Abstract
Over the last couple of decades, Social Networks have connected people on the web from across the globe and have become a crucial part of our daily life. These networks have also rapidly grown as platforms for propagating products, ideas, and opinions to target a wider audience. This calls for the need to find influential nodes in a network for a variety of reasons, including the curb of misinformation being spread across the networks, advertising products efficiently, finding prominent protein structures in biological networks, etc. In this paper, we propose Modified Community Diversity (MCD), a novel method for finding influential nodes in a network by exploiting community detection and a modified community diversity approach. We extend the concept of community diversity to a two-hop scenario. This helps us evaluate a node’s possible influence over a network more accurately and also avoids the selection of seed nodes with an overlapping scope of influence. Experimental results verify that MCD outperforms various other state-of-the-art approaches on eight datasets cumulatively across three performance metrics.

Keywords Community diversity · Community structure · Influential nodes · Independent cascade model · Social networks · Viral marketing

1 Introduction
Nowadays, Social networks have become an integral part of our daily lives and serve as an excellent medium for people to interact and propagate their views and ideas. Due to the large number of users available on various social networks across the globe, they act as expedient channels for advertisements and promotion of the products (Mislove et al., 2007; Garton
et al., 1997; Kumar et al., 2006). Over the years, advertisements over social networks, also
known as “Viral Marketing” have seen a significant boost due to their cost-effective nature
and ability to propagate faster and wider. This approach, also known as “word-of-mouth
advertising,” exploits people’s trust in a known or influential person’s suggestions over other
traditional means of advertisement (Ferguson, 2008; Zhou et al., 2019). Further, social net-
works have been used for various other uses, including building collaborations, friendships,
and other social or professional relations. Hence, the problem of finding nodes with wide
influence over a network is a prominent area of research in network analysis. This is fur-
ther required for various tasks like curbing rumors spread in social networks and identifying
prominent proteins in protein graphs and other graph-based networks.

The problem of finding influential nodes in a network can be delineated as the task of
identifying a set of \( k \) influential nodes in a network whose influence over the whole network
is maximum across any set of \( k \) nodes, where \( k \) is a small positive integral constant (Kempe
et al., 2003a; Liu et al., 2022). Each individual entity acts as a node in the network, whereas
the relationship between two nodes (friendship, follower-followee, co-authorship between
users, links between proteins, etc.) helps form edges in the network to connect the nodes.
For example, in social networks, the task entails maximizing the spread of an opinion by
communicating it across the maximum number of nodes possible. Formally, the problem of
finding the influential nodes in a network, modelled as a graph \( G(V, E) \) where \( V \) is the set
of all nodes in the network, and \( E \) is the set of edges connecting the nodes, can be defined as
shortlisting a small number of nodes (\( k \)) such that the spread of influence can be maximized
in the network. Mathematically, we can represent it as follows.

\[
S = \text{argmax } \sigma(S) \text{ where } S \in V, \|S\| = k, \|V\| \gg k
\]

Here, \( \text{argmax } \sigma(S) \) denotes the overall influence spread generated by a set \( S \) of \( k \)
number of nodes, and \( k \) is a small and positive constant with respect to the total number
of nodes in the network. In general, selecting influential nodes in a network consists of
two main components: choosing a set of \( k \) spreaders or seed nodes and then computing the
spread of influence across the network originating from the seed nodes utilizing a suitable
information propagation model. While selecting the set of seed nodes, we first ensure that
the number of seed nodes is significantly lesser than the number of nodes in the network to
use the resources available optimally. We also need to consider the influence or propagating
capability of each node in the network based on its locality and links within the network.
Owing to the presence of a community structuring system in social networks, involving
community-dependent features can be used to evaluate better the network structure, as well
as aid in accelerating the spread of the propaganda.

In this work, we propose the concept of Modified Community Diversity (MCD) for find-
ing influential nodes in a network. Our approach extends the idea of the previously proposed
Community based Spreaders Ranking algorithm (CSR) (Kumar et al., 2022) for two-hop
neighborhood information. The proposed work intends to enable better utilization of avail-
able seed nodes for information spread. We further propose the use of Extended Community
Diversity for gathering information about the neighborhood based upon a two-hop mech-
anism. This information is further used to avoid the overlapping influence of influential
nodes in a network, thus improving the efficacy of finding influential nodes in a network
and providing an optimized spread of information or influence. For further experimental
analysis, we compare the performance of our proposed approach with other state-of-the-art
approaches for influential node detection on eight datasets, seven of which are real-world
datasets. We use the Independent Cascade (IC) model (Goldenberg et al., 2001) to simulate
the influence propagation on all datasets. Our experimental evaluations over three performance metrics validate the superiority of our proposed approach over the other approaches across different domains and act as evidence of its efficacy and widespread applicability. The main contributions of our paper can be summarised as follows.

- We introduce Modified Community Diversity (MCD) method using community structure and extended community diversity to find influential nodes in social networks.
- The proposed work utilizes a two-hop concept which enables us to gather information about the neighborhood of a node better to judge its influence appropriately. This helps in avoiding the selection of influential nodes with overlapping influence.
- We verify the results of the proposed approach on eight network datasets against seven other state-of-the-art methods. The obtained results reveal that the proposed method outperforms different approaches cumulatively across three metrics. Statistical tests verify the superiority of MCD over competing approaches.

The breakdown of the remaining paper is as follows. Section 2 details the related works and the existing research gaps. Section 3 elaborates on our proposed methodology and algorithms. Section 4 consists of information about the various experimental details used in our paper, including the datasets used, the Information Propagation models, the performance metrics, and the implementation environment. Section 5 contains our experimental results, their analysis, and broad outcomes. Section 6 delineates the statistical tests we have done to verify the efficacy of MCD for influence maximization. Section 7 concludes our paper.

2 Related works

Due to the advent of the internet and the rise in the global usage of social networks, the task of finding influential spreaders in a network as a research problem has seen significant contributions from researchers. Researchers have proposed various approaches to find the most influential nodes in a network, to maximize the spread of information in a network (Banerjee et al., 2020; Li et al., 2018). These approaches can be predominantly categorized as diffusion process-based approaches, descriptive approaches, topological measures-based approaches, statistical and stochastic approaches, data mining-based approaches, and machine learning-based approaches (Mendonça et al., 2022; Kumar et al., 2023). Experimental evaluations have found topological measures-based approaches to be the most effective in identifying suitable spreader nodes. Topological measures-based approaches utilize predefined centrality measures for ranking nodes based on their efficacy in spreading information across a network. These centrality measures aid in selecting suitable spreader nodes with high influence on the other nodes in the network based on their specific topological properties. Centrality measures-based approaches generally consist of two main steps: generating a score for each node based on a defined centrality measure and selecting the top \( k \) nodes based on the score obtained for the measure.

