Journal of Industrial and Systems Engineering
Vol. 13, No. 3, pp. 16-40
Summer (July) 2021

JISE

Influence maximization in complex social networks based on community structure

Babak Amiri[1*], Mohammad Fathian[1], Elnaz Asaadi[1]

[1]*School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*

amiri.babak@gmail.com, fathian@iust.ac.ir, asadielnaz@gmail.com

Abstract

Many real-world networks, including biological networks, internet, information and social networks can be modeled by a complex network consisting of a large number of elements connected to each other. One of the important issues in complex networks is the evaluation of node importance because of its wide usage and great theoretical significance, such as in information diffusion, control of disease spreading, viral marketing and rumor dynamics. A fundamental issue is to identify a set of most influential individuals who would maximize the influence spread of the network. In this paper, we propose a novel algorithm for identifying influential nodes in complex networks with community structure without having to determine the number of seed nodes based on genetic algorithm. The proposed algorithm can identify influential nodes with three methods at each stage (degree centrality, random and structural hole) in each community and measure the spread of influence again at each stage. This process continues until the end of the genetic algorithm, and at the last stage, the most influential nodes are identified with maximum diffusion in each community. Our community-based influencers detection approach enables us to find more influential nodes than those suggested by page-rank and other centrality measures. Furthermore, the proposed algorithm does not require determining the number of k initial active nodes.

Keywords: Influential nodes, complex networks, community detection, influence maximization

1- Introduction

We live in a world surrounded by a variety of complex networks such as the internet, social networks, transportation networks, power grids and neural networks. Today social networks are playing an important role in the social life of people due to the fast growth of internet technology. With the growing popularity of online social networks, the way consumers and companies interact with others has changed. Consumers are looking for useful information from active people on online social networks; these people are commonly referred to as influencers (Liu et al., 2015). Finding a small subset of influential individuals in a complex network that would spread information to all the nodes in the network is a fundamental issue.

*Corresponding author

A company initially targets several influential people on the network by giving them an example of free products while hoping that the initially selected users will recommend a product to their friends who will in turn influence their own friends and so on; many individuals ultimately accept the new product owing to the powerful effect of word-of-mouth (also known as viral marketing) (Zhang et al., 2013).

A fundamental issue is how to choose the best initial influencers so that the product can be widely adopted by many people (Shang et al., 2017). The problem of influence maximization is how to identify the best nodes for initial activation in order to maximize the number of activated nodes at the end of the activation process under a particular diffusion model. In fact, the main purpose of influence maximization is to find a subset of key users to maximize the influence in the network. The influence spread of a node set S, as the expected number of active nodes, is defined based on the given seed set S, which is denoted by σ (S). The influence maximization is expressed as a problem on how to select k nodes from a network G so that the influence spread σ (S) is maximized (Gong et al., 2016). As proved in Zhang et al., (2013) the computation of influence spread σ (S) for a given set S is #P-hard and the influence maximization has been proved as a NP-hard optimization problem (Gong et al., 2016).

In social networks, the influence spread of an individual's opinion is limited within the circle of friends' friends' friends (i.e. three-hop area), because there exists an intrinsic-decay, which is called Three Degree Theory (Walker, 2011). Pei et al., (2016) pointed out that the global influence of a node in the network depends on the influence in the two-hop area. Following those studies, (Gong *et al.*, 2016) developed a local influence estimation to approximate the influence spread within the two-hop area (the neighbors' neighbors) for a node set S. Moreover, they introduced fast local influence estimation (LIE) function to compute the influence spread in both independent cascade and weighted cascade models. (Kempe and Kleinberg, 2003) proposed a greedy algorithm to solve the influence maximization problem, using the Monte Carlo simulation. It has a high execution time and is not scalable for large networks. To reduce the complexity of calculations, Leskovec et al., (2007) proposed a CELF (cost-effective lazy-forward) method.

One of the methods that can be replaced with the greedy method is identifying communities in the network to make sure that there is at least one initial activated node in any community or cluster. The advantage of this method compared to the greedy algorithm is its scalability and improved speed in the selection of the initial nodes.

In this paper, we propose an algorithm based on the genetic algorithm to identify influential nodes in social networks with community structure without having to determine the number of seed nodes. Our major contributions are summarized as follows:

- We employ genetics algorithm to detect communities throughout the network.
- In every step of the algorithm in each community, influential nodes are selected with three different methods: (a) randomly; (b) node with a maximum degree in a community; (c) structural holes.
- We compute fitness function with an LIE function as an influence measure that is proposed by Gong et al., (2016).
- In the last step of the algorithm, communities and nodes with maximum diffusion in each community are identified. The nodes that are identified at the last step of the algorithm are known as influential nodes.

