International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Influential Nodes Identification in Complex Networks: Sampling Approach

Karzan K. Abdulmajeed*[1], Abdulhakeem O. Mohammed[2],

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: Accurately identifying influential nodes within complex networks is crucial for understanding information and influence propagation. Existing state-of-the-art algorithms, while powerful, often rank all nodes, which can be computationally expensive and unnecessary for many applications. In this paper, we propose a simple yet efficient approach that overcomes these limitations. Initially, a systematic sampling methodology was employed to strategically select a subset of nodes from the network, representing a small fraction of its entirety. Subsequently, the betweenness centrality of these sampled nodes was estimated to facilitate their ranking. To assess the performance of our sampling method alongside alternative algorithms, we employ the stochastic Susceptible–Infected–Recovered (SIR) information diffusion model to compute various metrics including the infection scale, the final infected scale over time, and the average distance between spreaders. Our experimental findings, conducted on real-world networks, indicate that our proposed method accurately identifies influential nodes while maintaining significant computational efficiency.

Keywords: *Complex Network, Influential node, Centrality indices, Sampling, SIR model.*

1. Introduction

During the last few years, networks have gained immense traction, and there has been a notable surge in the user base of networks, increasing to billions of records while collecting a gigantic amount of data [1]. This data is used to identify influential spreaders in complex networks. With these huge records of data, the understanding of influential nodes must be more focused as they play a crucial role in targeting advertisements over media [2][3][4], spreading diseases [5], and word-of-mouth on social networks [6][7], which can be described by information spreading on networks. The importance of this huge data has become increasingly clear, and identifying influential nodes has become more important because they have a strong ability to affect other nodes. Therefore, the number of research activities has drastically increased in an attempt to obtain influential nodes in complex networks.

Centrality, a straightforward concept in network analysis, evaluates the importance of nodes through various metrics by assigning a real value to each node in the network, where the values produced are expected to provide a ranking of nodes to determine their importance, including degree [8], betweenness [9], closeness [10], and VoteRank centralities [11]. Degree centrality, is a fundamental local metric widely adopted for ranking user influence that focuses on the immediate neighborhood or local connection between nodes within the network., proves to be efficient in numerous scenarios. However, its reliance solely on node degree overlooks the broader topological structure of the network, leading to potential inaccuracies in identifying influential spreaders [11][12][13]. In contrast, betweenness and closeness centralities, esteemed global metrics which analyze the whole network to identify spreaders that play significant roles in connecting different parts of the network, and predicated on the premise of influence propagation along the shortest paths [14]. Nonetheless, these metrics entail high computational complexity as they necessitate the computation of shortest paths between all pairs of nodes in the network. VoteRank centrality[15], which is founded upon a voting mechanism, identifies a group of spreaders wherein each node possesses equal voting capacity and receives votes from its neighbors. Nevertheless, all centrality algorithms often compute some scores for all nodes and then rank them based on their scores [16], which is computationally expensive and unnecessary for many applications. In addition, many other approaches exist to identify influential nodes in complex network that play crucial role in reducing the time complexity such as, PageRank [17], ClusterRank[18], and LeaderRank [19][20].

However, it is important to consider the network structure when identifying influential spreaders and gaining a better understanding of how the network operates [21]. The specific location of a node within the network holds

[1] *Dep. of Computer Science-college of science - University of Zakho - Duhok, Iraq*
2 Dep. of Computer Science-college of science - University of Zakho - Duhok, Iraq
Corresponding Author Email: karzan.abdulmajeed@uoz.edu.krd

particular significance. Therefore, there are various approaches that rely on the network structure to identify influential nodes [13][22][23]these approaches evaluate the influential nodes in network based on the score of each node that achieved . There are several evaluation models that define the importance of a node in a network. Among these models, the most commonly used one is the SIR model [24]. In the SIR model, nodes in the network are classified as susceptible (S), infectious (I), or recovered (R). The model measures a node's spreading ability by infecting the node with the highest rank initially, and then assessing the number of nodes that have recovered in the network. The infection process stops if there are no nodes remaining in the infected list after a certain number of time steps.

In this paper, we propose a novel approach to accurately identifying influential nodes within complex networks while addressing the computational overhead associated with existing algorithms. Our methodology introduces a systematic sampling technique combined with an estimation method for betweenness centrality, enabling the strategic selection of a small subset of nodes from the network. This alleviates the need to rank all nodes, resulting in enhanced computational efficiency without compromising accuracy. Through empirical evaluation on real-world networks using the SIR information diffusion model, we demonstrate the effectiveness of our approach in accurately identifying the influential nodes in a network.

2. Related Work

In this section, we provide an overview to the most common measures and algorithms that are used to identify influential node in networks. Identifying influential nodes within complex networks is a pivotal endeavor in network analysis. Such nodes play a crucial role in expediting the dissemination of information, thereby optimizing the coverage of nodes within the network in fewer steps [25].

Maji et al. [21] conducted a survey to identify influential spreaders in a complex network through various centrality measures such as degree, closeness, coreness (k-shell centrality), etc. Additionally, a mathematical formulation is introduced to improve the K-shell method instead of guessing the value of the node through trial and error.

Lu et al. [19] developed the LeaderRank (LR) algorithm, which relies on random walkers and utilizes the stochastic matrix to identify the influence of nodes in a directed complex network. LR introduces a ground node (Leader) that is connected to every node (fans) in the network through bidirectional weighted edges, forming a leadership network. Additionally, LR is more resilient to noisy data and resistant to manipulations. LR only depends on the in-degree of each node for the identification of an influential node.

