IJICTR
International Journal of Information & Communication Technology Research

Volume 14- Number 1 – 2022 (38 -47)

A Community-Based Method for Identifying Influential Nodes Using Network Embedding

Narges Vafaei
Department of Computer Engineering
Faculty of Engineering
Alzahra University
Tehran, Iran
na.vafaei@student.alzahra.ac.ir

Mohammad Reza Keyvanpour*
Department of Computer Engineering
Faculty of Engineering
Alzahra University
Tehran, Iran
keyvanpour@Alzahra.ac.ir

Received: 8 February 2022 – Revised: 20 February 2022 - Accepted: 9 March 2022

Abstract— **People's influence on their friends' personal opinions and decisions is an essential feature of social networks. Due to this, many businesses use social media to convince a small number of users in order to increase awareness and ultimately maximize sales to the maximum number of users. This issue is typically expressed as the influence maximization problem. This paper will identify the most influential nodes in the social network during two phases. In the first phase, we offer a community detection approach based on the Node2Vec method to detect the potential communities. In the second phase, larger communities are chosen as candidate communities, and then the heuristic-based measurement approach is utilized to identify influential nodes within candidate communities. Evaluations of the proposed method on three real datasets demonstrate the superiority of this method over other compared methods.**

Keywords: social network mining; influence analysis; influence maximization problem; influential nodes

Article type: Research Article

Publisher: ICT Research Institute

[DOI: 10.52547/itrc.14.1.38]

I. INTRODUCTION

With the expansion of various social networks such as Facebook, Twitter, and Telegram, new methods have been created to discover and disseminate information. Influence analysis in social networks is one of the essential sub-branches of social network mining. Social influence occurs when a person's beliefs, feelings, and behaviors are influenced by others, whether consciously or unconsciously[6]. Businesses utilize social media to convince a smaller number of users to increase awareness and maximize sales to a maximum number of customers. This issue is typically formulated as the influence maximization problem. Influence maximization in social networks can offer magnificent advantage to many functional applications, such as viral marketing[7], solving social and political issues[8], rumor and inappropriate news control[9], plus medical and bioinformatics problems[10].

As shown in "Fig. 1", the general framework for the influence maximization problem consists of four

* Corresponding Author

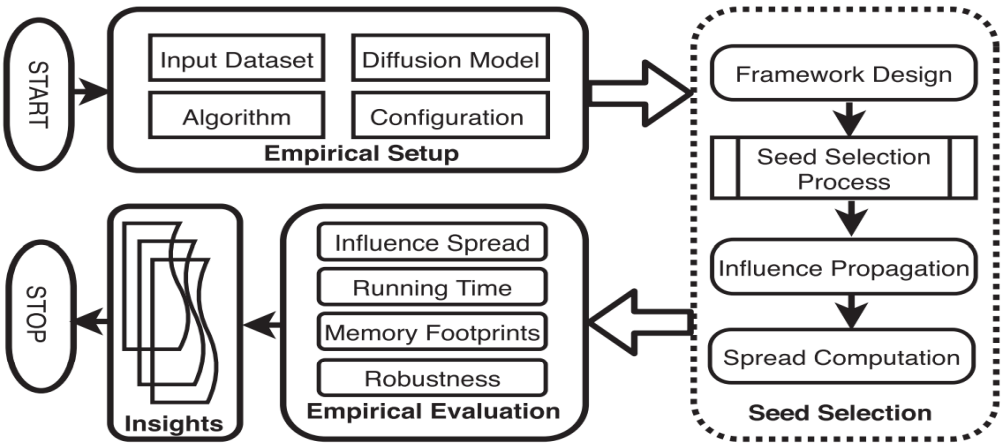

Fig 1. The basic framework of the IM problem[4]

steps[4]. The required information is collected in the empirical setup section. The second phase, or seed selection, is the main part of this problem and seeks to find influential nodes. In the evaluation section, the solution is estimated using evaluation criteria, and then, in the insight section, it is analyzed.

Information propagation on social networks is very complex. This complexity is due to the significant number of users and the essential need to consider the different requirements, restrictions, and the random character of the social network structure. Therefore, information diffusion models are utilized due to the complex structure and mapping of information propagation in social networks. The Independent Cascade (IC) model[11] and Linear Threshold (LT) model[12] are the substantial information diffusion models.

It has been proven that the influence maximization problem below the two diffusion models, LT and IC, is NP-hard [11] .Greedy algorithms are an approximation solution with a factor of $(1 - 1 / e)$ to the problem. However, they are very inefficient due to the simulation of the propagation process with the Monte Carlo method. Other approaches to solving the influence maximization problem include the heuristic algorithms, which have less time complexity than greedy, but guaranteeing approximation is challenging.

The community-based approach is another existing approach that has been more successful than the two mentioned approaches. In general, this approach supports parallelism by considering the non-overlapping construction of the community[13]. However, the accuracy of the approach, which depends on the structure of the community, can be a challenge.

In this paper, we propose the "CNE-IM" method, a Community-based method using a Network Embedding algorithm, for Influence Maximization. In summary, the contributions of our work are as follows:

- The high-quality features of each node were obtained as feature vectors using the node2vec, a network embedding method.
- By applying the K-Means clustering method to the obtained feature vectors, potential communities in the network are identified.
- Since more nodes are connected in larger communities, and the selection of influential nodes in these communities' spreads information across the network, larger communities are selected as candidate communities.
- The proposed algorithm CNE-IM is tested on real social networks, and the results show that this algorithm can achieve better results than other similar methods in the spread of influence.

The rest of this paper is organized in the following manner, Section II reviews previous work. Section III explains the prerequisites of the problem. Section IV examines the proposed method thoroughly. Section V evaluates the performance of the proposed method on two real networks and discusses the results. Finally, the paper ends in Section VI with a conclusion.