The rest of the paper is organized as follows: In section 2, a related work of influence maximization is overviewed; in section 3, we introduce our proposed method for identifying influential nodes based on a genetic algorithm; the experimental results are presented in section 4' and finally, section 5 is sums up the conclusion of the paper.

2- Literature review

The influence maximization problem has been presented by Kempe and Kleinberg, (2003). They proposed the greedy algorithm which is time-consuming and costly. To solve this problem, a large

number of research papers have been published. Generally, these works can be divided into three categories: (1) centrality-based algorithm; (2) community-based algorithms; (3) greedy algorithms. Here is a brief review of these research works.

2-1-Centrality-based algorithm

To get k "best" influential nodes in order to maximize diffusion, we can simply select the best k node based on some predefined centrality measures (e.g. degree centrality, closeness centrality, etc.). However, previous studies have shown that k node selection based on these centrality measures can produce false results. In order to solve this problem, some researchers provide new centralities that are more accurate (Shang et al., 2017).

Kempe and Kleinberg (2003) observed that diffusion function is monotone and submodular, and presented a simple greedy algorithm that uses large numbers of Monte-Carlo simulations to select the seed node with the maximal marginal gain. This simple greedy algorithm provides a factor of $(1-1/e-\varepsilon)$ optimal for the influence maximization problem. Zhao et al., (2015) reviewed the properties of the nodes and proposed a graph signal processing based centrality (GSPC) method, taking into account both node characteristics and network topology. Liu, Jing and Chang, (2016) proposed a weighted semi-local centrality measure, which conquers the defect of semi-local centrality. That not only combines the degree and the weight of a node but also considers the information pertaining to the multiple layers of node neighbors.

Fei et al., (2017) proposed a new approach based on existing centrality measures, which can identify vital nodes more effectively by combining the existing centrality measures with the characteristics of the network. Different centrality measures are considered as parameters for sorting nodes based on the mode of dissemination of information in the network. Zhao et al., (2017) proposed a novel measurement based on local centrality with a coefficient, where the local clustering coefficients of a node as well as its semi-local centrality are used to measure its impact. Bao et al., (2017) suggested a semi-local centrality indicator, which combines the shortest distance with the number of shortest paths and the reciprocal of average degree at the same time. Bian and Deng, (2017) presented a method based on the Analytic Hierarchy Process (AHP). Depending on the different measures of the network and the features of a complex network, AHP is used for evaluating the importance of nodes and identifying the influential nodes. Cai et al., (2017) identified the effective nodes by considering three measures: degree centrality, closeness centrality and betweenness centrality - all based on the evidence theory. Wenfeng Kang, Guangming Tang and Yifeng Sun, (2017) provided an algorithm using the weighted formal concept analysis, which is a smart computing technique. Ullah and Lee, (2017) proposed a temporal multi-hops social interaction-based centrality, considering the nodes' neighbors and neighbors-of-neighbors temporal modeled interactions and topological connections to select nodes that have the ability to spread further. Ahajjam and Badir, (2018) proposed a Hybrid Rank algorithm using a new hybrid centrality measure for detecting a set of influential spreaders using the topological features of the network.

2-2-Community-based algorithms

In social networks, some nodes have more communications with each other than the entire network nodes, which is referred to as the community. In fact, recognizing communities shows the division of the network and separates communities from a graph. Detecting communities helps us to better understand the network structure. Furthermore, by finding communities, the number of final activated nodes is maximized.

Wang et al., (2010) provided a community-based greedy algorithm for identifying top-k influential nodes. The proposed algorithm consists of two components: 1) an algorithm for identifying communities in a social network with respect to the dissemination of information, and; 2) a dynamic planning algorithm for selecting communities to find the influential nodes. Zhang et al., (2013) proposed a method with community structure that extracts the k most influential nodes in complex networks by applying the k-medoid clustering algorithm on the transfer probability matrix. Scripps, (2013) presented an algorithm based on community finding, which not only finds maximum activated nodes in the network but also