One of the first approaches using centrality scores for finding influential nodes in a network was proposed by Freeman (1978). The approach used the Degree Centrality measure, a local centrality measure that considers the number of neighbors of a node to generate its score. Degree Centrality has linear time complexity and was effective and efficient for small networks. However, due to its reliance on the local immediate neighbors for a given node, the effectiveness of Degree Centrality drops for nodes with few but influential neighbors. To counter the issues faced with local centrality measures, especially Degree Centrality,
researchers proposed semi-local centrality measures like LC (Chen et al. 2012) and LSC (Gao et al., 2014), which go beyond the immediate scope of a node, but up to a limited range. While LC ranks the nodes using one-hop and two-hop nearest neighbors, LSC also considers the clustering coefficient between immediate neighbors.

These local centrality-based approaches do not consider a node’s global location and only use the information corresponding to a certain locality, thus reducing their efficacy in real-life large networks. To counter this, researchers have also used global centrality measures like Betweenness Centrality, Closeness Centrality, and K-shell Decomposition, among others, to Influence Maximization. Freeman (1977) defined Betweenness for a specific node as the ratio of the shortest path which passes through the node and the total number of possible shortest paths in the network. Okamoto et al. (2008) proposed Closeness for a node as the inverse of the sum of the shortest path to all other nodes, thus determining the average closeness of a node to the rest of the graph. Betweenness and Closeness suffer from high time complexity, as they require the calculation of the shortest path between every node pair of the network.

Some other global centrality measures successfully used for Influence Maximization are Eigenvector Centrality (Bonacich, 2007) and Pagerank Centrality measure (Brin & Page, 1998). They aid in ranking nodes based on their importance among the neighborhood nodes. According to these measures, the presence of important nodes around the central node makes it a viable spreader node. Another global centrality-based measure, GLR (Salavati et al., 2019), divides the graph into communities and then selects each community’s local cores and gateway nodes. The nodes are ranked as influential spreaders based on their shortest paths to the local cores and the gateway nodes. Researchers have also proposed greedy search-based approaches (Kempe et al., 2003b; Leskovec et al., 2007) for ranking spreader nodes. These approaches use the sub-modularity and monotony of the Independent Cascade (IC) model along with Monte Carlo simulations to enhance the overall performance of the approach. However, greedy approaches are generally inefficient due to many simulations, and they require a lot of time to compute.

Some recent approaches also consider the bridge nodes, which connect two different groups of nodes in a network, as influential nodes. Liu et al. (2016) proposed the DIL centrality measure, where the importance of lines and the degree value are used to highlight a bridge node. DIL also considers the number of connected triangles when determining the relevance of a link, with the bridge node taking the primary role in controlling information propagation. Berahmand et al. (2019) introduced the DCL centrality measure, which finds the bridge nodes using a node’s degree, clustering coefficient, and the relationship between its one-hop neighbors. Wen and Deng (2019) introduced LID, an approach where prominent influencers in complex networks are extracted using quasi-local information, i.e., local structural properties.

Recently, researchers have proposed various community structure-based approaches (Kumar et al., 2022; Chen et al., 2022) for spreader ranking, which were found to be more effective in identifying suitable spreader nodes in a network (Kumar et al., 2021, 2022; Chen et al., 2021). Some recent community-based algorithms used for Influence Maximization are IM-ELPR (Kumar et al., 2021), CFIN (Khomami et al., 2021), and CAOM (He et al., 2020). CAOM divides the whole network into communities and selects only the most significant ones. Certain candidate nodes are then selected from each community using the one-hop measure. Hereafter, seed nodes are selected from candidate nodes by eliminating those with influence overlap. CFIN follows a similar approach, but the candidate nodes are
selected using degree centrality, and the seed nodes are selected using the clustering coefficient measure. IM-ELPR uses node seeding, community detection, and label propagation for influence maximization. The seed nodes are first identified using an extended h-index, and then the communities are detected using label propagation. The smaller, related communities are further merged into larger communities, and then the top $k$ influential nodes are identified.

The approaches mentioned above utilized local structures to identify the bridge nodes. However, unlike these approaches, the recently proposed Community-based Spreaders Ranking algorithm (CSR) (Kumar et al., 2022) for influence maximization selects global bridge nodes by considering the community bridge nodes as the most influential spreaders in the network. The CSR algorithm utilized three measures to identify bridge nodes, i.e., community diversity, community modularity, and community density. However, CSR didn’t consider the overlapping influence of the influential nodes in a network. Our proposed approach (MCD) considers this factor and thus enables better utilization of available seed nodes. Further, CSR considers only one-hop neighbors to detect influential nodes. In contrast, MCD uses a two-hop mechanism, which provides a better overview of the node’s neighborhood provided to the algorithm. This leads to markedly superior results by our approach as compared to CSR.

3 Proposed methodology

In this section, we present our proposed methodology for finding influential nodes in social networks using community detection and a modified community diversity approach. First, we uncover the community structure in the input network using Leiden’s algorithm. We consider the bridge nodes in the network to play a vital role in spreading the information, and thus, we identify the bridge nodes in the network using the Community Diversity (CD) score. To incorporate the two-hop mechanism for gathering more information regarding a node’s surroundings, we calculate the Extended Community Diversity (ECD) score. Then we calculate the Modified Community Diversity score using ECD and CD to ensure the proper distribution of the information. We rank the nodes based on the MCD score and select the top $k$ nodes, where $k$ signifies the constraint on the number of nodes that act as influencers. Finally, we evaluate the influence of the selected $k$ seed nodes as influential nodes over the entire network by using various evaluation parameters. The overall proposed work can be understood using Fig. 1, which depicts the process flow for calculating the MCD score for each node. The various phases of the proposed work are discussed below.