II. Related Works

Existing approaches to finding influential nodes in social networks can be categorized into simulation-based, heuristic, meta-heuristic, community-based, and hybrid approaches. This category is shown in "Fig. 2".

The simulation-based approach's main objective is to perform Monte Carlo simulations to assess the influence of $I_M(S)$ for each S seed set. Proving that the influence maximization problem is NP-hard, Kamp et al. [11] presented a greedy approach that approximates the solution by $(1 - 1/e)$.

The greedy method had a high time complexity, and for this reason, some researchers, such as Leskovec et al.[14], presented the Cost Effective Lazy Forward (CELF) algorithm, and Goyal et al.[15] introduced the CELF++ algorithm, which has addressed the execution time dependence issue on the number of graph nodes. The reported results show that CELF can accelerate the computational process up to 700 times compared to the greedy algorithm in the benchmark data set. CELF++ is an extension of CELF that eliminates unnecessary Monte Carlo simulations during the first step of CELF. The Staticgreedy [16] method proposed by Cheng et al. includes two steps of obtaining the R number of Monte Carlo snapshots and seed selection, providing high

[DOI: 10.52547/itrc.14.1.38]

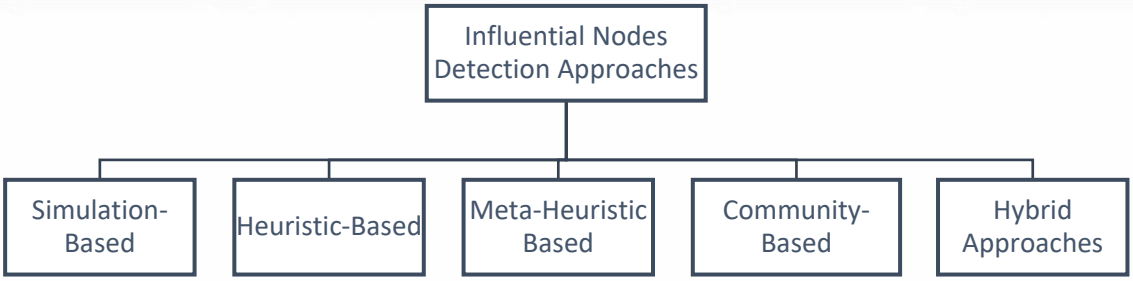

Fig 2. Approaches of influential nodes detection. Adopted from[2].

accuracy and scalability. As an alternative to the Monte Carlo simulation, Borgs et al.[17] presented a reverse reachable sampling technique for estimating influence spreads. In this method, first, a hypergraph is created from the input network, and then the seed nodes are selected repeatedly based on the highest degree. This method is mostly theoretically oriented and lacks any practical application. The Sketch-Based Influence Maximization (SKIM) [18], Two-phase Influence Maximization (TIM)[19], Influence Maximization via

Martingales (IMM)[20] and Stop-and-Stare Algorithm (SSA)[21] are among the other simulation-based methods.

Heuristic approach methods are faster than simulation-based algorithms; however, due to the instability of performance in various networks, there is no guarantee of accuracy. The Degree Discount [22] method focuses on the idea that once a node has been selected as a seed, its neighbors cannot influence it anymore. Therefore, the seed node is selected, and the degrees of its neighbors are reduced by one. Chen et al. [23]proposed a method based on the Local Directed Acyclic Graph (LDAG) models. In the first step, LDAGs are calculated, and in the second step, the seed set is selected by the greedy algorithm. The SimPath [24] algorithm uses the idea of the CELF method and utilizes path counting techniques instead of Monte Carlo simulations. The Shapley Value-Based Discovery of Influential Nodes (SPIN)[25], IRIE[26], ASIM[27], and EaSyIm[28] methods are among the other heuristic methods.

Metaheuristic methods often employ evolutionary computations and swarm intelligence techniques. Bucur et al.[29] used a genetic algorithm and showed that a primary genetic operator can estimate an approximate solution to the problem of influence spread in a reasonable run time. Jiang et al.[30] developed a simulated annealing-based algorithm that is 2–3 times faster than heuristic methods. Lotfi et al.[31] proposed a method based on the genetic algorithm to find influential nodes across dynamic networks. In this method, several graphs are modeled in a specified timestamp and then find the influential nodes in each of these graphs. Ma et al.[32] proposed an Evolutionary Deep Reinforcement Learning algorithm (EDRL-IM). In this algorithm, the influence maximization problem is first modeled under the deep Q network (DQN), and then this model is developed by combining an Evolutionary Algorithm (EA) and a Deep Reinforcement Learning algorithm (DRL).

Community-based influence maximization algorithms use community detection methods to shrink the network to the level of communities while increasing scalability. Wang et al.[33] proposed the greedy community-based algorithm (CGA) where networks were subdivided into subnetworks and influential users were identified across subnetworks. This method is not appropriate for large data sets. Chen et al.[34] also, used the community-based approach to find influential nodes by presenting the community-based influence maximization (CIM) method. Wilder et al.[35] also proposed a community-based method called Approximating with Random walks to Influence a Socially Explored Network (ARISEN). Li et al. in [36] investigated the maximization of influence in a community-based approach by considering the location. Singh et al.[13] proposed a Community-based Context-aware Influence Maximization (C2IM) method that uses a community-based approach to reduce search disclosure and consider user interest, examining the effectiveness of seeds. Pourkazemi et al.[37] proposed a community-based solution called CNLPSO-SL. In this method, a community detection method called CNLPSO-DE [38] has been used to identify potential communities. Then the most influential nodes have been identified using a semi-local centrality method. YE et al.[39] proposed a community-based method Using network embedding and the clustering algorithm for finding influential nodes appropriate for large-scale networks. This paper also presents a basic community-based approximation algorithm (BCRIM) to find influential people in communities.

Vafaei et al.[40] method can also be mentioned among the researchers that have used network embedding in this issue. They used the Word2Vec method to extract nodes' structural features and then used a heuristic method to find influential nodes.

