.6824	0.7755	0.8482	0.8261	0.8365	0.8088	0.8227	0.7786	0.7957
Email-EU-core	0.8801	0.8923	0.8885	0.9066	0.9350	0.9487	0.9322	0.9227	0.9504	0.9141	0.0576
CA-CSPhd	0.2660	0.2311	0.2754	-0.1671	0.7264	0.8396	0.6918	0.8242	0.8433	0.8021	0.8347
CA-HepTh	0.5692	0.5998	0.5776	0.7173	0.8242	0.8977	0.7869	0.8723	0.9063	0.8565	0.8895
CA-GRQC	0.6029949	0.6033341	0.6077032	0.7806882	0.8246283	0.8727	0.7940	0.8579	0.8740	0.8512	0.8591
CA-Condmat	0.5509	0.5674	0.5620	0.7844	0.7723	0.8761	0.7431	0.8421	0.8889	0.8236	0.8709
CA-AstroPh	0.7173	0.7317	0.7265	0.7927	0.8910	0.9139	0.8745	0.8747	0.9142	0.8600	0.90839
Facebook-ego	0.6283	0.6741	0.6449	0.7329	0.8586	0.8711	0.7719	0.7999	0.8657	0.7806	0.8342
Brightkite	0.56884	0.59507	0.5724	0.69815	0.81582	0.82692	0.78364	0.76996	0.82458	0.75943	0.80194

- K-shell and the improved method (θ) do not perform well in this setup. The average spreading is the lowest with the k-shell and slightly better than k-shell with the improved method. We shall see in Experiment 4 results that show that these two methods yield a maximum improvement in total spread with our 'improved seed selection technique'.

The monotonicity values for different heuristics for the benchmark networks generated ranking lists that are presented in Table 5. We can see that almost all recent methods generate a ranking that is more than 90% monotonous. The degree, K-shell and mdd method have comparatively lower monotonicity values while g_+,nc_+, and k_{sum} and all proposed variants give a monotonicity value of nearly 90%. Only the improved method (θ) has a moderate monotonicity value (80%) on most of the networks. This means that the ranking list generated by all these methods except the improved method and k-shell method are almost identical to the SIR model generated ranks.

7.3. Experiment 3

Here we attempt to understand if a minimum distance between seed nodes improve the total spreading on k-shell. Fig. 4 depicts the results of the experiment. We depict the total spreading influence for three different fractions of the initial seed nodes: 5%, 10% and 15%. The k-shell using our improved nodeset selection technique with a minimum distance of 2 hop between seed nodes are marked as ks-2 in the plots. Similarly ks-3 is the k-shell with a minimum distance of 3 hop between seed nodes. When the seed nodes are selected consecutively from the ranked list, i.e. using the top-k strategy, the minimum distance between nodes are 1 hop and we marked that as ks-1. With all the networks except Advogato, ks-2 generates higher total spreading influence than ks-1 and ks-3. For Advogato, only with a small β up to 0.016, the top-k strategy (i.e. ks-1) yields better spread but with any $\beta \geqslant 0.016$ ks-2 is better. The ks-3 have lowest performance in all cases except with the Hamsterster network with 5% initial seed nodes.

The conclusion is that the performance improves with increasing the minimum distance between the seed nodes up to a distance of 2 hop, and then it starts to deteriorate with a minimum distance of 3-hop. We observe that a minimum distance of two hops (ks-2) yields better total spread in almost all the cases over a broad range of the infection probability and initial fraction of seed nodes. Hence we establish the fact that while K-shell is used in selecting a minimal seed nodeset for a maximum spreadability, a minimum distance of 2-hop between the seed nodes offers greater collective influence.

7.4. Experiment 4

Here we use the minimum distance that yielded best improvement in Experiment 3 (i.e., a minimum distance of 2 hop between the seed nodes while choosing top-k nodes from the ranked list) on all other state-of-the-art methods to observe if an improvement occurs with them too. The plots depicting total influence without overlap with a different fraction of nodes as initial nodeset are shown in Fig. 5. Based on the work of (Guo et al., 2020) we also used the infection probability β values as 1.5 times of β_{th}, while simulating the spread of a node using the SIR model, as presented in Table 1. We observe from the plots that all four variants of the proposed methods over-performed other state-of-the-art methods due to the improved seed selection.

The last ranks are plotted in Fig. 6. A larger rank of the last seed node signifies that seeds are more dispersed and topologically distributed throughout the network allowing for a greater reach to the information propagation. In the plots, it is clear that with the improved seed selection, the last ranks go to much deeper in the rank list allowing the spreading to reach to (topologically) less important regions of the network.

We have computed the improvement in the total spreading influence (without overlap) with different fractions of seed nodes maintaining a minimum distance of 2 hops as

$$\%\sigma_I = \frac{\sigma_{total}^{md=2} - \sigma_{total}}{\sigma_{total}} * 100\%$$