3.1 Community detection

A community can be illustrated as a subset of a graph with nodes representing each member of the community (Yang et al., 2021; Chouchani & Abed, 2020; Ma et al., 2017). The nodes have similar characteristics and are connected to some of the other nodes by edges. These nodes can represent any entity in the world, and their meaning is dependent on the scope of the community and, to a major extent, the scope of the graph. Nodes present specifically in a community are expected to have meaningful and substantial connections with each other in comparison to the rest of the graph, thus ensuring a better spread of information within the community as compared to outside the community. In our proposed algorithm, we use Leiden’s algorithm (Traag et al., 2019) to identify the communities in a network. Leiden’s
algorithm is based on greedy optimization techniques, which results in efficient processing and quicker convergence. Leiden’s algorithm consists of three steps:

- the local moving of the nodes.
- refining the partition between communities.
- accumulation of the nodes in the network on the basis of the refined partition, with the non-refined partition creating an initial partition for the accumulated network.

### 3.2 Community diversity score

It is often observed in real-world networks that influential nodes are usually linked to multiple communities in the network. Thus, they have a significant influence over the state of the network, as they can simultaneously disseminate their influence to multiple communities. We consider these high-degree nodes, which are connected to multiple communities, as influential nodes by applying the concept of Community Diversity (Kumar et al., 2022), as they act as bridge nodes between communities. The community diversity score of a node can be formally defined as the measure of the probability of finding a certain community in the neighborhood of a node. Community diversity gives a high score to a node that is connected to multiple communities, as well as considers the number of communities to which
the node transmits information. The concept of community diversity revolves around the concept of Shannon’s entropy (Lin, 1991) over the term of the probability of finding the $v^{th}$ node in the $c^{th}$ community in the neighborhood of a certain node. The Community Diversity score for a node is computed as shown in (2), where $CD_{v_i}$ is the community diversity value of the $i^{th}$ node, and $CD_{vic}$ is the contribution of $c^{th}$ community in Community Diversity of $i^{th}$ node (Kumar et al., 2022). The neighborhood contribution of each community is shown in (3), where $\eta_{in}$ represents the number of nodes in the $c^{th}$ community in the neighborhood of node $i$, and $\eta_{i}$ represents the number of nodes in the neighborhood of node $i$.

$$CD_i = -\sum_{c=1}^{\text{comm}} CD_{vic} = -\sum_{c=1}^{\text{comm}} P_{ic} \ast \log(P_{ic})$$ (2)

$$P_{ic} = \frac{\eta_{ic}}{\eta_{i}}$$ (3)

**Algorithm 1** Community diversity score (Kumar et al., 2022).

Algorithm 1 represents the procedure used to compute the Community Diversity (CD) score for a graph $G$. It can be summarized as follows:

- We divide the graph into communities using Leiden’s algorithm.
- For each node, we calculate the number of neighbours. Further, for each community, the number of neighbours of a node that are part of the respective community is computed.
- We evaluate the ratio of the total number of neighbours of the node and the neighbour of the node in the respective community. This is further used to calculate the entropy score, i.e., the Community Diversity score.

If a node is connected to multiple communities and the connections are found to be uniformly distributed, then the community diversity score of that node is observed to be higher. However, if a node has connections to only a few communities or the distribution
of connections is not uniform, then the community diversity score is observed to be low. Hence, the concept of community diversity can be utilized to find out the nodes capable of efficiently spreading information to multiple parts of the network.

### 3.3 Extended community diversity score

We formally propose the Extended Community Diversity score as the ratio of the community diversity score of a node to that of its neighborhood. The concept of Extended Community Diversity revolves around considering a node along with its neighborhood for the prediction of its influence in a locality. The neighbors of a node contribute to helping the algorithm understand what relative position a node holds with respect to its neighborhood, which ultimately tells how much a node will benefit from being made the seed node. Extended Community Diversity gains information about the neighborhood of a node by taking two hops, i.e., one hop from the node to its neighbors, and the second hop from its neighbors to farther nodes. This provides more information for the proposed algorithm to process and deduces which node is most influential. The provided information is further used to avoid having many influential nodes in the same region and hence avoid the wastage of useful resources. Extended Community Diversity score for a node is computed as shown in (4) where $ECD_v$ is the Extended Community diversity value of $v^{th}$ node, $neigh$ is the neighbours of node $v$ and $CD_v$ is the community diversity score for node $v$ as computed in (2).

$$ECD_v = CD_v + \sum_{n=1}^{neigh} CD_n$$  \hspace{1cm} (4)

---

**Algorithm 2** Extended community diversity score.

Algorithm 2 represents the algorithmic procedure used to compute the Extended Community Diversity (ECD) score for graph G. The summary of the algorithm is as follows:

- We calculate the previously defined Community Diversity score (CD) in Algorithm 1.
• Then, for each node, we calculate the summation of its Community Diversity score, and its neighbor’s Community Diversity Score, which gives the Extended Community Diversity score.

3.4 Modified community diversity score

We formally define the Modified Community Diversity (MCD) score as $MCD = -P \log P$, where $P$ is the ratio of the Community Diversity of a node to its Extended Community Diversity. Modified Community Diversity uses the limited available resources judiciously by ensuring that no two influential nodes are in the same locality, which, if not taken care of, would result in the wastage of valuable spreading power and overlapping of the region of influence of influential nodes in the same locality. The problem of several influential nodes existing in the same neighborhood is known as the Rich Club effect. Our proposed approach tackles this problem by scoring the nodes based on their strength relative to their neighboring nodes, thus ensuring that the influential nodes are well spread throughout the network. The ratio of the results of the previously calculated CD score and ECD score is used for analyzing the MCD score. The concept of MCD maximizes when the distribution of spreading power is uniform, which makes sense here since the algorithm not only wants the respective node to be influential but also the neighborhood to be sufficiently influential in carrying on the spread of information further. This enables the node to maximize the spread of information in that locality and beyond it. This results in the respective node being somewhat equivalent to the whole neighborhood in terms of spreading power, thus making it the most influential node in the locality. This mitigates the problem of the rich club effect significantly by considering the overlap in the influence and handling it in a balanced way to avoid wastage of useful spreading strength. This results in an optimized spread of information with assured maximum throughput. Modified Community diversity score for a node is computed as shown in (6), where $P_v$ is as shown in (6). In (6), $CD_v$ is the community diversity score of $v^{th}$ node, as shown in (2), and $ECD_v$ is the Extended Community diversity value of $v^{th}$ node, as calculated in (4).

$$MCD_v = -P_v \log(P_v)$$  \hspace{1cm} (5)

$$P_v = \frac{CD_v}{ECD_v}$$  \hspace{1cm} (6)

Algorithm 3 represents the procedure used to calculate the Modified Community Diversity (MCD) score for graph $G$. The summary of the algorithm is as follows:

• The community diversity score (CD) is calculated using Algorithm 1, and the extended community diversity score (ECD) is calculated using Algorithm 2.
• Then, for each node, we calculate the ratio of the Community Diversity score and Extended Community Diversity Score, which gives us the MCD score for the respective node.