III. PRELIMINARIES

A. *Influence Maximization*

Definition 1 (influence maximization): The graph $G = (V, E)$ represents an online social network with the sets of nodes V and links E. By defining a probability function $p: E \rightarrow (0,1)$ that indicates the probability of propagation in each bond E, we are looking for K ($K \leqslant |V|$) nodes to maximize network spread. The K nodes are called the diffusion seed. The influence spread in the network, denoted by σ_M, is equal to the number of nodes affected by the set S on

[DOI: 10.52547/itrc.14.1.38]

the diffusion model M. The optimal seed set S^* is defined as (1).

$$S^* = arg \max_{S \subset V, |S|=K} \sigma_M(S) \quad (1)$$

B. Diffusion Models

Diffusion models map the information propagation. If the node is affected in the diffusion stages, it is active, and otherwise, it is inactive. Initially, the seed nodes are involved, and the remaining nodes are inactive. During the diffusion stages, each node tries to activate its neighboring nodes, and these steps will continue until the remaining nodes are activated. The Independent Cascade model (IC) [11] and Linear Threshold model (LT) [12] are the commonly-used information propagation models. In this paper, we utilize an independent cascade model for mapping information propagation.

Definition 2 (Independent Cascade (IC) model): Each activated node has only one opportunity to change the inactive nodes to active within this model. The edges' weight is a number between 0 and 1 and indicates the probability of their activation. The active node tries to activate that neighboring node at $t+1$ with the probability of weight between them.

An example of the information propagation in the IC model is shown in "Fig. 3". The numbers on the edges indicate propagation probability, and the red node indicates the active node.

Definition 3 (Linear threshold (LT) model): For each node, a threshold θ_v is defined, which indicates the node's desire to obtain a new idea from others. A larger threshold value means that the node is less likely to change position. In this method, to activate a node v, the neighbor's edges' total weight must be greater than the threshold of the desired node.

C. Monotonicity and Submodularity

The influence maximization problem under IC and LT diffusion models is NP-hard[11]. The optimal solution can be approximated if the influence function has the properties of submodularity and monotonicity[41]. The function is monotonic if adding more nodes to the seed set does not reduce its influence spread. It is also a submodular function if the marginal gain of influence spread from adding a node to the seed set is at least equal to the marginal gain of adding the same node to the seed superset.

Definition 4 (Monotonicity)[41]: The influence function $\sigma(.)$ has monotonicity property if $\sigma(S) \leq \sigma(S')$ for all $S \subseteq S' \subseteq V$.

Definition 5 (Submodularity)[41]: The influence function $\sigma(.)$ has submodularity property if $\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(S' \cup \{v\}) - \sigma(S')$ for all $S \subseteq S' \subseteq V$ and $v \in V \backslash S'$.

D. Network Embedding

The network embedding's objective is to extract nodes' low-dimensional attributes in a way that preserves the network structure. A vector is created for each node through network embedding those topological and structural features are coded in this vector. In the resulting vector space, node relations can be obtained through the distance between them[42].

Definition 6 (network embedding): The graph $G = (V, E)$ represents an online social network, V and E show the sets of nodes and links, respectively. The network embedding embeds each node $v \in V$ into a low-dimensional space R^d, that is, to learn a mapping function $f_G: V \rightarrow R\hat{}d$, where $d \ll V$ [39].

In the CNE-IM method, we utilize the Node2Vec [43] method to extract features for each node. Node2Vec is a Semi-supervised learning algorithm of network embedding that relies on random walking and deep learning. In this method, first, two search strategies, DFS (Depth First Search) and BFS (Breadth First Search) are used to generate random walks, and a path of direct and indirect neighbors is generated for each node. In order to extract the vector of each user, it utilizes the Skip-Gram[44] model after generating a number of paths for each user. In the Skip-Gram method, the probability of seeing the neighbors of a user depends on the vector of each of them. These steps are shown in "Fig. 4".

Node2Vec method learns the feature of networks as a search-based optimization problem and maintains graph features such as the criteria of centrality and similarity between nodes by considering the direct and quadratic neighbors of a node. There are several advantages to this condition. For example, one can describe search approaches based on a discovery and exploitation transaction and interpret the views learned in a prediction problem[43].

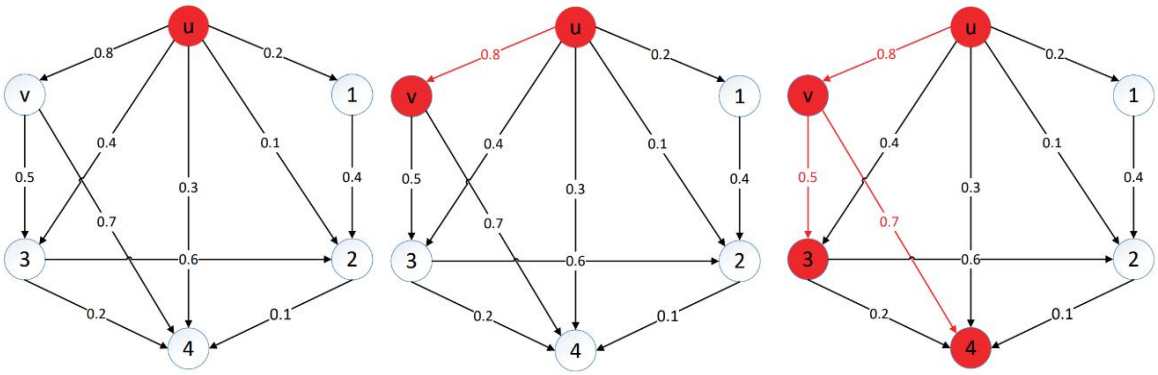

Fig 3. An example of information propagation under IC diffusion model[3]

[DOI: 10.52547/itrc.14.1.38]