covers as many communities as possible and maximizes the spread of influence. Wei Li and Jianbin Huang (2015) proposed a community-based greedy algorithm to identify top-k nodes in social networks, which consists of two separate steps: community detection and extraction of top-k nodes. In the first step, an efficient algorithm is used to discover the structure of a community in a network. Then a "split and conquer" process is applied to find top-k nodes in the network. Halappanavar, Sathanur and Nandi (2016) introduced a new idea by using community detection as a preprocessing step to accelerate the extraction of influential nodes in a complex network. Liu, Jing and Chang (2016) provided a new centrality measure based on the expansion factor that can identify nodes in different communities. The expansion factor centrality does not need to consider the structure of the community but depends on the results of the division of the community. Zhao, Li and Jin (2016) proposed the algorithm with community structure based on label propagation. The algorithm is parameter-free and requires no prior information about the community structure. Hosseini-Pozveh, Zamanifar and Naghsh-Nilchi (2017) considering the community structure of the social networks, proposed two new methods to maximize the spread of influence on the multiplication threshold, minimum threshold and linear threshold information diffusion models. Jaouadi and Ben Romdhane (2017) proposed a parameter less algorithm called DIN (Detecting Influential Nodes) in social networks that combines the structure and the semantic aspect. The main idea of the algorithm is to detect communities with overlap, to model the semantic of each community and select influential elements. Shang et al. (2017) presented a community-based framework on large-scale networks. In the proposed framework, the diffusion process is divided into two stages: (1) seeds expansion; and (2) propagation within the community. The first stage is a distribution of seed nodes among various communities at the beginning of the diffusion. The second stage is the influence propagation within the communities that are independent of each other. And to select seed nodes they developed a fast-greedy algorithm. Singh et al., (2019) proposed a Community based Context-aware Influence Maximization (C2IM) algorithm, which uses a community-based framework to improve the time-efficiency that reduces the search space significantly. Huang et al. (2019) proposed a Community-based Topic-aware Influence Maximization (CTIM), which exploits topic-aware and community-based strategy to improve the performance of influence maximization. B et al., (2019) proposed a Community-based Influence Maximization (CoIM) algorithm. CoIM focuses on the time-efficiency without much compromising on quality of seed.

2-3-Greedy algorithm

Leskovec et al. (2007) solved the time efficiency problem of the greedy algorithm by exploiting the submodular property of the spread function and presenting a CELF (cost-effective lazy-forward) method, which improves the running time according to the sub modularity property of the influence spread. Goyal, Lu and Lakshmanan, (2011) proposed the CELF++ algorithm, an extension of CELF, which is about 30% ~50% faster than the CELF algorithm. Zhou, Zhang and Cheng, (2014) proposed a two-stage algorithm called a greedy algorithm based on user preferences. To calculate user preferences in the first stage, two different models are studied: a LSI model to adjust the user preferences to a decreased hidden space, and a VSM based model, which is used in data retrieval. In the next stage, the greedy algorithm is used to find influential nodes in the network. Zhou et al. (2014) proposed an efficient Upper Bound based Lazy Forward algorithm by incorporating the bound into CELF. Kaur, Talluri and He (2015) proposed a new diffusion model based on the positive and negative effects of the host's view on social networks. Then, a novel problem called blocking selection of negative influential node set problem is presented to identify the positive node set so that the number of active negative nodes for all competitors is minimized in terms of the host. And a greedy algorithm for solving the problem is proposed. He, Kaur and Talluri (2016) presented a new problem called Positive Opinion Influential Node Set selection problem, which is useful for eliminating the negative effects while promoting products on social networks. And to solve the positive opinion influential node set selection problem, a greedy algorithm is proposed. Song et al., (2017) proposed the Upper Bound Interchange (UBI) and UBI + algorithms that improve the boundary of node replacement computation. Instead of creating a collection of seed nodes from the beginning, it starts from the previously found collection of influential seed nodes. Meantime, node replacement is used to

improve the influence coverage. Azaouzi and Romdhane (2018) proposed a Social Action-Based Influence Maximization Model (SAIM) for influence maximization in social networks, which compute an optimal set of influential nodes using a new concept named "influence-BFS tree".

3-Problem formulation

In this section, first, we introduce the basic concepts of influence calculation that are needed for a better understanding of the paper. Then the proposed algorithm is introduced.