Figure 1 illustrates the process flow for calculating the MCD score for each node. We first calculate the CD and ECD scores for each node. Using the generated values of CD and ECD scores, we compute the MCD scores for each node. We further rank the nodes on the basis of the MCD score and select the top $k$ nodes, where $k$ signifies the constraint on the number of nodes that act as influencers. We further evaluate the influence of the selected $k$ nodes over the entire network in Section 5.
4 Datasets, information propagation model, & performance metrics

4.1 Dataset and baseline models

We evaluate our proposed approach on seven real-life datasets and one synthetic dataset of varying sizes and types. Table 1 describes the different datasets used during our evaluation. Further, we do a comparative analysis of proposed approach with seven other state-of-the-art approaches, namely CSR (Kumar et al., 2022), GLR (Salavati et al., 2019), DCL (Berahmand et al., 2019), LID (Wen & Deng, 2019), DIL (Liu et al., 2016), H-Index (HI) (Lu et al., 2016) and PageRank (PR) (Brin & Page, 1998), over three distinct evaluation criteria. For the optimal performance for the baseline approaches, we took the input parameters as reported in the original papers throughout all our experiments.

- Community Spreaders Ranking (CSR) (Kumar et al., 2022): CSR measures a node’s spreading ability with respect to different communities on the basis of its connection to the communities.
- Gateway Local Rank (GLR) (Salavati et al., 2019): GLR extends over the closeness centrality measure and simplifies it by reducing the search set to the local and gateway nodes. To get the optimal performance, we take parameters $\alpha_1$ and $\alpha_2$ used for weighing the effect of core and gateway nodes to be 1 as reported (Salavati et al., 2019).
- Degree and Clustering Coefficient and Location (DCL) (Berahmand et al., 2019): DCL measures the spreading ability of nodes on the basis of a node’s degree, its neighbour, and the clustering coefficient.
- Local Information Dimensionality (LID) (Wen & Deng, 2019): LID measures the spreading ability of a node by considering the quasilocal structure of a node.
- Degree and Importance of Lines (DIL) (Liu et al., 2016): DIL ranks the nodes on the basis of the degree and the importance of lines.
- H-index (HI) (Lu et al., 2016): H-index is a node centrality measure that considers the neighborhood measures, the quality, as well as the quantity of the neighboring nodes simultaneously.

Algorithm 3  Modified community diversity score.
Table 1 The different datasets used for experimentation along with the various characteristics such as Number of Nodes, Edges, and Communities by Leiden Algorithm

| S.No. | Dataset     | Nodes (n) | Edges (m) | Communities | Source                             |
|-------|-------------|-----------|-----------|-------------|------------------------------------|
| 1     | BA          | 2,000     | 9,974     | 14          | (Barabasi & Albert, 1999)          |
| 2     | PGP         | 10,638    | 24,301    | 104         | (Boguna et al., 2004)             |
| 3     | Jazz Musicians | 198     | 2,742     | 4           | (Gleiser & Danon, 2003)           |
| 4     | Ca-hepth    | 7,126     | 35,324    | 481         | (Leskovec et al., 2007)           |
| 5     | GR-QC       | 5,242     | 14,496    | 394         | (Leskovec et al., 2007)           |
| 6     | Dolphins    | 62        | 159       | 5           | (Lusseau et al., 2003)            |
| 7     | Erdos992    | 6,100     | 7,500     | 50          | (Batagelj & Mrvar, 2000)          |
| 8     | p2p-Gnutella25 | 22,687   | 54,705    | 59          | (Ripeanu et al., 2002)            |

- PageRank (PR) (Brin & Page, 1998): PageRank measures the significance of a given node in the network based on its outgoing degree and damping factor. Similar to the work (Brin & Page, 1998) and various other works, we take the value of \( d \) to be 0.85 in all experiments to get optimal performance.

A brief description of different datasets is given below:

- **BA** (Barabasi & Albert, 1999): BA is a random graph generated using Barabási-Albert preferential attachment model. A network of \( n \) nodes is generated by adding new nodes, each having \( m \) edges that are preferentially coupled to existing nodes with a high degree.
- **PGP** (Boguna et al., 2004): Pretty Good Privacy (PGP) is an encrypted communication network. PGP algorithm is used for secure information interchange. Here, the users are nodes, and the edges of this huge network are the connectivity between users.
- **Jazz Musicians** (Gleiser & Danon, 2003): The Jazz musicians dataset is a real-world social network dataset, where each node represents a musician, and each edge between two nodes indicates that the two musicians have played together in a band.
- **Ca-hepth** (Leskovec et al., 2007): It represents a collaboration network of authors who have collaborated on papers submitted to the High Energy Physics - Theory Category.
- **General Relativity and Quantum Cosmology (GR-QC)** (Leskovec et al., 2007): GR-QC dataset is a real-world social network dataset from arXiv and covers scientific contributions by authors in General Relativity and Quantum Cosmology. The authors are nodes, and if they have collaborated with another author on a paper, then there is an undirected edge between them.
- **Dolphins dataset** (Lusseau et al., 2003): The dolphins dataset is a social network of bottlenose dolphins, where a node represents a dolphin, while an edge represents frequent associations between dolphins.
- **Erdos992** (Batagelj & Mrvar, 2000): This dataset is about the collaborations of Paul Erdős, a Hungarian Mathematician, with his co-authors. The co-authors are nodes while collaborations between the nodes signify the edges.
• p2p-Gnutella25 (Ripeanu et al., 2002): This dataset is one of the snapshots of the Gnutella peer-to-peer file sharing network from August 2002. Nodes represent hosts in the Gnutella network, while edges represent connections between the Gnutella hosts.