The percentage of the improvements in the total spread due to the

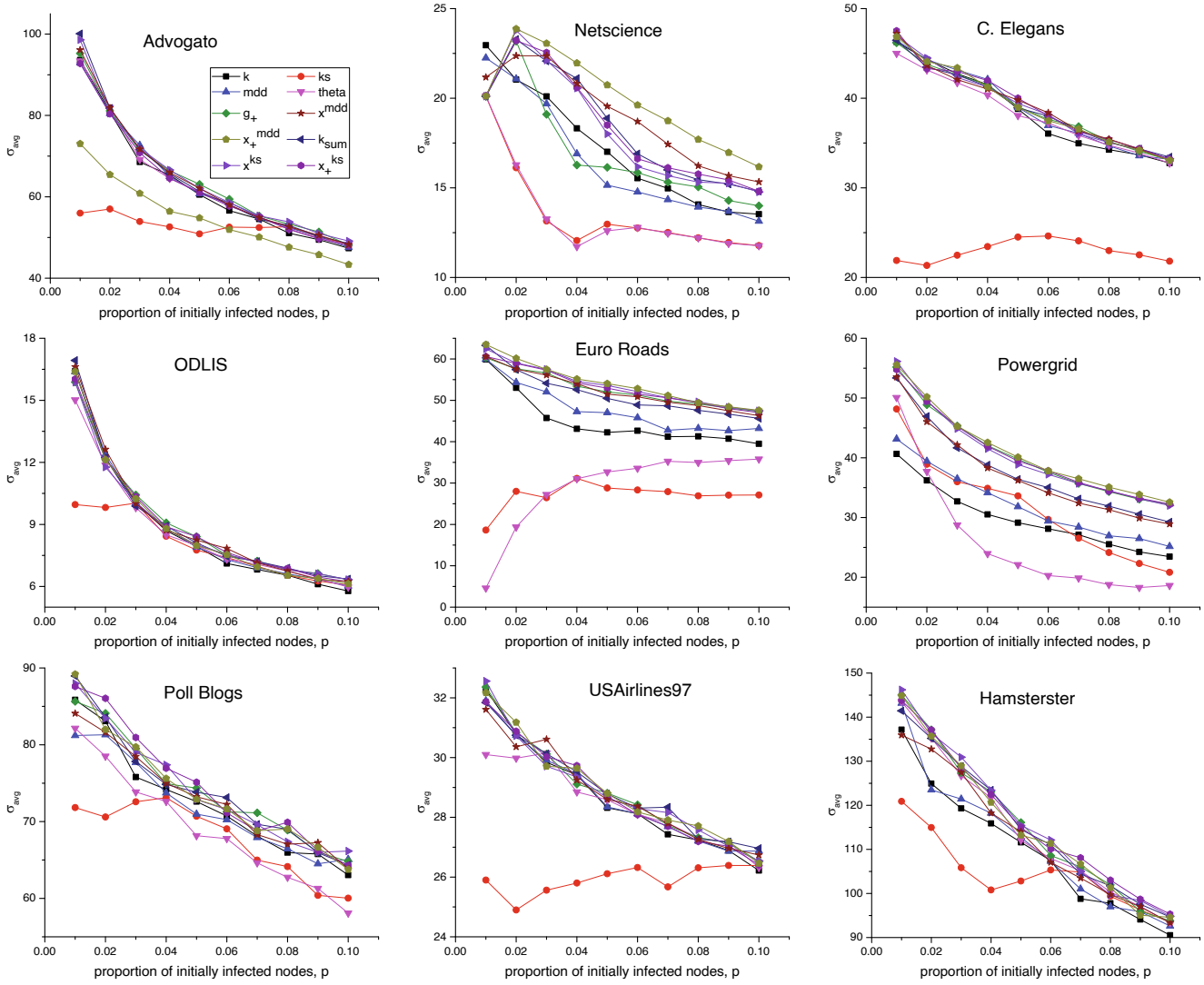


Fig. 3. Average spreadability per node with overlap(σ_{avg}) plotted for the different top fraction of nodes ($p * N$) by various node ranking methods like degree centrality (k), k-shell decomposition (ks), mixed degree decomposition (mdd), the improved method (θ), neighbors' degree method (k_{sum}), gravity plus (g_{+}) along with the proposed x^{ks}, x^{mdd} and their iterative versions x_{+}^{ks}, x_{+}^{mdd}. We have considered β values as shown in Table 1 that are nearly $1..5 * \beta_{th}$.

Table 5
Monotonicity values (M(*)) of ranked lists of real world networks generated by various node-ranking methods such as degree centrality (k), betweenness centrality (bc), closeness centrality (cc), k-shell (ks), mixed degree decomposition (mdd), improved method (θ), iterative neighborhood coreness (nc_+), iterative gravity method (g_+), neighbors degree method (k_{sum}) and four variants of the proposed method (one that uses k-shell (x^{ks}) and other using M-shell (x^{mdd})) along with their iterative versions (x_+^{ks}, x_+^{mdd}) respectively. A monotonicity value closer to 1 indicates more unique ranks of the nodes.

Network	M(k)	M(ks)	M(mdd)	M(θ)	M(nc_+)	M(g_+)	M(k_{sum})	M(x^{ks})	M(x_+^{ks})	M(x^{mdd})	M(x_+^{mdd})
NetScience	0.7069	0.6634	0.7397	0.6638	0.9125	0.9167	0.8966	0.9136	0.9162	0.9163	0.9171
C.elegans	0.9217	0.6094	0.9687	0.9892	0.9975	0.9977	0.9949	0.9975	0.9977	0.9977	0.9977
Pol-Blogs	0.9324	0.906	0.9443	0.9964	0.9992	0.9981	0.9986	0.9993	0.9993	0.9993	0.9993
Advogato	0.8445	0.8197	0.863	0.998	0.9985	0.9983	0.9969	0.9986	0.9986	0.9986	0.9986
Zachary	0.7079	0.4958	0.7536	0.8791	0.9472	0.9542	0.9403	0.9507	0.9542	0.9542	0.9542
Powergrid	0.5927	0.246	0.6928	0.9604	0.9419	0.9991	0.8866	0.9723	0.9921	0.9992	0.9998
Hamsterster	0.886	0.8489	0.911	0.996	0.9984	0.9989	0.9957	0.9987	0.9988	0.9988	0.9989
US-Airport	0.848	0.8278	0.8649	0.988	0.9982	0.9648	0.9969	0.9982	0.9983	0.9983	0.9983
US-Airline-97	0.8586	0.8114	0.8871	0.964	0.9945	0.9951	0.9924	0.9949	0.995	0.9951	0.9951
Jazz	0.9659	0.7944	0.9882	0.9345	0.9993	0.9993	0.9981	0.9993	0.9993	0.9993	0.9993
EURO-Road	0.4442	0.2129	0.6498	0.9882	0.9175	0.9962	0.84	0.9553	0.984	0.9955	0.9982
Macaques	0.9324	0.0447	0.8851	0.9263	1	1	0.9989	0.9989	1	0.9989	1
Odlis	0.8728	0.8151	0.9152	0.9969	0.9996	0.9999	0.9968	0.9999	0.9999	0.9999	0.9999
PGP	0.6193	0.4806	0.6678	0.9856	0.9851	0.9997	0.9552	0.9996	0.9997	0.9996	0.9997
DBLP-CITE	0.67	0.6427	0.6781	0.999	0.9983	0.9994	0.9909	0.9994	0.9994	0.9994	0.9994
CA-HepTh	0.7626	0.6741	0.8105	0.9552	0.9888	0.9936	0.9736	0.9935	0.9936	0.9935	0.9936
CA-GrQc	0.7459	0.663	0.7944	0.9577	0.9815	0.9871	0.9649	0.9871	0.9871	0.9871	0.9872
Wiki-Vote	0.7715	0.7625	0.7752	0.9976	0.9995	0.9996	0.9982	0.9996	0.9996	0.9996	0.9996
CA-CondMat	0.809	0.7409	0.8604	0.9776	0.9914	0.9948	0.9791	0.9948	0.9948	0.9948	0.9948
CA-AstroPh	0.9264	0.9083	0.9484	0.9946	0.9989	0.9991	0.9968	0.9991	0.9991	0.9991	0.9991
FB-NIPS	0.0043	0.004	0.0043	0.6804	0.687	0.687	0.687	0.687	0.687	0.687	0.687
FB-ego	0.9739	0.9419	0.9909	0.9859	0.9999	0.9999	0.9995	0.9999	0.9999	0.9999	0.9999
Brightkite	0.6782	0.6158	0.7127	0.9972	0.9979	0.9996	0.9857	0.9996	0.9996	0.9996	0.9996

improved seed selection over the top-k selection for the different networks are shown in Fig. 7. The average percentage improvement in the total spread over a range of initial fraction of seed nodes are plotted for the different networks and are shown in inset.