Lie et al. [26] introduced a novel metric called neighborhood centrality to identify influential spreaders in complex networks, which is based on the centrality or coreness of a node and its neighbors' centrality. The metric considers not only the importance of the node's direct neighbors but also its 2-step and even more steps neighborhood, in order to define the importance of these connections.

Zhang et al. [27] introduced voteRank to identify influential spreaders in unweighted and undirected complex network. The authors presented strategies for node selection to reduce the overlapping spreading influence from both individual and group perspectives. Kumar and Panda [28] introduced the Neighbor- hood Coreness algorithm to enhance the resolution of the VoteRank algorithm, leveraging k-shell values of neighbors for refinement. Despite these efforts, the challenge of overlapping influential regions of spreaders persisted. Liu et al. [29] proposed a novel method named VoteRank++, which is based on the foundation of the VoteRank method proposed in [27], to identify influential nodes within complex networks. Within the context of VoteRank++, nodes with different degrees are assigned distinct levels of voting weight. Further- more, A node can vote differently for its neighbors based on the different levels of closeness between nodes. In contrast, Chen et al. [25] proposed a local ranking approach named ClusterRank to identify influential nodes in directed networks. This approach includes both the local clustering coefficient and the influence of neighbors to determine the importance of the node in the network.

Ma et al. [30] introduced a novel approach named Extended Gravity Centrality (EGC) was developed for the identification of influential spreaders within networks, employing a gravity formula incorporating k-shell values and the shortest distance between nodes. Independently, Li et al. introduced the Local Gravity Model (LGM) [31], which relies on degree values and shortest path lengths between nodes. Notably, both EGC and LGM necessitate the computation of shortest paths, rendering them computationally intensive for large-scale graphs. To mitigate this, Yang and Xiao [32] proposed an enhanced gravity model called K-shell-based Gravity Centrality (KSGC), which leverages the K-shell algorithm to account for both local and global structural features, thereby improving efficiency in identifying influential nodes across complex networks.

Wang et al. [33]proposed a new measure called Efficiency centrality (EFFC) to identify influential spreaders in a complex network. EFFC considers the impact of the node in the network before and after removing nodes and its edges that the node is connected to neighbors. The removal process might change the degree of network efficiency and the structure of the network, particularly the removal of pivotal nodes, which will prominently change the whole network's efficiency.

Tulu et al.[34] introduced a novel approach called the Community-Aware Mediator (CAM) aimed at discerning influential nodes within intricate networks. This methodology gauges a node's influence through the entropy of random walks across various communities. Consequently, such nodes assume pivotal roles within their respective communities, facilitating the exchange and dissemination of information. Bae and Kim [35] propose a novel measure called coreness centrality (CC) to identify influential spreaders in unweighted and undirected networks. This measure uses both the degree and the coreness of each node by considering the k-shell of the neighbors or the neighbors of neighbors that are adjacent to a spreader. CC can estimate the powerful influential spreaders that have more connections and reside in the core network.

Li et al. [36] proposed a novel centrality named Clustered Local-Degree (CLD) to identify influential spreaders in a complex network based on the local clustering of a node and the degrees of its neighbors. It calculates the degrees of the nearest neighbors of a given node, combines the sum and the clustering coefficients of the nodes to rank spreaders.

Zareie et al. [37] proposed a method to identify influential spreaders in a network based on neighborhood diversity. Their proposed method uses k-shell to determine the centrality of nodes based on sphere diversity, which is define. Each node will obtain a ranking value, which will define the importance of the node. Another work by Zareie et al. [38] proposed an enhanced cluster rank approach aimed to find influential spreaders within a network. This method leverages neighborhood correlation coefficients alongside k-shell decomposition to discern influential entities. Identifying influential spreaders in a network depends on how nodes share connections with neighbors to provide a more detailed correlation structure between nodes and discover the influential spreaders.

Namtirtha et al. [39] proposed an indexing method named the k-shell hybrid method to identify highly influential spreaders not only from the core but also from lower shells. Their proposed method combines node degree and K-shell index, deriving benefits from both global measure (k-shell) and local measure (K), along with their combination. Another work by Namtirtha et al. [40] proposed a new method called weighted k-shell degree (KSD) neighborhood for identifying influential spreaders from a variety of complex network connectivity structures by assigning weights to the edges using the node degree and k-shell index of end nodes.

Guo et al. [41] proposed a method to identify a set of influential nodes in a complex network based on information entropy. The node with the highest number of connections with other nodes has a greater ability to be an influential node. Xu et al. [42] introduced adjacency information entropy to determine the vital node in a weighted and directed complex network.

Generalized mechanics model has been introduced by Liu et al. [43] to identify influential spreaders in networks. Utilizing a Weighted Gravity model (WGravity) [44], the significance of nodes is assessed by amalgamating local and global network connections. This model is based on the calculation of the node degree and largest value of the normalized eigenvector.

Zhao et al. [45] introduced a novel method called the global importance of nodes (GIN) metric to identify influential spreaders in unweighted and undirected complex networks based on the K-shell. The method considered both the local influence and the global influence of the nodes simultaneously, thereby identify nodes that may appear unimportant but are actually important in the complex network.

Liu et al. [46] proposed an approach to identify influential nodes based on graph traversal. The approach uses a breadth-first search (BFS) tree, in which the target node is the root node. The BFS tree assigns a score value to each node. The length of the tree should be short, and a node is considered influential when it is at the top level of the tree. Additionally, the local neighborhood of the root node has a higher expectation of being an influential one as well.

Gupta and Mishra [22] propose a novel method to identify the top-k influential spreaders in undirected and unweighted complex real-world networks using network structure. It computes the normalized global importance (NGI), which depends on the nodes degree, normalized iteration multiplier (NIM), and k-core decomposition.