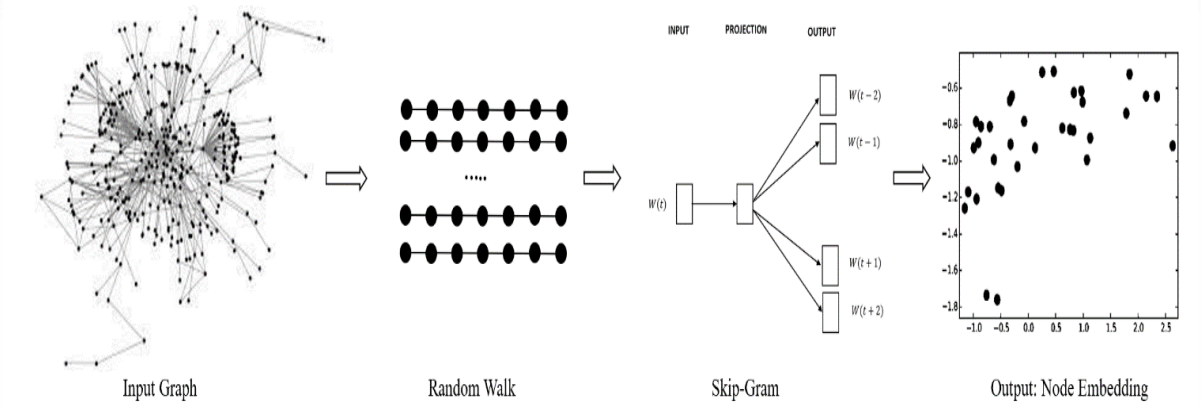

Fig 4. Steps of Node2Vec method to extract the structural feature vector of each node. Adopted from [1].

E. Semi-local Centrality

Centrality measurements are used as a heuristic approach for solving the influence maximization problem. Degree centrality, betweenness centrality, and closeness centrality are among the criteria of centrality. The degree centrality criterion is low-relevant, and the betweenness and closeness centrality also suffer from the time-consuming. The Semi-local centrality [5] criterion can be considered a trade-off between the characteristics of those centrality criteria. It considered both the nearest and the next nearest neighbors for each node and is defined as (2) and (3).

$$Q(u) = \sum_{w \in \Gamma(u)} N(w) \qquad (2)$$

$$C_L(v) = \sum_{w \in \Gamma(u)} Q(u) \qquad (3)$$

Where $\Gamma(u)$ is the set of nearest neighbors of node u and $N(w)$ is the number of the closest and the subsequent nearest neighbors of node w. An example of the semi-local centrality method is shown in "Fig. 5".

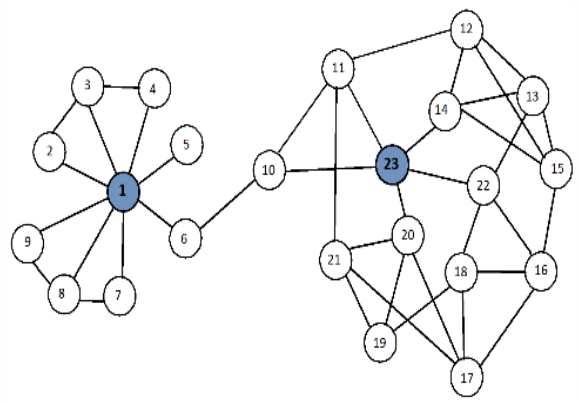

Fig 5. An example network consists of 23 nodes and 40 edges, where node 23 has a more significant influence than node 1, although it has a lower degree than node 1 [5].

IV. PROPOSED METHOD

In this paper, we present a community-based method CNE-IM for identifying influential nodes in social networks. The proposed method includes two phases of community detection and identification of influential nodes.

In the initial phase, by extracting the features of each node by the Node2Vec method and applying a clustering method, communities are identified. Subsequently, in the second phase, candidate communities are selected by applying the size criterion, and a semi-local centrality measure was used for identifying influential nodes within candidate communities. Finally, by applying the independent cascade diffusion model, the information propagation of these nodes is measured. The block diagram of the proposed method is shown in "Fig. 6".

A. Community detection based on network embedding

In this section, we first desire to learn each node's specific feature. For this purpose, we utilize the Node2Vec method. We give our dataset, which is in the form of a social network, as input to the algorithm, and for each node v, we obtain a d dimensional feature vector $[x_1, x_2, \ldots, x_d]$ as output.

After obtaining a vector for each node, we cluster the vectors using the K-Means algorithm. Indeed, by clustering vectors, we place network nodes in a different community. Each cluster represents a community with similar characteristics.

In general, in this section, using the network embedding and the K-Means clustering method, we were able to partition the network using the node's specific feature and detection the communities.

B. Selecting candidate communities

After detecting the potential communities within the network, a criterion must be used to select several organizations, followed by selecting the node with the most influence on candidate communities. Choosing the largest communities can be one of the leading measures for selecting candidate communities because, in larger communities, more nodes are interconnected, and the selection of influential nodes in these

[DOI: 10.52547/itrc.14.1.38]

communities will spread information throughout the network[45]. Therefore, we choose the K largest communities as candidate communities.

C. *Identifying influential nodes within candidate communities via semi-local centrality*

After selecting candidate communities, we are looking to find influential nodes in selected communities in this section. For this purpose, we utilize the Semi-local centrality measure. As stated in section III, the semi-local centrality considers both the nearest and the next nearest neighbors for each node to maximize the influence propagation.

The node with the highest semi-local centrality is selected as the influential node of each candidate community. These influential nodes are our seed nodes. In this way, an effective node is chosen from each community. Due to selecting one node from each candidate community, the number of seed nodes and candidate communities is equal. For the input value K, representing the number of seed nodes, K candidate communities are chosen.

After selecting the seed nodes, they are presented as active nodes to the independent cascade model to map information propagation and calculate influence spread.