Structural Holes: The concept of structural holes was first introduced by Burt (1992). Structural holes are nodes that act as a bridge in social networks. If they are deleted the network will be separated and the diffusion of information will be locked (Zhu, Liu and Yin, 2017). Burt (1992) introduced the network constraint coefficient to measure structural holes in the network. Obviously, the higher the coefficients are, the harder the nodes will form structural holes. Conversely, the smaller the coefficients are, the more important position the node has and the more influential the nodes will be. The network constraint coefficient of node *i* is:

$$C_i = \sum_{j \in \Gamma_i} \left(q_{ij} + \sum_{k\,(\neq i,j)} q_{ik} q_{kj} \right)^2 \tag{1}$$

$$q_{ij} = x_{ij} \,/\, \sum_{m \in \Gamma_i} x_{im} \tag{2}$$

Where node j is the neighbor of node i, and Γ_i is a set of neighbors of node i. Node k is the common neighbor of node i and node j. q_{ij} represents the ratio of node i's effort, which leads to maintaining the neighborhood relation to node j. $x_{ij} = 1$ if node i connects to node j. Otherwise, $x_{ij} = 0$. And r_{ij} can be simplified to simplify = $1/k(i)$. $k(i)$ is the degree of node i (Hu and Mei, 2018).

Genetic algorithm: The genetic algorithm is inspired by Darwin's theory of evolution (Yeh and Lin, 2007) and is used extensively in solving optimization problems and learning processes. Generally, in nature better generations arise from the combination of right chromosomes. Sometimes mutations occur in the chromosomes, which may lead to a better subsequent generation. That is based on such perception that the genetic algorithm solves the problem. The genetic algorithm uses various operators to solve problems described below. In the genetic algorithm, randomly, several solutions are generated for the problem. This set of solutions is called the initial population. Each solution is called a chromosome. Then, using the genetic algorithm operators, after selecting better chromosomes, the chromosomes are combined with the operators of the genetic algorithm and a mutation occurs in them. The chromosomes are compared to their fitness values, which are calculated through a fitness function. Eventually, the current population is combined with a new population that results from the combination and mutation in the chromosomes. This process continues until certain convergence criteria are met with fitness values, or until a certain number of repetitions.

Diffusion model: Influence maximization is based on the diffusion model. Currently, the most widely used models are Independent Cascade model (IC model) and Linear Threshold model (LT model). In both models, each node is in one of the following two states: active or inactive. Active nodes are those who have accepted the product and who will introduce it to their neighbors. Inactive nodes are those who have not heard anything about the product or refused to accept it. A node can switch from inactive to active under the influence of neighboring nodes in the network, but vice versa is not possible. First, all nodes are

inactive and then k nodes are selected to be activated and diffusion starts from an initial set of active nodes (Zhao, Li and Jin, 2016).

IC model: In Independent cascade model, some nodes are active by default and the rest are inactive. Whenever node v is activated with a degree of probability it can activate its neighbors. This probability is equal to the weight of the edge between node v and node u, $p_{u,v}$. Node v with the probability $p_{u,v}$ has only one chance to activate node u. This means if node v could not activate node u, it will not have a chance to activate node u. The diffusion process will continue until no nodes would remain active (Goldenberg et al., no date).

LT model: In Linear threshold model, like the IC model**,** some nodes are active by default and the rest are inactive. A threshold θ_v is considered for each node, and the impact of vertices on each other is assumed with a weight of $b_{u,v}$**,** which indicates the effect of node u on node v. This weight or probability is such that $\sum_{u\in N(v)} b_{u,v} \leq 1$, where N (v) denotes all v's neighbors. Then, each inactive node v is activated only if some of the influence of all active neighbors of node v exceeds its corresponding threshold, i.e., $\sum_{u\in \Gamma(v)} b_{u,v} \geq \theta_v$, where Γ(v) is the set of active neighbor nodes of node v and θ_v is the threshold for node v. The process runs until no more activations are possible (Watts, 2002).

3-1-Proposed algorithm

In this subsection, we design a new algorithm based on a genetic algorithm to identify the influential nodes for the problem of influence maximization from a community-based perspective. First, we provide an overview of the algorithm, and then we explain each stage in more details.

Algorithm overview: The proposed algorithm is presented in complex networks with community structure for identifying influential nodes based on the genetic algorithm with the aim of influence maximization. At each stage of the implementation of the algorithm, the communities are identified and the number of influential nodes is determined, which is equivalent to the number of communities obtained at the same stage. In each community, a node is selected in three ways as an influential node: (degree centrality, random and structural hole). And the amount of the spread diffusion of each influential node in its community is calculated by the LIE function. And this process continues until the end of the genetic algorithm, and at the last stage, the most influential nodes that have the highest diffusion in its community are identified.