Zhao et al. [23] proposed a ranking approach to identify influential spreaders in a network, which is based on structure holes and the k-shell algorithm. This approach can identify not just the core nodes with high k-shell indices but also the nodes that have lower k-shell indices yet play a vital role in connecting different sections of a network.

Curado et al. [47] introduced an innovative metric for pinpointing influential spreaders within intricate networks. Their approach integrates a random walk methodology with an effective distance gravity model, enabling the incorporation of local, global, and dynamic node interaction data. By amalgamating insights from effective distances within a gravity model framework, their metric offers enhanced efficacy in identifying pivotal nodes within complex networks. Expanding on this concept, Qiu et al. [48] also proposed a similar method to identify influential spreaders in a complex network based on the local and global position of the node. The degree centrality is used to measure the local influence, while the clustering coefficient is used to measure the global influence of nodes through the k-shell method. Then, it calculates the weight of local and global influence to define the influential node and obtain the importance of each node. Another similar approach proposed by Berberler [49] is called global and local structure. The proposed method not only depends on node influence to identify influential nodes in the network but also depends on how a node shares relation with other nodes. The basic idea of this method mainly depends on node connectivity and its location in the network.

Recently, deep learning approaches have been used to identify influential spreaders in networks, and they have made significant progress in this field. Bhattacharya et al. [13] proposed a novel deep learning framework named DeepInfNode for identifying influential spreaders with topological structure in graphs using Graph Convolutional Networks (GCN). The proposed framework predefines node neighbors in networks through Breadth-First Search (BFS) in order to learn hidden predictive signals before incorporating them into the learning layers. The framework analyzes both node properties and the shortest distances between nodes to identify influential nodes in the network. The top 10% most significant nodes in the network are considered influential, while the others are considered less significant.

3. Preliminaries

Considering an unweighted complex network $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, where V and E denote the set of nodes and edges in the network respectively. Interchangeably, we use n and m to denote the number of nodes and edges, respectively. The distance, denoted by $d(v, u)$ between nodes u and v corresponds to the number of edges in the shortest path connecting them. The eccentricity of a node v, denoted by $\mathrm{ecc}(v)$, is the largest distance from that node v to any other node, i.e., $\mathrm{ecc}(v) = maxu \in V\ d(v, u)$. For a node v of G, $N(v\ ) = \{u\ \in \mathrm{V}: uv\ \in \mathrm{E}\}$ is called the neighborhood of v . Peripheral nodes in networks are nodes that are located at the fringes or peripheries of the network structure. These nodes often have fewer connections compared to nodes in the core of the network and play less influential roles in information dissemination or network dynamics. Formally, peripheral nodes are nodes with the high eccentricity values, that is, using Breadth-First Search (BFS) algorithm from a node v, peripheral nodes are nodes at the maximum distance from v.

3.1. Degree Centrality

Degree centrality is a network centrality measure used to assess the relative importance of a node within a network. It is considered a local measure because it only considers the immediate connections (neighbors) of a particular node. In simpler terms, degree centrality reflects the number of direct connections a node possesses within the network structure. It is defined as the following:

$$Cd(v) = |N(v)| \tag{1}$$

3.2. Betweenness centrality

Betweenness centrality (BC) [9] measures the extent to which a node serves as a bridge for communication between other nodes in a network. It quantifies the number of shortest paths passing through a particular node, relative to the total number of shortest paths between all pairs of nodes in the network. It is defined as follows:

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

where σst is the total number of shortest paths from node s to node t and $\sigma st(v)$is the number of shortest paths from node s to node t that pass-through node v.

3.3. Closeness Centrality

The closeness centrality [10] quantifies how close a node is to all other nodes in the network, reflecting its ability to efficiently interact with others. It is defined as the reciprocal of the average shortest path length from the node to all other nodes in the network. It is defined as follows:

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

where $d(u, v)$is the topological distance between nodes u and v.

3.4. VoteRank Centrality

VoteRank centrality, introduced by Zhang et al. [27], leverages a voting scheme to identify influential nodes within a network. Every vertex $v \in \mathrm{V}$ is associated with a tuple (su, vau) where su denotes the voting score of vertex u and vau indicates the voting ability of vertex u. The voting score su, which is obtained from its adjacent neighbors, can be computed by adding the voting ability of all its neighbors, i.e.

$$Su = \sum_{i \in \mathrm{N(u)}} vai \tag{4}$$

4. The Proposed Method

In this section, we introduce a systematic sampling technique to strategically select a subset of nodes from the network for the purpose of identifying influential nodes. Specifically, our method aims to exclude a substantial portion of nodes deemed unlikely to exhibit influence. As previously noted, conventional centrality measures compute some scores for all nodes in the network and then rank them by their score which is prohibitive for large-scale networks. In contrast, our proposed approach streamlines the process by focusing solely on ranking nodes within our sampling set. Leveraging the betweenness centrality algorithm as outlined by Brandes [50], we expedite the estimation of betweenness scores for nodes within our sampling set, and then rank them based on their scores. Observations suggest that influential nodes often reside at the network's core. Accordingly, our method expeditiously identifies peripheral nodes by executing a limited number of breadth-first search (BFS) algorithms to identify a set P of peripheral nodes that exhibit considerable spatial separation. Subsequently, we initiate BFS from each node u in P and exclude all leaf nodes from the BFS tree, thereby refining our selection of influential nodes. It is important to note that prior to these steps, we repeatedly remove all nodes with a degree of one. Particularly, our method consists of three main steps:

Step 1: Identifying a set of peripheral nodes P with considerable spatial separation.