As can be seen in Algorithm 1, in line 1, a Node2Vec algorithm is applied to the input social network G, and a d dimensional vector is calculated for each node. A k-Means algorithm is used in line 2 to cluster the vectors obtained. In line 3, the obtained clusters are sorted by size, and k larger clusters in line 4 are selected as candidate communities. Lines 7 and 8 calculate the semi-local centrality for each node in the candidate communities, and the node with the highest semi-local centrality is selected as the influential node for each community. The influential nodes are given to the IC diffusion model in line 11. This model returns the influence spread of influential nodes. So, influential nodes and their influence spread are obtained in the algorithm's output.

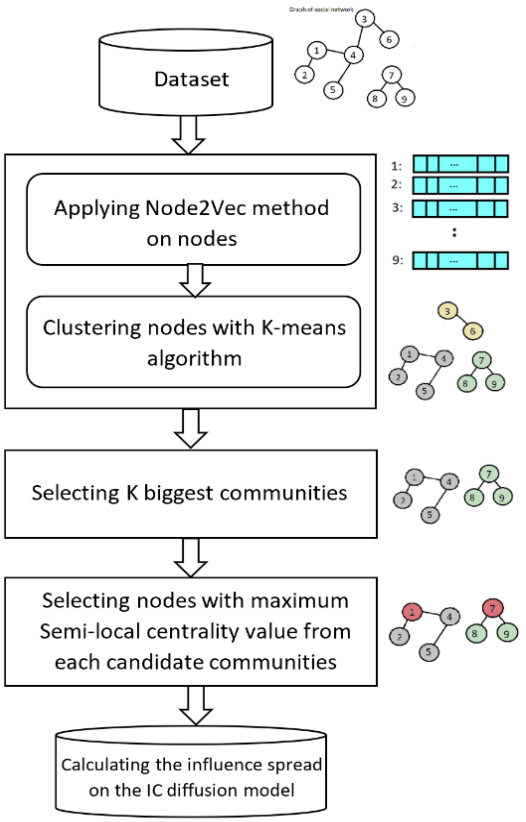

Fig 6. Block diagram of the CNE-IM method

Algorithm 1 CNE-IM

Input:
Social network $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$
Dimention $\boldsymbol{d}$
Size of clusters N
Seed set size K

Output:
Influential nodes and Influence spread of influential nodes

1: *Vector-of-nodes* ← Node2Vec(*G*,*d*)
2: Clusters ← K-Means(Vector-of-nodes , N)
3: *Sorted-Clusters* ←Sorting *Clusters* Based on size
4: *K-Clusters* ← Selecting *K* largest clusters from *Sorted-Clusters*
5: **for** each *c* **in** *K-Clusters*:
6: **for** each node **in** *c*:
7: *Sl*← (*node* , *Semi-local-Centrality(node)*)
8: *Influential-nodes* ← Selecting the node with largest *Semi-local-Centrality* from *SI*
9: **end for**
10: **end for**
11: Compute the influence spread of *Influential-nodes* under IC diffusion model

D. *Monotonicity and submodularity properties in CNE-IM*

This section illustrates the monotonicity and submodularity properties of the proposed CNE-IM method. According to the previous sections, the optimal solution can be approximated if the influence function has the properties of submodularity and monotonicity. The influence function of the proposed CNE-IM method is monotonic because each seed node is selected based on having the most centrality measure within each community, and definitely each node that is selected can activate a number of nodes. Additionally, the proposed method is submodular. Based on Definition 5, the condition for a function to be submodular is that the marginal gain of adding a node to the seed set is at least equal to the marginal gain of adding the same node to the seed superset. Since in the proposed method, the nodes in each community are sorted according to their effectiveness, the first selected nodes in each community activate more nodes. When node $v \in V \backslash S'$ is added to set S $(S \subseteq S' \subseteq V)$, it will increase the influence on the set $(S + v)$ more than the set $(S' + v)$. Accordingly, as the size of the seed set increases, the influence spread margin decreases.

E. *CNE-IM complexity*

CNE-IM time complexity is evaluated by calculating each step's complexity separately and then calculating the total complexity. The Node2Vec algorithm is applied first to the entire network, with time complexity of $O(b^2)$, where b is the graph's branching factor. The K-Means algorithm is then applied to the obtained vectors for each node with a

[DOI: 10.52547/itrc.14.1.38]

time complexity of $O(n^2)$, where n is the number of nodes. The m acquired communities are then sorted in $O(mlog\ m)$ to select k candidate communities. For v nodes in the candidate communities, a semi-local centrality measure with complexity $O(vl^2)$ is calculated, where l is the average degree of the network and then sorted in $O(vlog\ v)$. Overall, the time complexity of the CNE-IM algorithm is $O(b^2 + n^2 + mlog\ m + vl^2 + vlog\ v)$. So, the final time complexity is $O(b^2)$.

V. Result and Discussion

A. Dataset

The CNE-IM was evaluated using Dolphin social network [46] , Netscience [47] and HEP-physics[25, 37, 48] in the experiments. These social networks are widely used in various studies in community detection and influence maximization problems[37, 49-53]. The details of these networks are summarized in Table1.

TABLE I. Dataset Description

-	Number of nodes	Number of edges
Dolphin social network	62	159
Netscience	1589	2742
HEP-physics	8361	15751

B. Method comparison

For demonstrating the proposed algorithm's efficiency and effectiveness, three methods were selected for comparison, which are described as follows.

1) Degree centrality(DC): This method is widely used to calculate the influence of nodes in social networks. In this method, the node with the highest degree among network nodes is selected as the influential node.

2) Ranking random(RN): In this method, to select the specified seed, n nodes are randomly selected from the network and returned as seeds.

3) Semi-local centrality(SL): This method has been explained and subsequently selected as the alternative comparison method to investigate the effect of community detection on the proposed method.