4.2 Information propagation model

Information Propagation (IP) is described as the task of circulating information in complex networks (Chen et al., 2013; Guille et al., 2013). IP can be used to assess various spreader-ranking approaches. Generally, epidemic models are used to model information propagation in a complex network. In epidemic models, the population of users is represented as nodes, and the requisite information is regarded as an infectious disease, which propagates through the population upon the interaction between nodes.

While various epidemic models can be used for information propagation tasks, like SIR (Hethcote, 2000), IC (Goldenberg et al., 2001) and LT (Granovetter, 1978), we employ the Independent Cascade (IC) model to assess the performance of our proposed approach and evaluate it against other state-of-the-art approaches. The nodes in an IC model can be allocated one out of two states: infected node or susceptible node. The propagation of information in the network begins with setting the state of a few nodes as infected. These nodes further interact with their neighboring nodes and try to infect them. Further, the newly infected nodes further interact with their neighbors to further propagate the information to their neighbouring nodes until there are no new nodes in the network that are infected. A probability ($P$) is associated with these interactions between nodes, which elucidates the possibility of a susceptible node getting infected upon interaction with an infected node. The total influence of a seed node is calculated by taking the sum of all nodes which get infected by the end of information propagation.

4.3 Performance metrics

We use three performance metrics for comparing our approach with other state-of-the-art approaches. They are as follows:

(i) Final Infected Scale: It is defined as the ratio of activated or infected nodes in the network to the total number of nodes in the network at the end of information propagation. The nodes only get activated when the activation probability for a node is favourable for activation. The fraction of the total nodes taken as the initial spreaders is called the spreader fraction, and it plays a significant role in the final number of nodes that get infected. Generally, a high spreader fraction leads to a high percentage of the total nodes getting infected.

(ii) Average Shortest Path Length (ASPL): ASPL can be defined as the average of the shortest paths between two influential nodes for all sets of influential nodes. It enables us to evaluate the efficacy of information propagation in a complex social network. The higher the value of ASPL, the more even the spread of the influential nodes in the network. A high value of ASPL further indicates that influential nodes cover a wider part of the network instead of being concentrated in a small region, leading to the poor spread of information in the network.

(iii) Execution Time: Execution Time: We evaluate our approach against other approaches based on the time taken to rank nodes based on their ability to propagate information in a network. The time consumption for all centralities was assessed in a Google Colaboratory environment. The code was run on an Intel Xeon CPU with one core, with 13GB of RAM available for use.
5 Experimental results and analysis

This section presents the experimental results and analysis performed on the eight datasets using three performance metrics as mentioned in Section 4. We conducted a comparative analysis of the proposed approach with seven other state-of-the-art approaches for finding influential nodes, namely, H-Index (HI) (Lu et al., 2016), PageRank (PR) (Brin & Page, 1998), Community-based Spreader Ranking (CSR) (Kumar et al., 2022), Gateway Local Ranking (GLR) (Salavati et al., 2019), Degree and Clustering Coefficient and Location (DCL) (Berahmand et al., 2019), Local Information Dimensionality (LID) (Wen & Deng, 2019), and Degree and Importance of Lines (DIL) (Liu et al., 2016). We utilized the Information Cascade (IC) model to simulate information propagation. Our analysis of the results obtained over three metrics, illustrated in Figs. 2, 3, and 4, is elaborated below.

5.1 Final infected state

The Final Infected Scale illustrates the final degree of propagation or infection in the network at the end of the simulation, for varying values of fraction of nodes taken as initial spreaders. We evaluated our proposed approach against other competing approaches using the IC simulation model for 100 test runs and plotted the average results in Fig. 2. For small datasets, i.e., datasets with the number of nodes less than 2000, we take the values of the initial spreader fraction from the set 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, while for larger datasets with the number of nodes greater than 2000 nodes, we take the values of the initial spreader fraction from the set 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04. We have taken larger values of the initial spreader fraction for smaller datasets, as the absolute number of initial spreaders will be small for smaller values of the initial spreader fraction in smaller datasets. We have taken the infection probability, i.e., the probability of influencing a susceptible node as 0.1 for all combinations of datasets and approaches.

The graphs for Final Infected Scale with variations in Spreader Fraction for all approaches and datasets are illustrated in Fig. 2. We observed that the final infected scale generally increased with an increase in spreader fraction, irrespective of the dataset. We observed that our proposed approach, MCD, outperformed competing approaches significantly across all datasets, except the Dolphins dataset, where PR outperforms our approach. HI was consistently among the worst-performing approaches of the lot for most datasets. Thus, we conclude that our approach handily outperforms other approaches in maximizing information propagation in a network for a wide set of values of the initial spreader fraction.

5.2 Average distance between spreaders

It is desirable for the initial spreaders to have the least overlapping area of influence so that a wider portion of the network can be activated more efficiently. Thus, it is preferable that the average distance between spreaders is maximized. To compare our proposed MCD approach against other approaches, we calculated the average shortest distance between the initial spreaders \( L_s \) for a wide set of values of the spreader fraction.

The graphs for the Average Distance between Spreaders vs. Initial Spreader Fraction for all approaches and datasets are illustrated in Fig. 3. We observed that MCD significantly outperforms other approaches across datasets for almost all values of the initial spreader fraction. This ensures that MCD selects nodes that are far apart, thus avoiding overlapping of influence as stated in the manuscript. It verifies that MCD is able to cater to the rich
Fig. 2 Final infected scale plots for various values for initial spreader fraction on the (a) PGP dataset (b) BA dataset (c) Ca-hepth dataset (d) GR-QC dataset (e) Jazz Musicians dataset (f) Dolphins dataset (g) Erdos992 dataset (h) p2p-Gnutella25 dataset. The results are averaged for 100 independent simulations of the Information Cascade model with an activation probability (P) equal to 0.1
Fig. 3  Average distance between spreaders with respect to different initial spreaders fraction values for (a) PGP dataset (b) BA dataset (c) Ca-hepth dataset (d) GR-QC dataset (e) Jazz Musicians dataset (f) Dolphins dataset (g) Erdos992 dataset (h) p2p-Gnutella25 dataset. The results are averaged for 100 independent simulations of the Information Cascade model with an activation probability (P) equal to 0.1
Fig. 4 The execution time plots for different methods for ranking the spreaders for (a) PGP dataset (b) BA dataset (c) Ca-hepth dataset (d) GR-QC dataset (e) Jazz Musicians dataset (f) Dolphins dataset (g) Erdos992 dataset (h) p2p-Gnutella25 dataset
club effect, which other competing approaches are unable to counter. The significant gap in the figure between our approach and competing approaches indicates that MCD selects nodes that are different from the ones selected by other approaches, and significantly farther apart. This further ensures nodes selected by MCD may not belong to the same region, thus having a wider spread of influence and lesser overlap between the regions of influence for each seed node. Further, we observe that on increasing the spreaders fraction, the difference between seed nodes decreases due to the increased absolute number of seed nodes for the whole network. This is observed in the case of all approaches, as increasing the absolute number of spreaders in a graph is bound to decrease the average distance between the spreaders. However, MCD still has a higher average spreader distance as compared to other approaches, even for greater values of Spreader Fraction, verifying MCD’s superiority over competing approaches. Among the other approaches, we found DCL and PR to perform reasonably well. However, there was still a significant gap between these approaches and MCD. We found GLR consistently performs the worst in this metric across almost all datasets. Thus, we conclude that MCD suffers from a significantly lesser rich club effect and outperforms all other approaches significantly.