1. Select an arbitrary vertex z from the network.

2. Perform a breadth-first search BFS starting from vertex z to determine a vertex x at the greatest distance from z. Node x is designated as a peripheral node and added to the list P.

3. To ensure substantial spatial separation between subsequent peripheral nodes, execute BFS from vertex x to identify a vertex y not yet included in P, positioned at the greatest distance from x. Add y to the list P. Repeat this process for t iterations.

Step 2: Determining the sampling set S of nodes.

1. Initially, include all nodes into the sampling set $S = G.V$.

2. For each node u in peripheral node set P, execute a BFS starting from u to identify all leaf nodes L from the BFS tree.

3. Remove all nodes in L from the sampling set S.

Step 3: Rank the nodes in S.

1. Estimate the betweenness centrality for each node $u \in S$ using the accumulation algorithm presented by Brandes [50].

2. Return a ranking list of all nodes in S.

Let l be the size of our sampling list S. Thus, the running time of our sampling algorithm is $O\ (lm)$. Indeed, we will show later in Section 6, that the size of our sampling set is very small compared to the whole number of nodes in the network.

Algorithm1: Pseudo code of our Sampling Method

Input: A connected graph $G = (V, E)$, an integer t.

Output: A set S of influential nodes.

```
Remove all nodes with degree one and repeat until no node with degree one exists.
$S = G.V$ # initially set the sampling list to include all nodes.
$P = [\,]$ # A set of peripheral nodes.
$x$= an arbitrary vertex in $V$
$\boldsymbol{for}\ i: = 1\ to\ t\ do$
    Run a $BFS$ at $x$,and let $y$ be a vertex at the largest distance from $x$ that is not yet included into $P$.
    Append $y$ into the peripheral list $P$.
    $x = y$
end
$\boldsymbol{for\ each}\ u$ in $P$ do
    Run a $BFS$ at u and return a list $L$ of leaf nodes.
    Remove all nodes $w \in L$ from $S$
$\boldsymbol{end}$
$R = betweenness(G, sampling = S)$
$\boldsymbol{return}\ R.$
```

5. SIR Model and Performance Metrics

In this section, we will describe the SIR model and the metrics used to see the performance of our sampling approach.

5.1. SIR Model

To assess the effectiveness of various method in capturing influential nodes within networks, we employ the Susceptible-Infected-Recovered (SIR) model [51] as a standard evaluation framework which has been frequently used for this purpose in the literature [27][29][33][52]. The SIR model partitions cases into three statuses [51]: (i) Susceptible (S), representing individuals not yet infected by the disease; (ii) Infected (I), denoting the infected cases that are capable of spreading the disease; and (iii) Recovered (R), indicating previously infected individuals who have recovered and gained immunity. Initially, a node under examination is designated as infected, and at each time step, infected nodes randomly infect susceptible neighbors at a spreading rate λ (ranging from 0 to 1), which is also called the infection rate. The number of infected nodes at a certain time step is the ability of the initially infected node to influence others, also called node infection ability.

Subsequently, infected nodes may be removed (either deceased or recovered with immunity) with a probability of β as a recover rate, set to 1 without loss of generality. The dynamic process continues until no further nodes can be infected.

5.2. Performance Metrics

To assess the performance of the proposed approach along with other methods such as betweenness, closeness, degree, and voteRank centralities for identifying influential nodes, we use the following metrics that have been presented in [27]. The first two metrics are based on spreading scale under SIR model, while the third one is based on structural properties of selected influential nodes.

(i) Infection Scale $F(t)$: Throughout the information diffusion process governed by the SIR model, the quantities of infected and recovered nodes dynamically fluctuate over time within the system. At any given moment time t, the infection scale $F(t)$ denotes the total number of infected nodes and recovered nodes at time t. This metric serves as a pivotal gauge of the efficacy of the spreader selection algorithm, offering insights into the extent of information dissemination within the network over time, emanating from the designated seed nodes. The following equation calculates the Infection scale, $F(t)$:

$$F(T) = N_I(t) + N_R(t) \qquad (5)$$

where $N_I\ (t), N_R(t)$ represents the number of infected, recovered nodes at time t, respectively.

(ii) Final infected scale $F(t_c)$: Expresses the affected scale when stable state is reached. $F(t_c)$ enumerates all those nodes that became infected and then recovered at time tc, where tc represents the final time when there is no infected node exits in network. The following equation is used to calculate the final infected scale, $F(t_c)$:

$$F(t_c) = N_R(t) \qquad (6)$$

where $N_R(t)$ denotes the number of recovered nodes at time t.

(iii) Average distance between spreaders L_s: It is crucial that the spreader nodes distribute strategically across diverse parts of the network. When spreader nodes cluster together, certain network parts may be left unaffected. Maximizing the distance between selected spreaders enhances the potential for broader information dissemination and coverage. Hence, we employ the average shortest path length Ls among the selected spreaders S, which is formally defined as the following.

$$L_s = \frac{1}{|S|(|S|-1)} \sum_{\mathrm{u,v \in S, u \neq v}} d(u,v) \qquad (7)$$

6. Experimental Results

6.1. Datasets

To examine our proposed sampling method, a comparative analysis will be conducted alongside established network centrality measures including degree, closeness, betweenness, and voteRank centralitities, using different network datasets of a variety of sizes, structural properties, and domains. All datasets considered are listed either in SNAP [53] or Konect [54] project. All of the networks are treated as undirected, unweighted and we only considered the largest connected components.

Table 1. Table 1: Network datasets and their parameters: the number of nodes n; number of edges m, and the number of nodes in our sampling method S.