We observe that for the most of the heuristics, an improvement is apparent on most of networks. Except for the advogato and pollblogs networks, all other have a positive improvement. With the *Netscience* network, the average % of improvement goes beyond 200% with the k-shell and the improved method; also with the *Powergrid* network, the improvement reaches nearly 100% with the improved method. The *Advogato* network has the lowest improvements (2%-4%) including negative improvements (-2% with degree centrality method). The *Pollblogs* network has the lowest but positive improvements in the range of 2–3% for all methods except for the k-shell and the improved method in the range of 14–18%. In rest of the networks, improvement ranges between 10% - 20%. Effectively, many heuristics that perform moderately in multiple node selection using top-k nodes, become competitive using our minimum distance based improved seed node selection. We also observe that the maximum improvement is achieved with the k-shell and the *improved method* in all networks.

8. Conclusions and future work

Many single influential node identification methods that rank the nodes based on some heuristic measure, to proxy their actual spreadability, perform well compared with SIR simulated spreadability. The same methods are often employed to select a minimal seed nodeset in applications such as viral marketing, rumor spreading, or constraint budget problems. The primary objective in all such cases is to reach a greater portion of the population, starting with minimum nodes. Most of the existing studies overlook the problem of overlap in spreading influence while computing the total spreadability of the selected seed nodes. This paper presents two heuristics using K-shell and M-shell values as integral components along with their iterative variants for single super spreader identification and multiple seed node selection. The classical SIR model is also modified to compute the total spreadability of multiple seed nodes without overlap, i.e., without considering the same node multiple times even if it is under the spreading influence of multiple seed nodes. This work proposes a minimum distance-based improved seed node selection process, that improves the total spreading influence compared to the top-k seed node strategy, where top-k nodes are taken as initial seeds from the ranked list produced by single(most) influential node identification methods. It is observed that the proposed methods are at par with the state-of-the-art when applied to identify a single super spreader. It is also seen that when the different minimum distance technique is applied on seed node selection using classical K-shell, a minimum distance of 2 hops between the seed nodes offers the most improvement in total influence. All 4 variants (x^{ks}, x^{mdd} and their extended iterative versions x_+^{ks}, x_+^{mdd}, respectively) of the proposed methods over-performed the state-of-the-art when improved seed selection is employed over other state-of-the-art with top-k strategy. Finally, it is established that a minimum distance criterion during seed node selection improves the total spreading influence of most of the state-of-the-art single spreader identification heuristics when applied to multiple initial seed node selection.

The present work focused on unweighted networks where all the links/edges are of the same importance, but in practice it is not always the case. As future work we intend to extend the proposed heuristics and the 'improved seed nodeset selection technique' to weighted networks. The COVID-19 pandemic situation fuelled a large number of complex

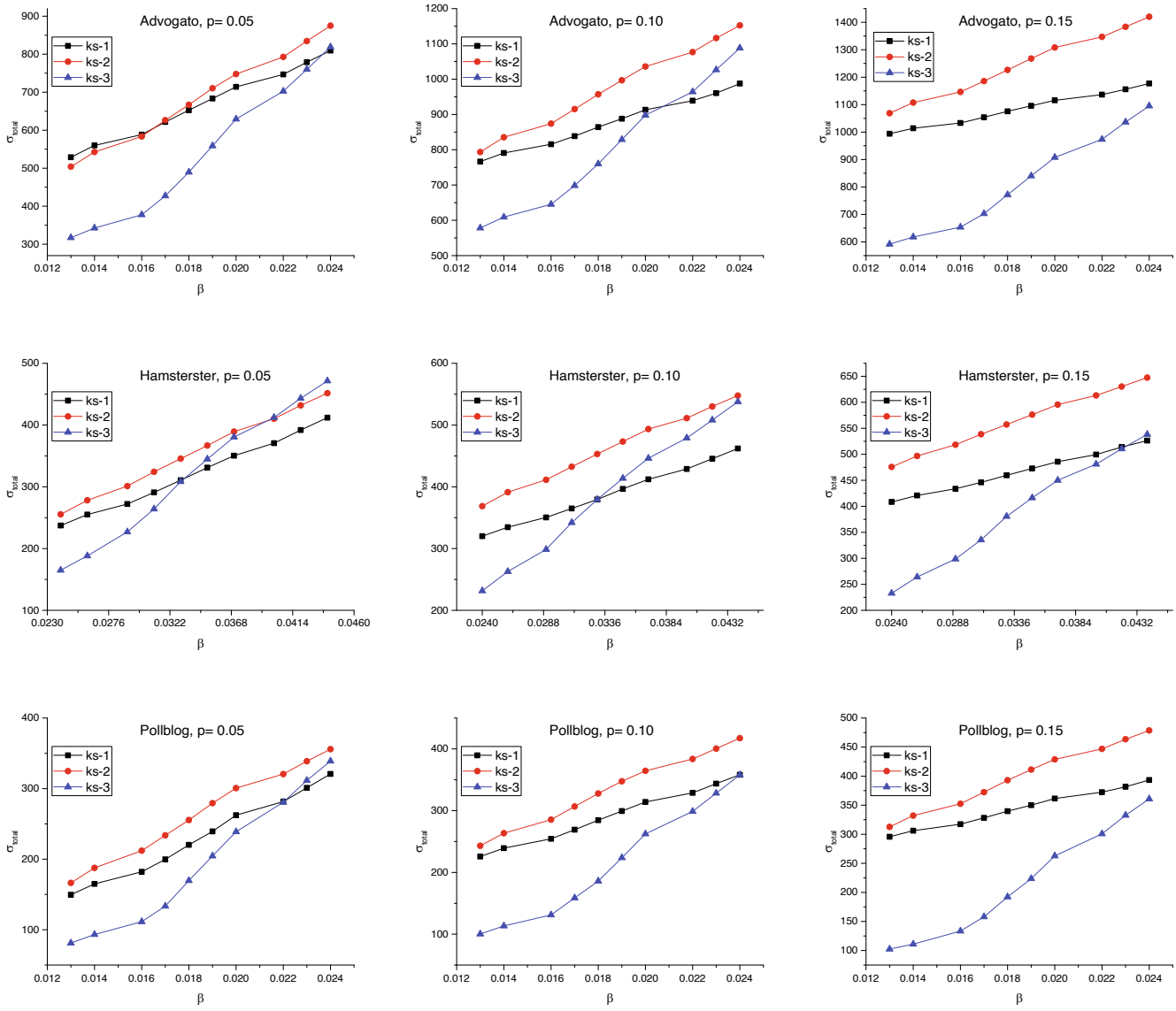


Fig. 4. Improvement on total spreading influence of K-shell due to different minimum distance constraint while selecting seed nodes with varying infection probability (β). Initial seed nodeset sizes considered are (p=) 5%, 10% and 15% of total nodes. We observe that a minimum distance of 2 hops between seed nodes (ks-2) yields higher total spreading influence (σ_{total}) in terms of total number of recovered nodes once the infection dies off.