5.3 Execution time

We compared our proposed approach, MCD, against other competing approaches based on the execution time required by each approach for ranking the nodes for all datasets. Figure 4 illustrates the line graph comparing MCD with other approaches based on the execution time for all datasets. DIL performed well on larger datasets but was relatively more computationally expensive for smaller datasets. We found LID and PR to be the least expensive computationally, while GLR was consistently the most expensive. Our proposed approach was computationally the least expensive for smaller datasets, while it was more expensive than local hybrid centrality-based approaches like LID for larger datasets. Thus, we observe that MCD performs well in terms of execution time when compared to other approaches, especially when considering the improved performance in other metrics.

6 Statistical testing

To further evaluate the performance of MCD using statistical tests, we have conducted the Friedman test (Friedman, 1937), which is a non-parametric statistical test used for multiple comparisons (more than two given methods). We have performed this test to determine whether the performance of our proposed method is significantly different than the other approaches for two or more data sets. The Friedman test attempts to detect significant differences based on the ranking of the methods, rather than their errors. It consists of two hypotheses: the null \( H_0 \) and the alternate hypothesis \( H_1 \). The former states that there are no prominent differences between these algorithms (equality of medians) and the latter insists that the algorithms have significant differences in their median of the population, thus negating the null hypothesis. We have performed the following test on the Final Infected Scale metric as well as the Average Distance between Spreaders metric.

In order to perform the Friedman Statistical Test, we execute the following procedure:

- Gather the generated results for each problem pair.
- Rank the values in ascending order from 1 (best value) to \( n \) (worst value) for each problem \( i \) for a particular algorithm \( j \).
For algorithm \( j \), the average rank for each problem \( i \) is calculated using the (7):

\[
R_j = \frac{1}{n} \sum_{i=1}^{n} r_i^j
\]  

(7)

where \( r_i^j \) is the rank \((1 < j < k)\) and \( R_j \) is the average rank.

Now that all algorithms are ranked according to their priority, compute the Friedman Statistic \( F_f \) using the following equation. \( F_f \) is based on a Chi-square distribution, with \( k - 1 \) degree of freedom. \( F_f \) is computed as shown in (8).

\[
F_f = \frac{12n}{k(k+1)} \left[ \sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right]
\]  

(8)

The previous statistic produced relatively conservative results, which were not desired. Thus, Iman and Davenport (1980) introduced a new statistic \( F_{id} \), as shown below, which follows the \( F \) - distribution with the degree of freedom as \( k - 1 \) and \((n - 1)(k - 1)\), given in (9):

\[
F_{id} = \frac{(n - 1) \chi^2_F}{n(k - 1) - \chi^2_F}
\]  

(9)

Tables 2 and 3 show the average ranking calculated from the Friedman test for all the approaches used in this paper. MCD clearly has a better average ranking than the other competing approaches.

The unadjusted P-value or Holm P-value, obtained from the Iman-Davenport statistic, corresponding to the performance of MCD, advocates the rejection of the null hypothesis \( H_0 \). All of the computed P-values are less than the standard significance level \( \alpha = 0.05 \). This indicates that there is a significant difference between the performance of the baseline approaches and our proposed approach. Hence, the obtained P-values conclude the negation of null hypothesis \( H_0 \) but are not appropriate for comparison with different methods. In order to compare these approaches with each other, we must calculate their respective adjusted P-values keeping MCD as the control algorithm.

Hence, we use adjusted P-values (APVs) for evaluation in order to perform a comparative study between the control method and benchmark methods on a statistical ground. Adjusted P-value provides the correct correlation between these algorithms, taking into account the accumulated family error with respect to the MCD control algorithm. Moreover, the adjusted P-values can be directed compared with the significance level \( \alpha \), which is equivalent to 0.05. In order to calculate the adjusted P-values, a few post-hoc procedures

### Table 2

| Method | Average Rank |
|--------|--------------|
| MCD    | 1.567        |
| PR     | 2.983        |
| DCL    | 3.467        |
| CSR    | 4.675        |
| LID    | 5.033        |
| DIL    | 5.225        |
| GLR    | 6.517        |
| HI     | 6.533        |
Table 3: Average ranking of algorithms using the Friedman test based on Average Distance between Spreaders

|  | Approach | Average Rank  |
|---|-----------|-------------|
| 1 | MCD       | 1.067       |
| 2 | PR        | 2.983       |
| 3 | DCL       | 3.517       |
| 4 | LID       | 4.675       |
| 5 | HI        | 5.3         |
| 6 | CSR       | 5.367       |
| 7 | DIL       | 5.508       |
| 8 | GLR       | 7.583       |

need to be defined. Various post-hoc procedures such as Dunn (1961) and Holland and Copenhaver (1987) differ as they adjust the value of $\alpha$ to compensate for multiple comparisons for multiple methods. In this paper, we have used the common Holm’s procedure (Holm, 1979) to evaluate the respective APVs. The values of Holm P-values and APVs are always sorted in ascending order. The equation for the same is given below, where indices $i$ and $j$ refers to the main hypothesis whose APVs are being computed and different hypothesis in the set respectively. $P_j$ is the P-value for the $j^{th}$ hypothesis. Holm APV is computed as shown in (10).

$$Holm\ APV_i = \min\{v, 1\}, \text{ where } v = \max\{(k - j)p_j : 1 \leq j \leq i\}$$  

(10)

The results of the APVs using Holm’s procedure are shown in Tables 4 and 5. The intended APVs are less than the significance level, thus rejecting the null hypothesis. It is clearly observed from the results obtained that MCD is significantly better than the competing approaches on these statistical tests as well.