Network	***N***	***M***	***S***
hamster	2000	16098	239
p2p-Gnutella08	6299	20776	578
as-733	6474	13895	315
PGP	10680	24316	1327
musae-wiki	11631	170918	334
CA-AstroPh	17903	197031	1443
musae-facebook	22740	171002	1917
loc-brightkite	56739	212945	3978

The details of the datasets are listed in Table 1 along with the number of nodes in our sampling list S generated by our sampling method 1. It is noteworthy that for hamster, p2p-Gnutella08, as-733 and musae-wiki networks, we selected 5 peripheral vertices, corresponding to t = 5 in Algorithm 1. Conversely, for the remaining networks, we selected 10 peripheral nodes, that is t = 10.

6.2. Evaluation

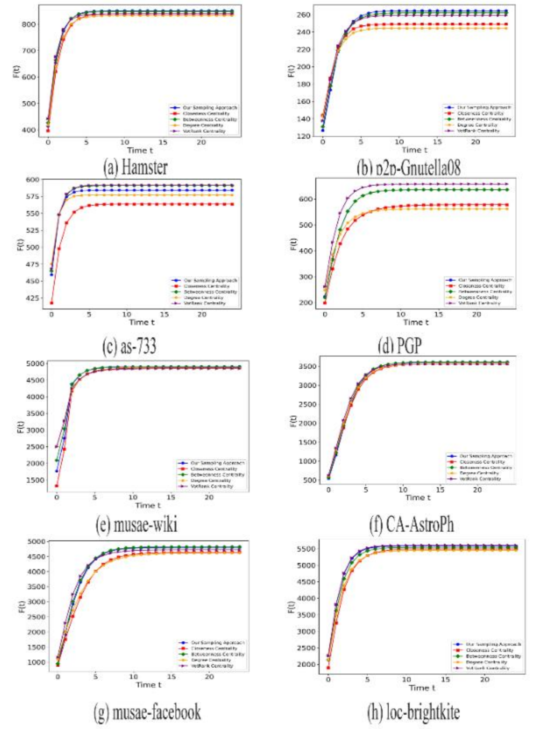
Utilizing the SIR model and the metrics detailed in Section 5.1 and Section 5.2, respectively, we evaluate the performance of our sampling method compared with the results achieved by other methods including closeness centrality, betweenness centrality, degree centrality, and voteRank centrality. This evaluation is conducted on eight real-world networks of varying application domains and sizes, as listed in Table 1. To avoid the randomness involved in the SIR model, the simulations are repeated multiple times, and the final results represent the average outcome across all iterations.

Figure 1 illustrates the evolution of the infected scale $F(t)$ across eight networks, employing different methodologies, with an infection rate of $\lambda = 0.05$ and $p \ = \ 0.005$, where p represents the ratio of initial spreaders. Figure 1 reveals that our proposed methodology for identifying initial spreaders facilitates robust information dissemination, notably impacting a larger scale compared to alternative methods across networks such as hamster, p2p-Gnutella08, musae-wiki, Ca-AstroPh, Musae-facebook, and loc-brighkite. However, in the as-733 network, VoteRank and betweenness centralities exhibit marginal superiority over our approach, although our methodology still outperforms other methods. Similarly, in the PGP network, VoteRank centrality displays a slight advantage over our method and betweenness centrality, albeit our technique demonstrates superior outcomes compared to closeness and degree centralities.

Figure 2 provides additional insights into the impact of varying the number of initial spreaders on the Final Infected Scale. The graph displays the results of $F(t_c)$ plotted against spreader fraction, spanning from 0.004 to 0.009 across all networks except for hamster, a smaller network, where the range extends from 0.01 to 0.06. Notably, our method demonstrates superior performance over other methodologies in musae-wiki and CA-AstroPh networks. Moreover, it is observed that centrality measures such as degree, closeness, and VoteRank exhibit inconsistent performance with changes in the initial spreader fraction. In hamster and musae-facebook networks, betweenness centrality marginally outperforms our sampling method, while our approach yields superior outcomes compared to other methodologies. Conversely, in the locbrightkite network, both our method and VoteRank achieve similar results, surpassing other approaches. In the remaining network scenarios, betweenness and VoteRank centrality exhibit slightly better performance compared to our method.

Figure 3 presents the Final Infected Scale $F(t_c)$ across various infection rates λ and methods on networks. Notably, our proposed methodology and betweenness centrality exhibit the capability to achieve a broader spread scale compared to alternative methods across diverse values of λ, particularly evident in networks such as hamster, musae-wiki, CA-AstroPh, musae-facebook, and loc-brighlite. In the remaining networks, VoteRank slightly outperforms our method.

Figure 4 illustrates the values of L_s associated with source spreaders identified by various methodologies across varying scales of spreaders. Notably, our sampling technique demonstrates notable superiority in networks such as hamster, musae-wiki, and CA-AstroPh, exhibiting larger L_s values compared to all other methods. In the case of Ca-AstroPh, Musae-facebook, as-733, and p2p-Gnutella networks, our sampling method outperforms all others, with the exception of VoteRank centrality, which yields larger Lsvalues. Conversely, in the loc-brightkite network, VoteRank centrality attains larger L_s values surpassing all other methodologies, while degree centrality marginally outperforms our method.