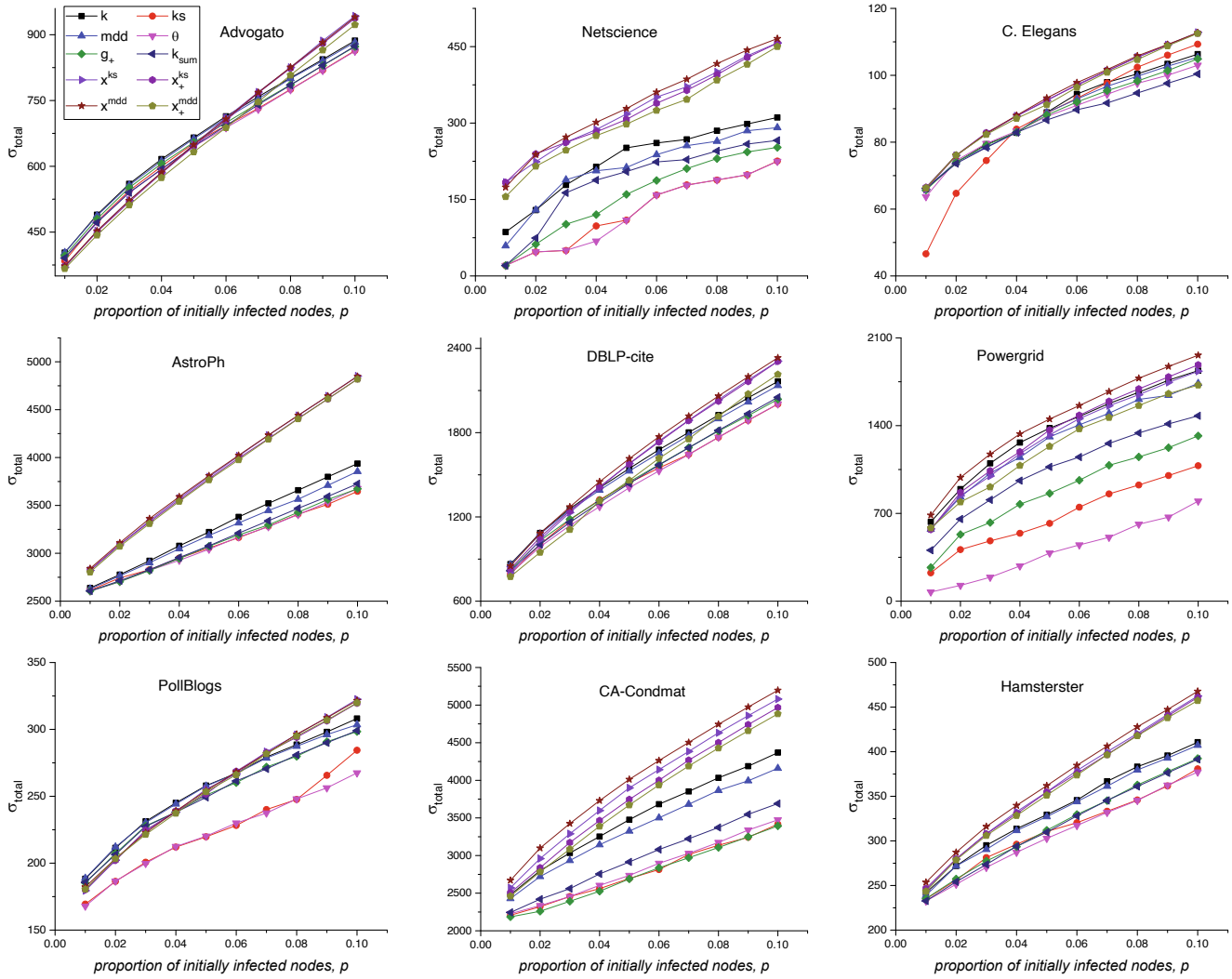


Fig. 5. Total spreading influence of a seed nodeset (σ_{total}) is plotted for different proportions of nodes ($p * N$) taken as initial seed nodes from the ranked list yielded by various node-ranking methods like degree(k), k-shell decomposition (ks), *mdd*, the *improved method*(θ), neighbors' degree (k_{sum}), gravity plus (g_{+}). The proposed x^{ks}, x^{mdd} and their iterative versions x_{+}^{ks}, x_{+}^{mdd} use the improved seed nodeset selection as per Algorithm 2 by maintaining a minimum 2 hop distance between seed nodes. We have considered infection probability (β) values as around 1.5 times of β_{th} as shown in Table 1 during individual node's spreadability computation using the SIR model simulation (Ref. Algorithm 3).