4) Degree discount (DD) : The degree of the seed node's neighbors is reduced by one in this method since it is based on the idea that once a node has been selected as a seed, its neighbors cannot influence it.

C. Evaluation and Results

For measuring the performance of the proposed algorithm, an influence spread evaluation metric employ that uses in influence maximization problem researches[37, 54-56]. It is defined as (4).

$$R(A) = \frac{V_A}{N} \qquad (4)$$

In (4), Considering $\acute{A}$ as an initial set of active nodes, the V_A and $R(A)$ show the total number of nodes influenced by A during the diffusion process and the influence spread, respectively[33].

The experiments' results of the CNE-IM and comparative methods to different seed sizes on the first data set, dolphin, are shown in Table 2 and "Fig. 7". The probability value of the diffusion model is considered to be 0.05. As it is clear the proposed method has competed with the degree discount method and has in some cases been better. The random method has a minor spread of influence among the methods, and the degree and semi-local centrality results are almost similar. These results are shown schematically in "Fig. 6," where the x-axis and y-axis denote the number of seeds and the average number of influence spread, respectively.

Table 3 shows the evaluation results of the CNE-IM and other comparative methods based on the number of seeds on the Netscience network. In this evaluation, the probability of the diffusion model is considered to be 0.01. As it is clear, the CNE-IM has been superior to other methods, which can be seen schematically in "Fig. 8".

Table 4 also shows the evaluation results of the CNE-IM and other comparative methods based on the number of seeds on the HEP-physics network. In this evaluation, the probability of the diffusion model is considered to be 0.01. In this dataset, Degree centrality, Degree discount, and CNE-IM are in competition with each other, and their results are similar. However, the proposed method has performed better than others, as shown in "Fig. 9".

TABLE II. Influence Spread of Dolphin Social Network

Methods	Number of seed			
	1	2	3	4
CNE-IM	**2.86**	4.44	5.90	**7.50**
Degree centrality	2.52	4.06	5.38	6.48
Ranking random	1.50	3.00	4.30	5.50
Semi-local centrality	2.62	4.08	5.38	6.48
Degree discount	2.62	4.55	6.13	7.05

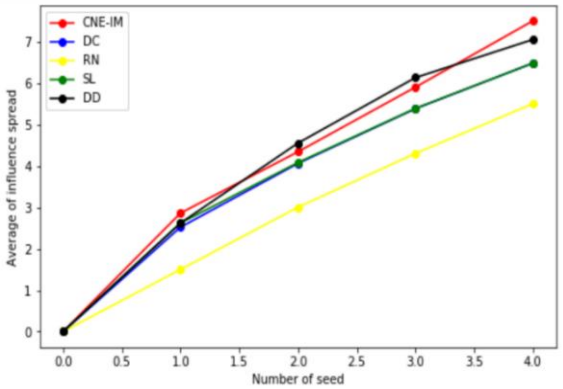

Fig 7. Influence spread of Dolphin social network for different seed size with p=0.05

[DOI: 10.52547/itrc.14.1.38]

TABLE III. INFLUENCE SPREAD OF NETSCIENCE SOCIAL NETWORK

Methods	Number of seed				
	5	10	15	20	25
CNE-IM	**7.03**	**14.00**	**19.00**	**25.70**	**30.50**
Degree centrality	6.59	12.60	17.50	22.40	27.30
Ranking random	3.03	8.50	13.01	17.00	25.00
Semi-local centrality	5.30	10.20	15.10	20.01	26.25
Degree discount	6.871	12.60	18.26	23.98	29.49

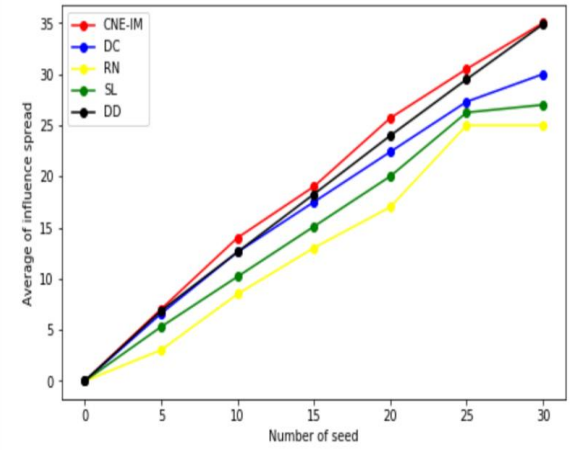

Fig 8. Influence spread of Netscience social network for different seed size with p=0.01

TABLE IV. INFLUENCE SPREAD OF HEP-PHYSICS SOCIAL NETWORK

Methods	Number of seed				
	5	10	15	20	25
CNE-IM	**13.01**	**20.03**	**27.21**	**34.14**	**41.01**
Degree centrality	12.10	19.61	26.46	33.47	40.09
Ranking random	5.58	10.75	19.24	26.25	31.26
Semi-local centrality	11.40	17.48	22.37	27.25	32.13
Degree discount	11.59	19.10	26.16	32.85	40.02

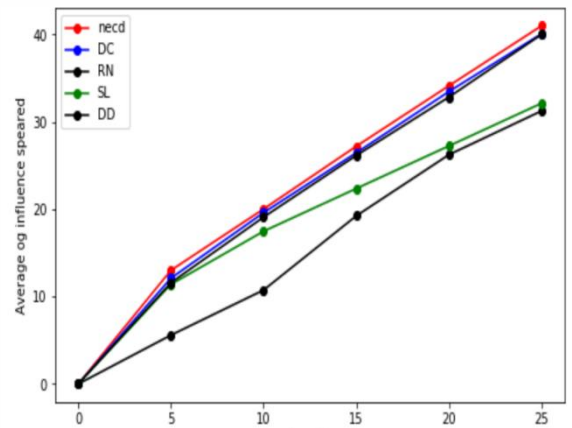

Fig 9. Influence spread of HEP-physics social network for different seed size with p=0.01