(a) Hamster

(b) p2p-Gnutella08

(c) as-733

(d) PGP

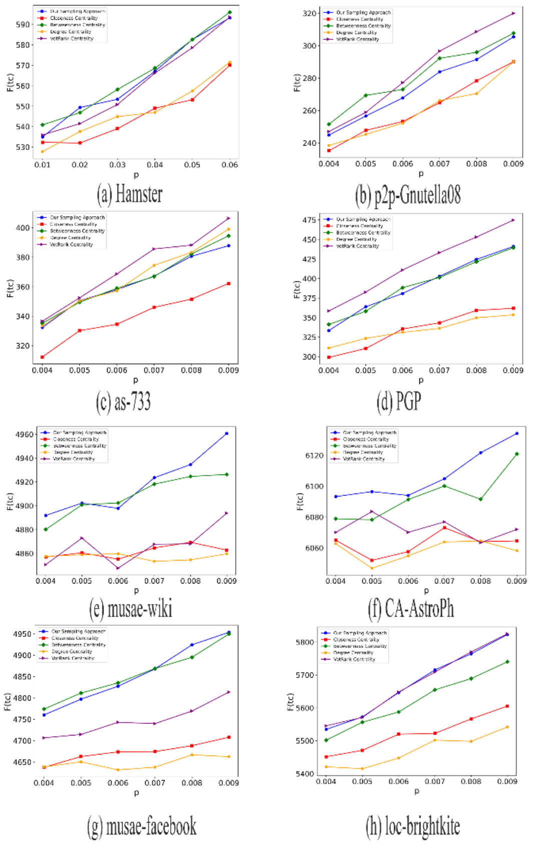
(e) musae-wiki

(f) CA-AstroPh

(g) musae-facebook

(h) loc-brightkite

Fig. 1. The infected scale F(t) (t = 25) on all networks under different methods, where the infection rate $\lambda \ = 0.05$ and with ratio of initial infected nodes $p \ = \ 0.005$. The results are averaged over 100 independent runs.



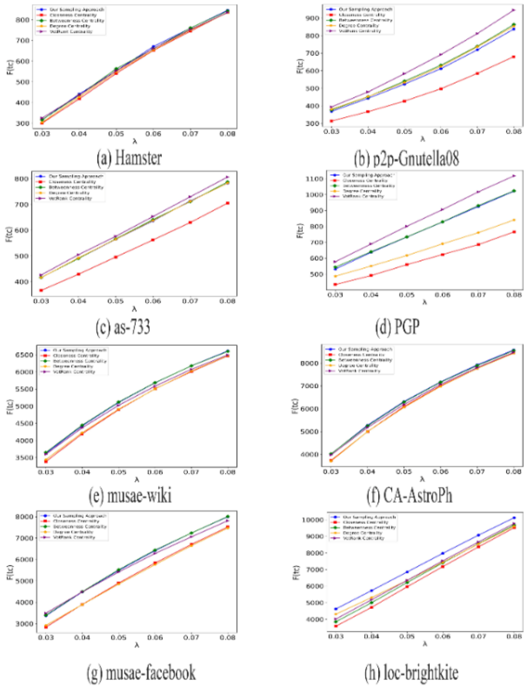
(a) Hamster (b) p2p-Gnutella08

(c) as-733 (d) PGP

(e) musae-wiki (f) CA-AstroPh

(g) musae-facebook (h) loc-brightkite

Fig. 2. The final affected scale $F(t_c)$ on all networks under different methods, where the infection rate $\lambda = 0.05$. p is the ratio of initial infected nodes. The results are averaged over 100 independent runs.



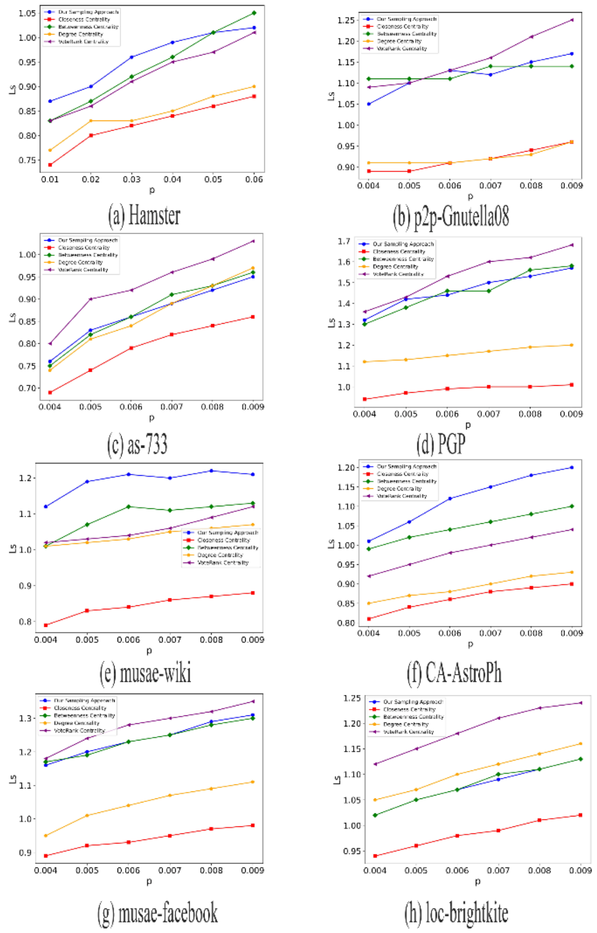
(a) Hamster (b) p2p-Gnutella08

(c) as-733 (d) PGP

(e) musae-wiki (f) CA-AstroPh

(g) musae-facebook (h) loc-brightkite

Fig. 3. The final affected scale $F(t_c)$ with the different infection rate λ on all networks under different method.



(a) Hamster (b) p2p-Gnutella08

(c) as-733 (d) PGP

(e) musae-wiki (f) CA-AstroPh

(g) musae-facebook (h) loc-brightkite

Fig. 4. The average shortest path length L_s of nodes selected by different methods. p is the ratio of initial infected nodes.