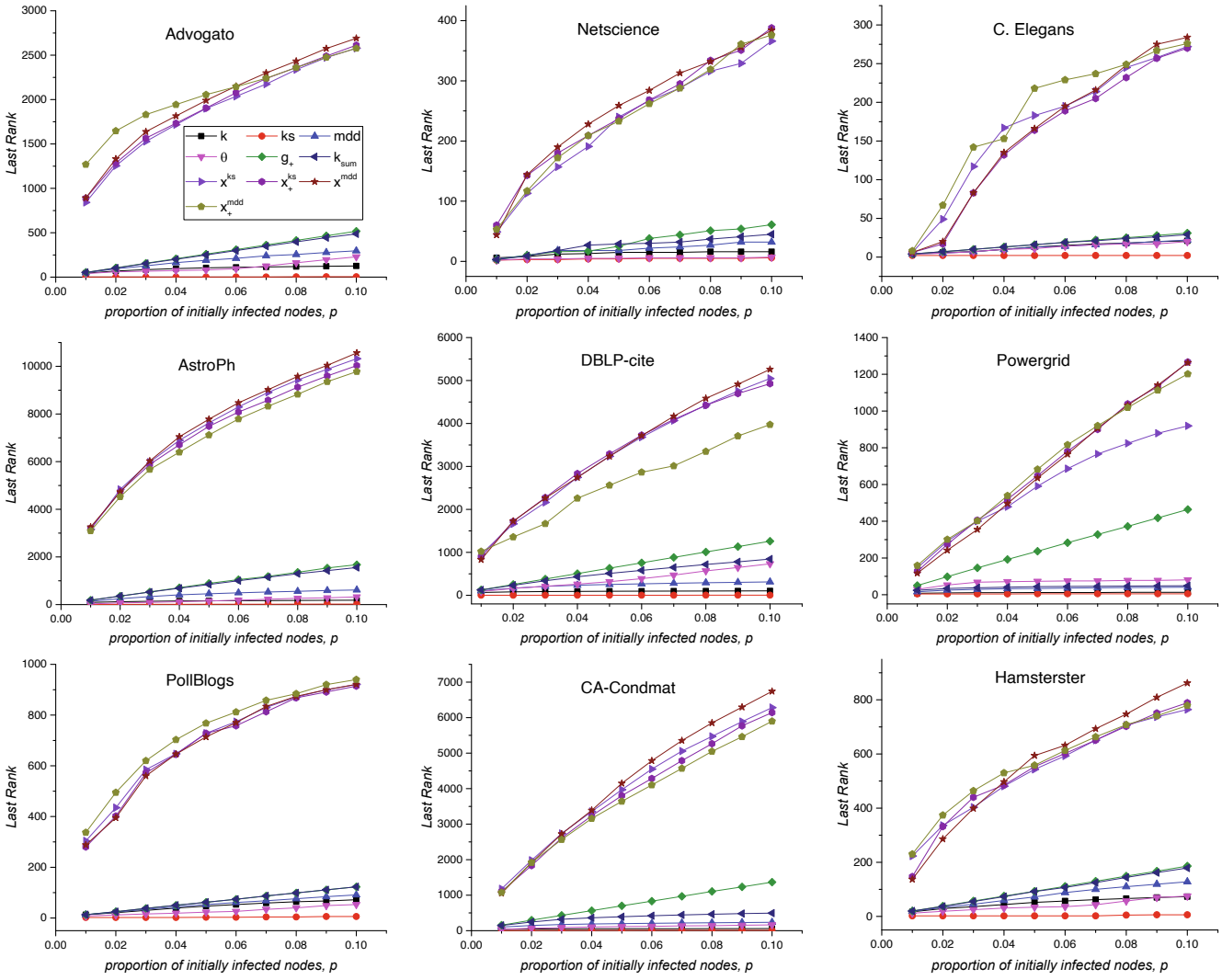
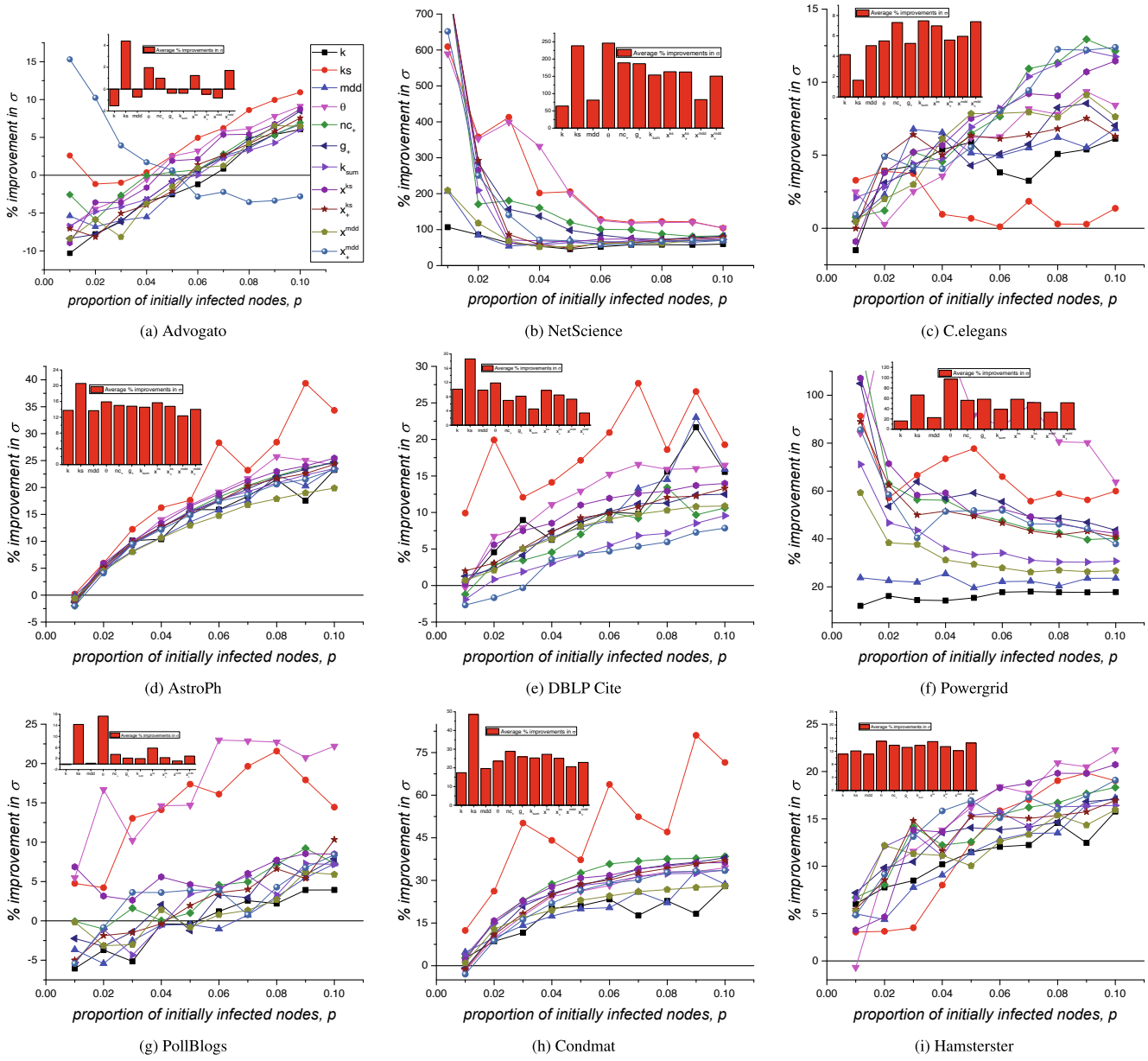


Fig. 6. The rank of the last seed nodes are plotted for different proportion of nodes ($p * N$) for various node ranking heuristics like degree (k), k-shell decomposition (ks), mixed degree decomposition (mdd), the *improved method* (θ), neighbors' degree (k_{sum}), gravity plus (g_+). The proposed x^{ks}, x^{mdd} and their iterative versions x^{ks}_+, x^{mdd}_+ use the improved seed nodeset selection as per Algorithm 2 by maintaining a minimum 2 hop distance between seed nodes. The more the last rank, the more dispersed the nodes get selected as seed nodes and yields more in total spread.



(a) Advogato (b) NetScience (c) C.elegans

(d) AstroPh (e) DBLP Cite (f) Powergrid

(g) PollBlogs (h) Condmat (i) Hamsterster

Fig. 7. Percentage improvement in the total spreading influence ($\%\sigma_I$) due to the adoption of the improved seed node selection technique for all the methods for different proportion of seed nodes ($p * N$) on real networks. The infection probability (β) is considered as 1.5 times of β_{th}. Inset shows the average percentage improvement over all p. It is seen that in most of the cases a positive improvement is achieved with almost all the methods.

epidemic models that closely map and predict future patterns of the infection spreading (Arefin et al., 2019; Kabir & Tanimoto, 2020). Another line of future work is to study the relative performance of different node ranking heuristics to understand if it changes when more complex recent epidemic models are considered as benchmark, instead of the standard SIR model.