7 Conclusion

Finding influential nodes in complex networks like social networks play a vital role in achieving maximum information spread across the network. It has found numerous applications in business, marketing, and generating awareness for a social cause. In this work, we proposed a modified Community Diversity (MCD) method for finding influential nodes. First, we introduced the concept of Extended Community Diversity, which represents the community diversity score of the extended neighborhood of a node. Next, we introduced Modified Community Diversity, built upon the concepts of Community Diversity and

Table 4: Adjusted P-values (APVs) using Holm procedure for Final Infected State

|  | Approach | Z-score | P-value  | APV     |
|---|----------|---------|----------|---------|
| 1 | HI       | -11.11  | 5.88E-29 | 4.11E-28 |
| 2 | GLR      | -11.06  | 8.91E-29 | 5.35E-28 |
| 3 | DIL      | -8.18   | 1.42E-16 | 7.08E-16 |
| 4 | LID      | -7.75   | 4.53E-15 | 1.81E-14 |
| 5 | CSR      | -6.95   | 1.82E-12 | 5.46E-12 |
| 6 | DCL      | -4.24   | 1.08E-05 | 2.15E-05 |
| 7 | PR       | -3.16   | 7.68E-04 | 7.68E-04 |
Table 5  Adjusted P-values (APVs) using Holm procedure for Average Distance between Spreaders

| Approach | Z-score | P-value  | APV   |
|----------|---------|----------|-------|
| 1 GLR    | −14.57  | 2.13E-48 | 1.49E-47 |
| 2 DIL    | −9.93   | 1.51E-23 | 9.08E-23 |
| 3 CSR    | −9.61   | 3.45E-22 | 1.73E-21 |
| 4 HI     | −9.46   | 1.45E-21 | 5.81E-21 |
| 5 LID    | −8.06   | 3.56E-16 | 1.07E-15 |
| 6 DCL    | −5.47   | 2.15E-08 | 4.29E-08 |
| 7 PR     | −4.28   | 9.10E-06 | 9.10E-06 |

Extended Community Diversity. To verify the efficacy of our approach in determining the most effective influential nodes in a network, we compared MCD with seven recent state-of-the-art methods over eight datasets. We found that MCD cumulatively outperformed the competing approaches across three metrics. Further, we observed that seed nodes selected by MCD were reasonably close together for extended activation due to a node’s influence but far enough to not significantly overlap with each other’s scope of influence, thus mitigating the Rich Club effect significantly.

Acknowledgements  We would like to thank Aryaman Rao for his help in creating the revised manuscript.

Author Contributions  Aryan Gupta and Inder Khatri- Conceptualization, Methodology, Visualization, Investigation, Writing, Reviewing and Editing. Arjun Choudhary- Conceptualization, Methodology, Software, Writing- Original draft preparation. Sanjay Kumar- Conceptualization, Supervision, Reviewing, and Editing.

Funding  No funding was received for conducting this study.

Data Availability  All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval  Not Applicable

Competing interests  The authors have no competing interests to declare that are relevant to the content of this article.

References

Banerjee, S., Jenamani, M., & Pratihar, D. (2020). A survey on influence maximization in a social network. Knowledge and Information Systems, 62, 3417–3455. https://doi.org/10.1007/s10115-020-01461-4.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512. https://doi.org/10.1126/science.286.5439.50.

Batagelj, V., & Mrvar, A. (2000). Some analyses of Erdos collaboration graph. Social Networks, 22(2), 173–186. https://doi.org/10.1016/S0378-8733(00)00023-X.

Berahmand, K., Bouyer, A., & Samadi, N. (2019). A new local and multidimensional ranking measure to detect spreaders in social networks. Computing, 101, 1711–1733. https://doi.org/10.1007/s00607-018-0684-8.

Boguna, M., Pastor-Satorras, R., Diaz-Guilera, A., & et al. (2004). Models of social networks based on social distance attachment. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 70, 056122. https://doi.org/10.1103/PhysRevE.70.056122.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. Social Networks, 29, 555–564. https://doi.org/10.1016/j.socnet.2007.04.002.
Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks, 30*, 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Chen, X., Deng, L., Zhao, Y., & et al. (2021). Community-based influence maximization in location-based social network. *World Wide Web, 24*(6), 1903–1928. https://doi.org/10.1007/s11280-021-00935-x.

Chen, W., Lakshmanan, L., & Castillo, C. (2013). Information and influence propagation in social networks. *Synthesis Lectures on Data Management, 5*, 1–177. https://doi.org/10.2200/S00527ED1V01Y201308DTM037.

Chen, D., Lu, L., Shang, M. S., et al. (2012). Identifying influential nodes in complex networks. *Fuel and Energy Abstracts, 391*, 1777–1787. https://doi.org/10.1016/j.physa.2011.09.017.

Chen, L., Zheng, H., Li, Y., & et al. (2022). Enhanced density peak-based community detection algorithm. *Journal of Intelligent Information Systems, 1–22*. https://doi.org/10.1007/s10844-022-00702-y.

Chouchani, N., & Abed, M. (2020). Online social network analysis: Detection of communities of interest. *Journal of Intelligent Information Systems, 54*, 1–17. https://doi.org/10.1007/s10844-018-0522-7.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090.

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

Freeman, L. C. (1978). Centrality in social networks: conceptual clarification. *Social Networks, 1*, 215–239. https://doi.org/10.1016/0378-8733(78)90017-7.

Free, L. C. (1977). Set of measures of centrality based on betweeness. *Sociometry, 40*(1), 35–41. https://doi.org/10.2307/3033543.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*, 215–239. https://doi.org/10.1016/0378-8733(78)90017-7.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522.

Gao, S., Ma, J., Chen, Z., & et al. (2014). Ranking the spreading ability of nodes in complex networks based on local structure. *Physica A: Statistical Mechanics and its Applications, 403*, 130–147. https://doi.org/10.1016/j.physa.2014.02.032.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication, 3*(1), JCMC313. https://doi.org/10.1111/j.1083-6101.1997.tb0062x.