Conclusion

Identifying influential nodes within complex networks is a cornerstone for comprehending the intricate dynamics of information dissemination. While numerous state-of-the-art algorithms exist, they are often hindered by two significant drawbacks: (1) ranking all nodes in the network based on some calculated scores, which is unnecessary for many applications, and (2) high computational complexity. This paper proposes a novel methodology that addresses these limitations. We introduce a simple yet powerful systematic sampling approach coupled with an approximation technique for estimating the betweenness centrality of strategically selected nodes. This approach effectively identifies influential nodes within complex networks with demonstrably improved computational efficiency. To evaluate our proposed method, we compared its performance against centrality measures, including degree centrality, closeness centrality, betweenness centrality, and VoteRank centrality. We used the Susceptible-Infected-Recovered (SIR) information diffusion model to analyze key metrics like infection scale, final infected scale over time, and average distance between spreaders across real-world network datasets. Our empirical validation demonstrated that the proposed method identified the influential nodes accurately, achieving comparable results to existing algorithms while significantly reducing computational time.

References

[1] K. Taha, "Static and Dynamic Community Detection Methods That Optimize a Specific Objective Function: A Survey and Experimental Evaluation," *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 98330–98358, 2020. doi: 10.1109/ACCESS.2020.2996595.

[2] F. Zhu *et al.*, "A context-aware trust-oriented influencers finding in online social networks," in *2015 IEEE International Conference on Web Services*, IEEE, 2015, pp. 456–463.

[3] A. Sheikhahmadi and M. A. Nematbakhsh, "Identification of multi-spreader users in social networks for viral marketing," *J. Inf. Sci.*, vol. 43, no. 3, pp. 412–423, 2017.

[4] J. Gu, L. C. Abroms, D. A. Broniatowski, and W. D. Evans, "An investigation of influential users in the promotion and marketing of heated tobacco products on Instagram: a social network analysis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 3, p. 1686, 2022.

[5] X. Wei, J. Zhao, S. Liu, and Y. Wang, "Identifying influential spreaders in complex networks for disease spread and control," *Sci. Rep.*, vol. 12, no. 1, p. 5550, 2022.

[6] A. Susarla, J.-H. Oh, and Y. Tan, "Influentials, imitables, or susceptibles? Virality and word-of-mouth conversations in online social networks," *J. Manag. Inf. Syst.*, vol. 33, no. 1, pp. 139–170, 2016.

[7] F. J. Arenas-Márquez, M. del R. Martínez-Torres, and S. L. Toral, "How can trustworthy influencers be identified in electronic word-of-mouth communities?," *Technol. Forecast. Soc. Change*, vol. 166, p. 120596, 2021.

[8] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Soc. Netw. Crit. concepts Sociol. Londres Routledge*, vol. 1, pp. 238–263, 2002.

[9] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[10] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[11] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Phys. a Stat. Mech. its Appl.*, vol. 391, no. 4, pp. 1777–1787, 2012.

[12] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[13] R. Bhattacharya, N. K. Nagwani, and S. Tripathi, "Detecting influential nodes with topological structure via Graph Neural Network approach in social networks," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2233–2246, 2023.

[14] K. Hajarathaiah, M. K. Enduri, S. Anamalamudi, and A. R. Sangi, "Algorithms for finding influential people with mixed centrality in social networks," *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 10417–10428, 2023.

[15] S. M. V Reddy, D. Annapurna, and A. Narasimhamurthy, "Influence node analysis based on neighborhood influence vote rank method in social network," *Sci. Temper*, vol. 14, no. 04, pp. 1537–1543, 2023.

[16] F. Bloch, M. O. Jackson, and P. Tebaldi, "Centrality measures in networks," *Soc. Choice Welfare*, vol. 61, no. 2, pp. 413–453, 2023.

[17] Q. Liu *et al.*, "An influence propagation view of pagerank," *ACM Trans. Knowl. Discov. from Data*, vol. 11, no. 3, pp. 1–30, 2017.

[18] Y. Wang, G. Yan, Q. Ma, Y. Wu, and D. Jin, "Identifying influential spreaders on weighted networks based on ClusterRank," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, IEEE, 2017, pp. 476–479.

[19] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS One*, vol. 6, no. 6, p. e21202, 2011.

[20] G. Huang, J. Liu, X. Chen, and J. Ren, "A New Method of Identifying Influential Nodes in Complex Software Network Based on LeaderRank," 2016.

[21] G. Maji, S. Mandal, and S. Sen, "A systematic survey on influential spreaders identification in complex networks with a focus on K-shell based techniques," *Expert Syst. Appl.*, vol. 161, p. 113681, 2020.

[22] M. Gupta and R. Mishra, "Spreading the information in complex networks: Identifying a set of top-N influential nodes using network structure," *Decis. Support Syst.*, vol. 149, p. 113608, 2021.

[23] Z. Zhao, D. Li, Y. Sun, R. Zhang, and J. Liu, "Ranking influential spreaders based on both node k-shell and structural hole," *Knowledge-Based Syst.*, vol. 260, p. 110163, 2023.

[24] M. E. J. Newman, "Analysis of weighted networks," *Phys. Rev. E*, vol. 70, no. 5, p. 56131, 2004.

[25] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: the role of clustering," *PLoS One*, vol. 8, no. 10, p. e77455, 2013.

[26] Y. Liu, M. Tang, T. Zhou, and Y. Do, "Identify influential spreaders in complex networks, the role of neighborhood," *Phys. A Stat. Mech. its Appl.*, vol. 452, pp. 289–298, 2016.