CRediT authorship contribution statement

Giridhar Maji: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Animesh Dutta:** Conceptualization, Writing - review & editing, Resources. **Mariana Curado Malta:** Methodology, Writing - review & editing, Formal analysis. **Soumya Sen:** Supervision, Methodology, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been supported by portuguese national funds through FCT – Fundação para a Ciência e Tecnologia within the Projects' Scope: UIDB/05422/2020 and UID/CEC/00319/201

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36–43). ACM. doi: 10.1145/1134271.1134277, 2005.

Ahajjam, S., & Badir, H. (2018). Identification of influential spreaders in complex networks using HybridRank algorithm. *Scientific Reports, 8*(1), 1–10. https://doi.org/10.1038/s41598-018-30310-2

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2017). Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. *Physica A: Statistical Mechanics and its Applications, 468*, 278–288. https://doi.org/10.1016/j.physa.2016.11.002

Arefin, M. R., Masaki, T., Kabir, K. A., & Tanimoto, J. (2019). Interplay between cost and effectiveness in influenza vaccine uptake: a vaccination game approach. *Proceedings of the Royal Society A, 475*(2232), 20190608. https://doi.org/10.1098/rspa.2019.0608

Bae, J., & Kim, S. (2014). Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A: Statistical Mechanics and its Applications, 395*, 549–559. https://doi.org/10.1016/j.physa.2013.10.047

Banerjee, S., Jenamani, M., & Pratihar, D. K. (2019). ComBIM: A community-based solution approach for the Budgeted Influence Maximization Problem. *Expert Systems with Applications, 125*, 1–13. https://doi.org/10.1016/j.eswa.2019.01.070

Basaras, P., Katsaros, D., & Tassiulas, L. (2013). Detecting influential spreaders in complex, dynamic networks. *Computer, 46*(4), 24–29. https://doi.org/10.1109/MC.2013.75

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 3* (pp. 361–362).

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections, 21*(2), 47–57.

Batagelj, V., & Zaveršnik, M. (2011). Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification, 5*(2), 129–145. https://doi.org/10.1007/s11634-010-0079-y

Bhat, N., Aggarwal, N., & Kumar, S. (2020). Identification of Influential Spreaders in Social Networks using Improved Hybrid Rank Method. *Procedia Computer Science, 171*, 662–671. https://doi.org/10.1016/j.procs.2020.04.072

Boguná, M., Pastor-Satorras, R., Díaz-Guilera, A., & Arenas, A. (2004). Models of social networks based on social distance attachment. *Physical Review E, 70*(5). https://doi.org/10.1103/PhysRevE.70.056122, 056122.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*(1), 113–120. https://doi.org/10.1080/0022250X.1972.9989806

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Cao, T., Wu, X., Wang, S., & Hu, X. (2011). Maximizing influence spread in modular social networks by optimal resource allocation. *Expert Systems with Applications, 38*(10), 13128–13135. https://doi.org/10.1016/j.eswa.2011.04.119

Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., & Shir, E. (2007). A model of Internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences, 104*(27), 11150–11154. https://doi.org/10.1073/pnas.0701175104

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and Its Applications, 391*(4), 1777–1787. https://doi.org/10.1016/j.physa.2011.09.017

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks, in. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1082–1090).

Da Silva, L. N., Malacarne, A., e Silva, J. W. S., Kirst, F. V., De-Bortoli, R., et al. (2018). The scientific collaboration networks in university management in Brazil, *Creative Education, 9* (09), 1469. doi: 10.4236/ce.2018.99109.

Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E, 72*(2). https://doi.org/10.1103/PhysRevE.72.027104, 027104.

Ferguson, R. (2008). Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing. *Journal of Consumer Marketing, 25*(3), 179–182. https://doi.org/10.1108/07363760810870671

Franchi, E., Poggi, A., & Tomaiuolo, M. (2020). Social media for online collaboration in firms and organizations. In *Information Diffusion Management and Knowledge Sharing: Breakthroughs in Research and Practice* (pp. 473–489). IGI Global. doi: 10.4018/978-1-7998-0417-8.ch023.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Gao, C., Lan, X., Zhang, X., & Deng, Y. (2013). A bio-inspired methodology of identifying influential nodes in complex networks. *PloS One, 8*(6). https://doi.org/10.1371/journal.pone.0066732. e66732.

Gao, L., Yu, S., Li, M., Shen, Z., & Gao, Z. (2019). Weighted h-index for Identifying Influential Spreaders. *Symmetry, 11*(10), 1263. https://doi.org/10.3390/sym11101263

Ghoshal, A. K., Das, N., & Das, S. (2019). Misinformation containment in OSNs leveraging community structure. In *2019 IEEE 10th international conference on awareness science and technology (iCAST)* (pp. 1–6). IEEE. doi: 10.1109/ICAwST.2019.8923277.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Gleiser, P. M., & Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems, 6*(04), 565–573. https://doi.org/10.1142/S0219525903001067

Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., & Arenas, A. (2003). Self-similar community structure in a network of human interactions. *Physical Review E, 68*(6). https://doi.org/10.1103/PhysRevE.68.065103, 065103.

Guo, C., Yang, L., Chen, X., Chen, D., Gao, H., & Ma, J. (2020). Influential nodes identification in complex networks via information entropy. *Entropy, 22*(2), 242. https://doi.org/10.3390/e22020242

Han, L., Zhou, Q., Tang, J., Yang, X., & Huang, H. (2021). Identifying Top-k influential nodes based on discrete particle swarm optimization with local neighborhood degree centrality. *IEEE Access, 9*, 21345–21356. https://doi.org/10.1109/ACCESS.2021.3056087

He, S., Peng, Y., & Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics, 101*(3), 1667–1680. https://doi.org/10.1007/s11071-020-05743-y

Hong, W., Qian, C., & Tang, K. (2020). Efficient minimum cost seed selection with theoretical guarantees for competitive influence maximization. *IEEE Transactions on Cybernetics*, 1–14. https://doi.org/10.1109/TCYB.2020.2966593

Hu, Z.-L., Liu, J.-G., Yang, G.-Y., & Ren, Z.-M. (2014). Effects of the distance among multiple spreaders on the spreading, *EPL (Europhysics Letters), 106*(1), 18002. doi: 10.1209/0295-5075/106/18002.

Kabir, K. A., & Tanimoto, J. (2020). Evolutionary game theory modelling to represent the behavioural dynamics of economic shutdowns and shield immunity in the COVID-19 pandemic. *Royal Society Open Science, 7*(9). https://doi.org/10.1098/rsos.201095, 201095.

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). ACM. doi: 10.1145/956750.956769.

Kempe, D., Kleinberg, J., & Tardos, É. (2015). Maximizing the spread of influence through a social network. *Theory of Computing, 11*(4), 105–147. https://doi.org/10.4086/toc.2015.v011a004

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika, 33*(3), 239–251. https://doi.org/10.2307/2332303

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893. https://doi.org/10.1038/nphys1746

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604–632. https://doi.org/10.1145/324133.324140

Kleinberg, J. (2007). Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic Game Theory, 24*, 613–632.