Gleiser, P. M., & Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems, 06*(04), 565–573. https://doi.org/10.1142/S0219525903001067.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters, 12*, 211–223. https://doi.org/10.1023/A:1011122126881.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*, 1420–1443. https://doi.org/10.1086/226707.

Guille, A., Hacid, H., Favre, C., & et al. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record, 42*(2), 17–28. https://doi.org/10.1145/2503792.2503797.

He, Q., Wang, X., Mao, F., & et al. (2020). CAOM: A community-based approach to tackle opinion maximization for social networks. *Information Sciences, 513*(C), 252–269. https://doi.org/10.1016/j.ins.2019.10.064.

Hetherly, W. (2000). The mathematics of infectious diseases. *SIAM Review, 42*(4), 599–653. https://doi.org/10.1137/S0036144500371907.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43*, 417–423. https://doi.org/10.2307/2531823.

Holm, S. A. (1979). Simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://doi.org/10.2307/4615733.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the fbietkan statistic. *Communications in Statistics - Theory and Methods, 9*(6), 571–595. https://doi.org/10.1080/0361092808827904.

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 137–146). Association for Computing Machinery. https://doi.org/10.1145/956750.956769.

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining*. KDD’03 (pp. 137–146). New York: Association for Computing Machinery. https://doi.org/10.1145/956750.956769.

Khomami, M. M. D., Rezvanian, A., Meybodi, M. R., & et al. (2021). CFIN: A community-based algorithm for finding influential nodes in complex social networks. *Journal of Supercomputing, 77*(3), 2207–2236. https://doi.org/10.1007/s11227-020-03355-2.
Kumar, S., Gupta, A., & Khatri, I.C.S.R. (2022). A community-based spreaders ranking algorithm for influence maximization in social networks. World Wide Web. https://doi.org/10.1007/s11280-021-00996-y.

Kumar, S., Kumar, A., & Panda, B.S. (2022). Identifying influential nodes for smart enterprises using community structure with Integrated Feature Ranking. IEEE Transactions on Industrial Informatics, 19(1), 703–11. https://doi.org/10.1109/TII.2022.3203059.

Kumar, S., Mallik, A., & Panda, B.S. (2023). Influence maximization in social networks using transfer learning via graph-based LSTM. Expert Systems with Applications, 212, 118770. https://doi.org/10.1016/j.eswa.2022.118770.

Kumar, S., Mallik, A., & Sengar, S.S. (2022). Community detection in complex networks using stacked autoencoders and crow search algorithm. The Journal of Supercomputing, 1–28. https://doi.org/10.1007/s11227-022-04767-y.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. KDD ’06 (pp. 611–617). New York: Association for Computing Machinery. https://doi.org/10.1145/9781-4419-6515-8.13.

Kumar, S., Singhla, L., Jindal, K., & et al. (2021). IM-ELPR: Influence maximization in social networks using label propagation based community structure. Applied Intelligence, 51, 7647-7665. https://doi.org/10.1007/s10489-021-02266-w.

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

Leskovec, J., Krause, A., Guestrin, C., & et al. (2007). Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. KDD ’07 (pp. 420–429). New York: Association for Computing Machinery. https://doi.org/10.1145/1281192.1281239.

Li, Y., Fan, J., Wang, Y., & et al. (2018). Influence maximization on social graphs: A survey. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1852–1872. https://doi.org/10.1109/TKDE.2018.2807843.

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151. https://doi.org/10.1109/18.61115.

Liu, Y., Qiu, L., & Sun, C. (2022). An influence maximization algorithm based on low-dimensional representation learning. Applied Intelligence, 1–18. https://doi.org/10.1007/s10489-022-03178-z.

Liu, J., Xiong, Q., Shi, W., & et al. (2016). Evaluating the importance of nodes in complex networks. Physica A: Statistical Mechanics and its Applications, 452, 209–219. https://doi.org/10.1016/j.physa.2016.02.049.

Lu, L., Zhou, T., Zhang, Q. M., & et al. (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.

Lusseau, D., Schneider, K., Boisseau, O. J., & et al. (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 54(4), 396–405. https://doi.org/10.1007/s00265-003-0651-y.

Ma, L., Chiew, K., Huang, H., & et al. (2017). Evaluation of local community metrics: from an experimental perspective. Journal of Intelligent Information Systems, 51, 1–22. https://doi.org/10.1007/s10844-017-0480-5.

Mendonça, R. F. M., da Motta Salles Barreto, A., & Ziviani, A. (2022). Efficient information diffusion in time-varying graphs through deep reinforcement learning. World Wide Web. https://doi.org/10.1007/s11280-021-00998-w.

Mislove, A., Marcon, M., Gummad, K. P., & et al. (2007). Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on internet measurement (pp. 29–42). Association for Computing Machinery. https://doi.org/10.1145/1298306.1298311.

Okamoto, K., Chen, W., & Li, X.Y. (2008). Ranking of closeness centrality for large scale social networks. In Frontiers in algorithmics, (Vol. 5059 pp. 186–195). Springer. https://doi.org/10.1007/978-3-540-69311-6.21.

Ripeanu, M., Foster, I., & Iamnitchi, A. (2002). Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Computing Journal, 6.

Salavati, C., Abdollahpour, A., & Manbari, Z. (2019). Ranking nodes in complex networks based on local structure and improving closeness centrality. Neurocomputing, 336(C), 36–45. https://doi.org/10.1016/j.neucom.2018.04.086.

Traag, V. A., Waltman, L., & Van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, 9(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z.

Wen, T., & Deng, Y. (2019). Identification of influencers in complex networks by local information dimensionality. Information Sciences, 512, 549–562. https://doi.org/10.1016/j.ins.2019.10.003.
Yang, H., Cheng, J., Su, X., & et al. (2021). A Spiderweb model for community detection in dynamic networks. *Applied Intelligence, 51*(7), 5157–5188. https://doi.org/10.1007/s10489-020-02059-7.

Zhou, J., Fan, J., Wang, J., & et al. (2019). Cost-efficient viral marketing in online social networks. *World Wide Web, 22*, 1–24. https://doi.org/10.1007/s11280-018-0651-5.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.