[27] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Sci. Rep.*, vol. 6, no. 1, p. 27823, 2016.

[28] S. Kumar and B. S. Panda, “Identifying influential nodes in Social Networks: Neighborhood Coreness based voting approach,” *Phys. A Stat. Mech. its Appl.*, vol. 553, p. 124215, 2020.

[29] P. Liu, L. Li, S. Fang, and Y. Yao, “Identifying influential nodes in social networks: A voting approach,” *Chaos, Solitons & Fractals*, vol. 152, p. 111309, 2021.

[30] L. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, “Identifying influential spreaders in complex networks based on gravity formula,” *Phys. A Stat. Mech. its Appl.*, vol. 451, pp. 205–212, 2016.

[31] Z. Li *et al.*, “Identification of a promoter element mediating kisspeptin-induced increases in GnRH gene expression in sheep,” *Gene*, vol. 699, pp. 1–7, 2019.

[32] X. Yang and F. Xiao, “An improved gravity model to identify influential nodes in complex networks based on k-shell method,” *Knowledge-Based Syst.*, vol. 227, p. 107198, 2021.

[33] S. Wang, Y. Du, and Y. Deng, “A new measure of identifying influential nodes: Efficiency centrality,” *Commun. Nonlinear Sci. Numer. Simul.*, vol. 47, pp. 151–163, 2017.

[34] M. M. Tulu, R. Hou, and T. Younas, “Identifying influential nodes based on community structure to speed up the dissemination of information in complex network,” *IEEE access*, vol. 6, pp. 7390–7401, 2018.

[35] J. Bae and S. Kim, “Identifying and ranking influential spreaders in complex networks by neighborhood coreness,” *Phys. A Stat. Mech. its Appl.*, vol. 395, pp. 549–559, 2014.

[36] M. Li, R. Zhang, R. Hu, F. Yang, Y. Yao, and Y. Yuan, “Identifying and ranking influential spreaders in complex networks by combining a local-degree sum and the clustering coefficient,” *Int. J. Mod. Phys. B*, vol. 32, no. 06, p. 1850118, 2018.

[37] A. Zareie, A. Sheikhahmadi, and M. Jalili, “Influential node ranking in social networks based on neighborhood diversity,” *Futur. Gener. Comput. Syst.*, vol. 94, pp. 120–129, 2019.

[38] A. Zareie, A. Sheikhahmadi, M. Jalili, and M. S. K. Fasaei, “Finding influential nodes in social networks based on neighborhood correlation coefficient,” *Knowledge-based Syst.*, vol. 194, p. 105580, 2020.

[39] A. Namtirtha, A. Dutta, and B. Dutta, “Identifying influential spreaders in complex networks based on kshell hybrid method,” *Phys. A Stat. Mech. its Appl.*, vol. 499, pp. 310–324, 2018.

[40] A. Namtirtha, A. Dutta, and B. Dutta, “Weighted kshell degree neighborhood: A new method for identifying the influential spreaders from a variety of complex network connectivity structures,” *Expert Syst. Appl.*, vol. 139, p. 112859, 2020.

[41] C. Guo, L. Yang, X. Chen, D. Chen, H. Gao, and J. Ma, “Influential nodes identification in complex networks via information entropy,” *Entropy*, vol. 22, no. 2, pp. 1–19, 2020, doi: 10.3390/e22020242.

[42] X. Xu, C. Zhu, Q. Wang, X. Zhu, and Y. Zhou, “Identifying vital nodes in complex networks by adjacency information entropy,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-59616-w.

[43] F. Liu, Z. Wang, and Y. Deng, “GMM: A generalized mechanics model for identifying the importance of nodes in complex networks,” *Knowledge-Based Syst.*, vol. 193, p. 105464, 2020.

[44] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, and T. Zhou, “Identifying influential spreaders by gravity model,” *Sci. Rep.*, vol. 9, no. 1, p. 8387, 2019.

[45] J. Zhao, Y. Wang, and Y. Deng, “Identifying influential nodes in complex networks from global perspective,” *Chaos, Solitons & Fractals*, vol. 133, p. 109637, 2020.

[46] Y. Liu, X. Wei, W. Chen, L. Hu, and Z. He, “A graph-traversal approach to identify influential nodes in a network,” *Patterns*, vol. 2, no. 9, 2021.

[47] M. Curado, L. Tortosa, and J. F. Vicent, “A novel measure to identify influential nodes: return random walk gravity centrality,” *Inf. Sci. (Ny).*, vol. 628, pp. 177–195, 2023.

[48] L. Qiu, Y. Liu, and J. Zhang, “A New Method for Identifying Influential Spreaders in Complex Networks,” *Comput. J.*, vol. 67, no. 1, pp. 362–375, 2024.

[49] M. E. Berberler, “Global and local structure-based influential nodes identification in wheel-type networks,” *Numer. Methods Partial Differ. Equ.*, vol. 40, no. 1, p. e22709, 2024.

[50] U. Brandes, “A faster algorithm for betweenness centrality,” *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.

[51] R. M. May, *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

[52] G. Xu and C. Dong, “CAGM: A communicability-based adaptive gravity model for influential nodes identification in complex networks,” *Expert Syst. Appl.*, vol. 235, p. 121154, 2024.

[53] J. Leskovec and A. Krevl, “Stanford large network dataset collection (snap),” *URL http//snap. stanford. edu/data/index. html*, 2010.

[54] J. Kunegis, “Konect: the koblenz network collection,” in *Proceedings of the 22nd international conference on world wide web*, 2013, pp. 1343–1350.