Knight, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association, 61*(314), 436–439. https://doi.org/10.1080/01621459.1966.10480879

Konect (2017). Hamsterster friendships network dataset – KONECT, url:http://konect.uni-koblenz.de/networks/petster-friendships-hamster.

Kumar, P., Verma, P., & Singh, A. (2018). A study of epidemic spreading and rumor spreading over complex networks. In *Towards extensible and adaptable methods in computing* (pp. 131–143). Springer. doi: 10.1007/978-981-13-2348-5_11.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 641–650). ACM. doi: 10.1145/1772690.1772756.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 2–43. https://doi.org/10.1145/1217299.1217301

Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval* (pp. 1–10). Springer. doi: 10.1007/3-540-45735-6_1.

Li, Z., Ren, T., Ma, X., Liu, S., Zhang, Y., & Zhou, T. (2019). Identifying influential spreaders by gravity model. *Scientific Reports, 9*(1), 1–7. https://doi.org/10.1038/s41598-019-44930-9

Li, H., Shang, Q., & Deng, Y. (2021). A generalized gravity model for influential spreaders identification in complex networks. *Chaos, Solitons & Fractals, 143*. https://doi.org/10.1016/j.chaos.2020.110456, 110456.

Liu, D., Jing, Y., Zhao, J., Wang, W., & Song, G. (2017). A fast and efficient algorithm for mining top-k nodes in complex networks. *Scientific Reports, 7*, 43330. https://doi.org/10.1038/srep43330

Liu, W., Li, Y., Chen, X., & He, J. (2020). Maximum likelihood-based influence maximization in social networks. *Applied Intelligence*, 1–16. https://doi.org/10.1007/s10489-020-01747-8

Liu, J.-G., Ren, Z.-M., & Guo, Q. (2013). Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications, 392*(18), 4154–4159. https://doi.org/10.1016/j.physa.2013.04.037

Liu, Y., Tang, M., Zhou, T., & Do, Y. (2015). Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Scientific Reports, 5*, 13172. https://doi.org/10.1038/srep13172

Liu, F., Wang, Z., & Deng, Y. (2020). GMM: A generalized mechanics model for identifying the importance of nodes in complex networks. *Knowledge-Based Systems, 193*. https://doi.org/10.1016/j.knosys.2019.105464, 105464.

Li, M., Zhang, R., Hu, R., Yang, F., Yao, Y., & Yuan, Y. (2018). Identifying and ranking influential spreaders in complex networks by combining a local-degree sum and the clustering coefficient. *International Journal of Modern Physics B, 32*(6), 1850118. https://doi.org/10.1142/S0217979218501187

Li, Q., Zhou, T., Lü, L., & Chen, D. (2014). Identifying influential spreaders by weighted LeaderRank. *Physica A: Statistical Mechanics and its Applications, 404*, 47–55. https://doi.org/10.1016/j.physa.2014.02.041

Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS One, 6*(6). https://doi.org/10.1371/journal.pone.0021202. e21202.

Lü, L., Zhou, T., Zhang, Q.-M., & Stanley, H. E. (2016). The H-index of a network node and its relation to degree and coreness. *Nature Communications, 7*, 10168. https://doi.org/10.1038/ncomms10168

Maji, G. (2020). Influential spreaders identification in complex networks with potential edge weight based k-shell degree neighborhood method. *Journal of Computational Science, 39*. https://doi.org/10.1016/j.jocs.2019.101055, 101055.

Maji, G., Mandal, S., & Sen, S. (2020). A systematic survey on influential spreaders identification in complex networks with a focus on K-shell based techniques. *Expert Systems with Applications, 161*. https://doi.org/10.1016/j.eswa.2020.113681, 113681.

Maji, G., Namtirtha, A., Dutta, A., & Malta, M. C. (2020). Influential spreaders identification in complex networks with improved k-shell hybrid method. *Expert Systems with Applications, 144*. https://doi.org/10.1016/j.eswa.2019.113092, 113092.

Ma, L., Ma, C., Zhang, H.-F., & Wang, B.-H. (2016). Identifying influential spreaders in complex networks based on gravity formula. *Physica A: Statistical Mechanics and its Applications, 451*, 205–212. https://doi.org/10.1016/j.physa.2015.12.162

Massa, P., Salvetti, M., & Tomasoni, D. (2009). Bowling alone and trust decline in social network sites. In *Proc. int. conf. dependable, autonomic and secure computing* (pp. 658–663). doi: 10.1109/DASC.2009.130.

McAuley, J. J., & Leskovec, J. (2012). Learning to discover social circles in ego networks. In *NIPS*, Vol. 2012 (pp. 548–56). Citeseer.

Moreno, Y., Pastor-Satorras, R., & Vespignani, A. (2002). Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 26*(4), 521–529. https://doi.org/10.1140/epjb/e20020122

Morone, F., & Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature, 524*(7563), 65–68. https://doi.org/10.1038/nature14604

Nagarajan, K., Muniyandi, M., Palani, B., & Sellappan, S. (2020). Social network analysis methods for exploring SARS-CoV-2 contact tracing data. *BMC Medical Research Methodology, 20*(1), 1–10. https://doi.org/10.1186/s12874-020-01119-3

Namtirtha, A., Dutta, A., & Dutta, B. (2018). Identifying influential spreaders in complex networks based on kshell hybrid method. *Physica A: Statistical Mechanics and its Applications, 499*, 310–324. https://doi.org/10.1016/j.physa.2018.02.016

Namtirtha, A., Dutta, A., & Dutta, B. (2020). Weighted kshell degree neighborhood: A new method for identifying the influential spreaders from a variety of complex network connectivity structures. *Expert Systems with Applications, 139*. https://doi.org/10.1016/j.eswa.2019.112859, 112859.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical Review E, 66*(1). https://doi.org/10.1103/PhysRevE.66.016128, 016128.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social Networks, 27*(1), 39–54. https://doi.org/10.1016/j.socnet.2004.11.009

Opsahl, T. (2013). Why anchorage is not (that) important: Binary ties and sample selection, online] http://toreopsahl. com/2011/08/12/why-anchorage-is-not-that-important-binary-tiesand-sample-selection (accessed September 2013).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web, Tech. Rep., Stanford InfoLab.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics, 87*(3), 925. https://doi.org/10.1103/RevModPhys.87.925

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters, 86*(14), 3200. https://doi.org/10.1103/PhysRevLett.86.3200

Pei, S., Muchnik, L., Andrade, J. S., Jr, Zheng, Z., & Makse, H. A. (2014). Searching for superspreaders of information in real-world social media. *Scientific Reports, 4*, 5547. https://doi.org/10.1038/srep05547

Pittel, B., Spencer, J., & Wormald, N. (1996). Sudden emergence of a giant k-core in a random graph. *Journal of Combinatorial Theory, Series B, 67*(1), 111–151. https://doi.org/10.1006/jctb.1996.0036

Plageras, A. P., Psannis, K. E., Stergiou, C., Wang, H., & Gupta, B. B. (2018). Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings. *Future Generation Computer Systems, 82*, 349–357. https://doi.org/10.1016/j.future.2017.09.082

Rath, M. (2019). Application and Impact of Social Network in Modern Society. In *Hidden link prediction in stochastic social networks* (pp. 30–49). IGI Global. doi: 10.4018/978-1-5225-9096-5.ch002.