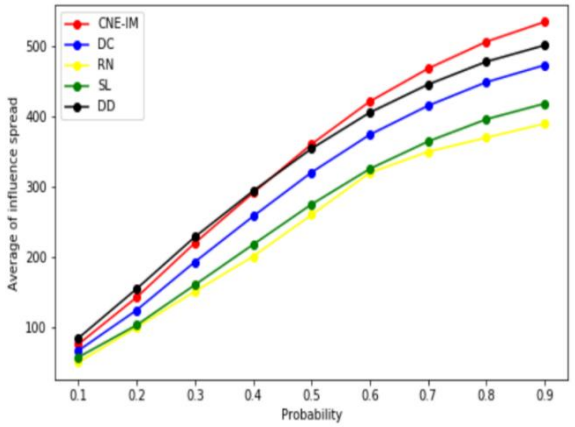

Fig 10. Influence spread of Netscience social network for the different probability of IC diffusion model with seed size=30

"Fig. 10" shows the influence spread of Netscience social network for the different probability of IC diffusion model. In this experiment, first, a seed set of size 30 is extracted by CNE-IM on the Netscience dataset. Then, the influence spread of the seed set is measured on the IC diffusion model. According to "Fig. 10", in the smaller probabilities, most methods have similar influence spreads; however, when the activation probability of edges increases, more users are influenced. Due to higher probabilities, more influence paths are activated, and CNE-IM will be able to transfer influence via these newly discovered paths.

D. Discussion

In this study, community detection obtained from the combination of network embedding and clustering positively affects the influence maximization problem. It has gained positive attention in comparison to other stated approaches. Rather than exploring the total network for identifying influential nodes, only selected candidate communities were surveyed, and communities that were less likely to have influential nodes were excluded. This model proves cost-effective compared to other methods such as degree centrality, ranking random method, semi-local centrality, and degree discount that scan the entire network. The CNE-IM was implemented under the IC diffusion model and was run on real social networks, where the results show good efficiency compared to the other approaches. According to the experiments, the random ranking method had the worst performance in almost all datasets, and the degree centrality results were close to the degree discount results. Because semi-local centrality is sensitive to the network structure and suitable for heterogeneous networks, the degree discount and degree centrality methods are more effective than the semi-local centrality method, using the Netscience and HEP-physics datasets. Finally, in the proposed method, considering the network topology using network embedding and also identifying communities accordingly, better results were obtained. We also proved that the proposed method has monotonicity and submodularity.

[DOI: 10.52547/itrc.14.1.38]

VI. Conclusion and Future work

In this paper, utilizing community recognition methods based on network embedding and Centrality measurements, the CNE-IM method was proposed for identifying influential nodes. First, the specific feature of each node was initially studied via the Node2Vec method, and subsequently, by using the K-Means algorithm, nodes were clustered. Then, the candidate communities were selected according to the size, and the semi-local centrality method was used to identify influential nodes in each selected community. Experimental results verify the performance of the proposed method on two networks in comparison to other algorithms. Furthermore, this method is cost-effective compared to other methods due to the search for nodes being limited to selected communities and not requiring an entire network search.

Regarding the first and second neighbors of each node, the proposed method considers the influential users based on the network topology and does not consider the network content. Content on social networks provides us with significant information, and their use can help identify influential people in specific fields. The application of this method can be seen in the discussion of advertising and how to make it more targeted. As our further research work, is to provide a suitable method in social networks by considering the content of the network. Furthermore, we will also investigate the influence maximization problem in multiple networks. Since users tend to operate on multiple social networks simultaneously, we can expand our spread across multiple networks.

References

[1] M. Xu, "Understanding graph embedding methods and their applications," *SIAM Review*, vol. 63, no. 4, pp. 825-853, 2021.

[2] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *Knowledge and Information Systems*, pp. 1-39, 2020.

[3] L. Qiu, X. Tian, S. Sai, and C. Gu, "LGIM: A global selection algorithm based on local influence for influence maximization in social networks," *IEEE Access*, vol. 8, pp. 4318-4328, 2019.

[4] S. S. Singh, D. Srivastva, M. Verma, and J. Singh, "Influence maximization frameworks, performance, challenges and directions on social network: A theoretical study," *Journal of King Saud University-Computer and Information Sciences*, 2021.

[5] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777-1787, 2012.

[6] W. Xu and W. Wu, *Optimal Social Influence*. Springer, 2020.

[7] S. Tian, S. Mo, L. Wang, and Z. Peng, "Deep reinforcement learning-based approach to tackle topic-aware influence maximization," *Data Science and Engineering*, pp. 1-11, 2020.

[8] H. Huang, Z. Meng, and S. Liang, "Recurrent Neural Variational Model for Follower-based Influence Maximization," *Information Sciences*, 2020.

[9] D. A. Vega-Oliveros, L. da Fontoura Costa, and F. A. Rodrigues, "Influence maximization by rumor spreading on correlated networks through community identification," *Communications in Nonlinear Science and Numerical Simulation*, vol. 83, p. 105094, 2020.

[10] W. Ju, L. Chen, B. Li, W. Liu, J. Sheng, and Y. Wang, "A new algorithm for positive influence maximization in signed networks," *Information Sciences*, vol. 512, pp. 1571-1591, 2020.

[11] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137-146.

[12] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5766-5771, 2002.

[13] S. S. Singh, A. Kumar, K. Singh, and B. Biswas, "C2IM: Community based context-aware influence maximization in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 514, pp. 796-818, 2019.

[14] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 420-429.

[15] A. Goyal, W. Lu, and L. V. Lakshmanan, "Celf++ optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 47-48.

[16] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, "Staticgreedy: solving the scalability-accuracy dilemma in influence maximization," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 509-518.

[17] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, 2014, pp. 946-957: SIAM.

[18] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 629-638.

[19] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 75-86.

[20] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1539-1554.