Rehman, A. U., Jiang, A., Rehman, A., Paul, A., Sadiq, M. T., et al. (2020). Identification and role of opinion leaders in information diffusion for online discussion network. *Journal of Ambient Intelligence and Humanized Computing*, 1–13. https://doi.org/10.1007/s12652-019-01623-5

Reitz, J. M. (2002). *ODLIS: Online dictionary of library and information science*. Western Connecticut State University Libraries.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*(4), 581–603. https://doi.org/10.1007/BF02289527

Shang, Q., Zhang, B., Li, H., & Deng, Y. (2021). Identifying influential nodes: A new method based on network efficiency of edge weight updating. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 31*(3), 033120. doi: 10.1063/5.0033197.

Shao, Z., Liu, S., Zhao, Y., & Liu, Y. (2019). Identifying influential nodes in complex networks based on Neighbours and edges. *Peer-to-Peer Networking and Applications, 12*(6), 1528–1537. https://doi.org/10.1007/s12083-018-0681-x

Sheikhahmadi, A., & Zareie, A. (2020). Identifying influential spreaders using multi-objective artificial bee colony optimization. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106436, 106436.

Song, G., Zhou, X., Wang, Y., & Xie, K. (2014). Influence maximization on large-scale mobile social network: A divide-and-conquer method. *IEEE Transactions on Parallel and Distributed Systems, 26*(5), 1379–1392. https://doi.org/10.1109/TPDS.2014.2320515

Šubelj, L., & Bajec, M. (2011). Robust network community detection using balanced propagation. *European Physical Journal B, 81*(3), 353–362. https://doi.org/10.1140/epjb/e2011-10979-2

Sun, Y., Liu, C., Zhang, C.-X., & Zhang, Z.-K. (2014). Epidemic spreading on weighted complex networks. *Physics Letters A, 378*(7–8), 635–640. https://doi.org/10.1016/j.physleta.2014.01.004

Takahata, Y. (1991). *Diachronic changes in the dominance relations of adult female japanese monkeys of the Arashiyama B Group* (pp. 123–139). Albany: The Monkeys of Arashiyama. State University of New York Press.

Tang, M., Liu, Q., Ma, T., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Dhelaan, M. (2019). $\mathscr{X}$-lowest-influence overlapping nodes based community detection in complex networks. *IEEE Access, 7*, 109646–109661. https://doi.org/10.1109/ACCESS.2019.2930474

Tong, G. A., Li, S., Wu, W., & Du, D.-Z. (2016). Effector detection in social networks. *IEEE Transactions on Computational Social Systems, 3*(4), 151–163. https://doi.org/10.1109/TCSS.2016.2627811

Wang, Z., Bauch, C. T., Bhattacharyya, S., d'Onofrio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M., & Zhao, D. (2016). Statistical physics of vaccination. *Physics Reports, 664*, 1–113. https://doi.org/10.1016/j.physrep.2016.10.006

Wang, J., Hou, X., Li, K., & Ding, Y. (2017). A novel weight neighborhood centrality algorithm for identifying influential spreaders in complex networks. *Physica A: Statistical Mechanics and its Applications, 475*, 88–105. https://doi.org/10.1016/j.physa.2017.02.007

Wang, M., Li, W., Guo, Y., Peng, X., & Li, Y. (2020). Identifying influential spreaders in complex networks based on improved k-shell method. *Physica A: Statistical Mechanics and its Applications*. https://doi.org/10.1016/j.physa.2020.124229, 124229.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'Small-world' networks. *Nature, 393*(1), 440–442. https://doi.org/10.1038/30918

Wei, B., Liu, J., Wei, D., Gao, C., & Deng, Y. (2015). Weighted k-shell decomposition for complex networks based on potential edge weights. *Physica A: Statistical Mechanics and its Applications, 420*, 277–283. https://doi.org/10.1016/j.physa.2014.11.012

Wen, T., Pelusi, D., & Deng, Y. (2020). Vital spreaders identification in complex networks with multi-local dimension. *Knowledge-Based Systems, 195*. https://doi.org/10.1016/j.knosys.2020.105717, 105717.

Weskida, M., & Michalski, R. (2019). Finding influentials in social networks using evolutionary algorithm. *Journal of Computational Science, 31*, 77–85. https://doi.org/10.1016/j.jocs.2018.12.010

Yang, G., Benko, T. P., Cavaliere, M., Huang, J., & Perc, M. (2019). Identification of influential invaders in evolutionary populations. *Scientific Reports, 9*(1), 1–12. https://doi.org/10.1038/s41598-019-43853-9

Yang, G., Cavaliere, M., Zhu, C., & Perc, M. (2020). Ranking the invasions of cheaters in structured populations. *Scientific Reports, 10*(1), 1–13. https://doi.org/10.1038/s41598-020-59020-4

Yang, L., Li, Z., & Giua, A. (2020). Containment of rumor spread in complex social networks. *Information Sciences, 506*, 113–130. https://doi.org/10.1016/j.ins.2019.07.055

Yin, H., Benson, A. R., Leskovec, J., & Gleich, D. F. (2017). Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 555–564). ACM. doi: 10.1145/3097983.3098069.

Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research, 33*(4), 452–473. https://doi.org/10.1086/jar.33.4.362975

Zareie, A., & Sheikhahmadi, A. (2018). A hierarchical approach for influential node ranking in complex social networks. *Expert Systems with Applications, 93*, 200–211. https://doi.org/10.1016/j.eswa.2017.10.018

Zareie, A., & Sheikhahmadi, A. (2019). EHC: Extended H-index Centrality measure for identification of users' spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications, 514*, 141–155. https://doi.org/10.1016/j.physa.2018.09.064

Zareie, A., Sheikhahmadi, A., & Jalili, M. (2020). Identification of influential users in social network using gray wolf optimization algorithm. *Expert Systems with Applications, 142*. https://doi.org/10.1016/j.eswa.2019.112971, 112971.

Zeng, A., & Zhang, C.-J. (2013). Ranking spreaders by decomposing complex networks. *Physics Letters A, 377*(14), 1031–1035. https://doi.org/10.1016/j.physleta.2013.02.039

Zhang, J.-X., Chen, D.-B., Dong, Q., & Zhao, Z.-D. (2016). Identifying a set of influential spreaders in complex networks. *Scientific Reports, 6*, 27823. https://doi.org/10.1038/srep27823