[21] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proceedings of the 2016 international conference on management of data*, 2016, pp. 695-710.

[22] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 199-208.

[23] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *2010 IEEE international conference on data mining*, 2010, pp. 88-97: IEEE.

[24] A. Goyal, W. Lu, and L. V. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *2011 IEEE 11th international conference on data mining*, 2011, pp. 211-220: IEEE.

[25] R. Narayanam and Y. Narahari, "A shapley value-based approach to discover influential nodes in social networks," *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 1, pp. 130-147, 2010.

[26] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks," in *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 918-923: IEEE.

[27] S. Galhotra, A. Arora, S. Virinchi, and S. Roy, "Asim: A scalable algorithm for influence maximization under the independent cascade model," in *Proceedings of the 24th*

[DOI: 10.52547/itrc.14.1.38]

International Conference on World Wide Web, 2015, pp. 35-36.

[28] S. Galhotra, A. Arora, and S. Roy, "Holistic influence maximization: Combining scalability and efficiency with opinion-aware models," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 743-758.

[29] D. Bucur and G. Iacca, "Influence maximization in social networks with genetic algorithms," in *European conference on the applications of evolutionary computation*, 2016, pp. 379-392: Springer.

[30] Q. Jiang, G. Song, C. Gao, Y. Wang, W. Si, and K. Xie, "Simulated annealing based influence maximization in social networks," in *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

[31] J. J. Lotf, M. A. Azgomi, and M. R. E. Dishabi, "An improved influence maximization method for social networks based on genetic algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 586, p. 126480, 2022.

[32] W. Li, Y. Li, W. Liu, and C. Wang, "An influence maximization method based on crowd emotion under an emotion-based attribute social network," *Information Processing & Management*, vol. 59, no. 2, p. 102818, 2022.

[33] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 1039-1048.

[34] Y.-C. Chen, W.-Y. Zhu, W.-C. Peng, W.-C. Lee, and S.-Y. Lee, "CIM: Community-based influence maximization in social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, pp. 1-31, 2014.

[35] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 1029-1038.

[36] X. Li, X. Cheng, S. Su, and C. Sun, "Community-based seeds selection algorithm for location aware influence maximization," *Neurocomputing*, vol. 275, pp. 1601-1613, 2018.

[37] M. Pourkazemi and M. Keyvanpour, "CNLPSO-SL: A two-layered method for identifying influential nodes in social networks," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 22, no. 2, pp. 109-123, 2018.

[38] M. Pourkazemi and M. R. Keyvanpour, "Community detection in social network by using a multi-objective evolutionary algorithm," *Intelligent Data Analysis*, vol. 21, no. 2, pp. 385-409, 2017.

[39] F. Ye, J. Liu, C. Chen, G. Ling, Z. Zheng, and Y. Zhou, "Identifying influential individuals on large-scale social networks: A community based approach," *IEEE Access*, vol. 6, pp. 47240-47257, 2018.

[40] N. Vafaei, M. R. Keyvanpour, and S. V. Shojaedini, "Influence Maximization in Social Media: Network Embedding for Extracting Structural Feature Vector," in *2021 7th International Conference on Web Research (ICWR)*, 2021, pp. 35-40: IEEE.

[41] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852-1872, 2018.

[42] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 5, pp. 833-852, 2018.

[43] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855-864.

[44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[45] B. Wilder12, N. Immorlica, E. Rice24, and M. Tambe12, "Maximizing influence in an unknown social network," 2018.

[46] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396-405, 2003.

[47] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[48] B. Hou, Y. Yao, and D. Liao, "Identifying all-around nodes for spreading dynamics in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 15, pp. 4012-4017, 2012.

[49] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on evolutionary computation*, vol. 18, no. 1, pp. 82-97, 2013.

[50] J. Liu, Q. Xiong, W. Shi, X. Shi, and K. Wang, "Evaluating the importance of nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 209-219, 2016.

[51] Z. Wang, Y. Zhao, J. Xi, and C. Du, "Fast ranking influential nodes in complex networks using a k-shell iteration factor," *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 171-181, 2016.

[52] X. Liu, N. Ding, C. Liu, Y. Zhang, and T. Tang, "Novel social network community discovery method combined local distance with node rank optimization function," *Applied Intelligence*, pp. 1-18, 2021.

[53] A. Kumari, R. K. Behera, A. S. Shukla, S. P. Sahoo, S. Misra, and S. K. Rath, "Quantifying Influential Communities in Granular Social Networks Using Fuzzy Theory," in *International Conference on Computational Science and Its Applications*, 2020, pp. 906-917: Springer.

[54] L. Li, Y. Liu, Q. Zhou, W. Yang, and J. Yuan, "Targeted influence maximization under a multifactor-based information propagation model," *Information Sciences*, vol. 519, pp. 124-140, 2020.

[55] F. Wang *et al.*, "Maximizing positive influence in competitive social networks: A trust-based solution," *Information sciences*, vol. 546, pp. 559-572, 2021.

[56] F. Ghayour-Baghbani, M. Asadpour, and H. Faili, "MLPR: Efficient influence maximization in linear threshold propagation model using linear programming," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1-10, 2021.

Narges Vafaei received her B.Sc. degree in Software Engineering from the Shahrood University of Technology. She is currently pursuing the M.Sc. degree in Artificial Intelligence at the Alzahra University, Tehran, Iran. Her research interests include Social Network Analysis, Natural Language Processing, and Data Mining.

Mohammad Reza Keyvanpour is an Associate Professor at Alzahra University, Tehran, Iran. He received his B.Sc. degree in Software Engineering from Iran University of Science and Technology, Tehran, Iran. He received his M.Sc. and Ph.D. degrees in Software Engineering from Tarbiat Modares University, Tehran, Iran. His research interests include Image Retrieval and Data Mining.

[DOI: 10.52547/itrc.14.1.38]