1)Habitat representation

Each generated habitat represents a solution for the proposed method. In the proposed method the locus-based adjacency representation proposed by Park and YoungJa, (1998) and used by Handl, (2007), is used to encode each habitat. In this graphical representation, each habitat consists of N variables, where N denotes the number of nodes in the corresponding network. For each of these N variables, there is a set of possible values that can be deduced from the adjacency matrix of the network. For example, if node 1 is associated with nodes 2, 3 and 4 in the corresponding network, then a set of possible values for node 1 is a set {2, 3, 4}. Therefore, in the locus-based adjacency representation, the variables and their values represent the node of network nodes. On the other hand, if the value of j is assigned to the i^{th} variable, it is interpreted from the corresponding network as a link (edge) between nodes i and j. It means that in the partition of the network which is represented by the corresponding habitat, nodes i and j are in the same cluster (or community). After encoding habitat, the decryption stage is required to identify all the clusters (communities) that are generated by each habitat. At this stage, the nodes that are part of the same component are assigned to one cluster. The main advantage of the locus-based adjacency representation is that the number of clusters (or communities) will be automatically determined by the number of

components of each habitat in the decoding step (Reihanian, Feizi-Derakhshi and Aghdasi, 2017). An example of the encoding and decoding process in the locus-based adjacency representation for a habitat is shown in figure 1.

2)Primary population generation

To generate primary population in the proposed method, the random production method of chromosomes is used. For each of the N nodes in the network, a neighbor from among its neighbors is randomly selected and stored at the position of that node in the corresponding chromosome. This action is repeated until the N^{th} node is reached and eventually the corresponding chromosome is formed. The above operation is repeated for all members of the initial population, and ultimately the initial population matrix is formed.

3)Parent selection

The roulette wheel mechanism is used to select the parent in the proposed method. The fitness function assigns fitness to possible chromosomes.

If the chromosome has more fitness, it is more likely to be selected for the next generation production. In this method, we first calculate the probability of selecting (P) for each chromosome. This probability is calculated by the following equation:

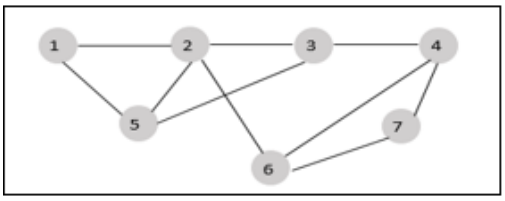
$$P_i = \frac{F_i}{\sum_{j=1}^{N} F_i} \tag{3}$$

Where F_i is the fitness of chromosome i in the population and N is the number of chromosomes in the population. Then, for each chromosome, we calculate the cumulative probability (Pm).

In the selection stage, it is required to select Nc number parent. The value of Nc is obtained from the following equation:

$$Nc = 2 \times round\left(pc \times \frac{N_{pop}}{2} \right) \tag{4}$$

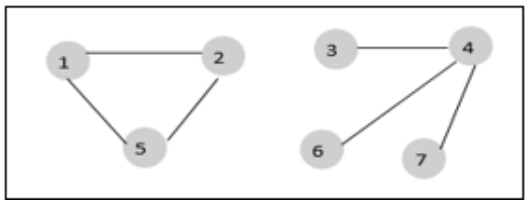
Where Pc is the probability of the combination and N_{pop} is the number of members of the population. To select parents, a random number is initially generated and, if its value is smaller than the cumulative probability of each of the chromosomes, that chromosome is selected for the combination.



(a) An example of a network with 7 nodes and 10 edges.

SIV number: 1 2 3 4 5 6 7
SIV value(habitat): 5 1 4 7 2 4 4

(b) An example of encoding a habitat with Locus-based adjacency representation.



(c) The final communities of the network shown in part (a) after decoding the habitat illustrated in part (b).

Fig 1. An example of encoding and decoding process in the locus- based adjacency representation (Reihanian et al., 2017)

4)Uniform Crossover

We use uniform crossover because it ensures the maintenance of the effective communication of the nodes in the social network in each individual child. In fact, due to the biased initialization, each individual with a feature has a safe place in the population. If the gene i contains a value of j, then there is an edge (i , j). Therefore, with respect to two safe parents, a random binary vector is created. Then the uniform crossover from the first parent chooses the genes where the vector is a 1, and from the second parent selects the genes where the vector is a 0, and combines the genes to form the child. The child in each position i contains the value of j, inherited from one of its two parents. So, there is an edge (*i* , *j*), and this means that two safe parents get a safe child (Pizzuti, 2008).

5)Mutation

To apply the mutation operator, for each chromosome selected for mutation, a gene of chromosomes (i) is selected randomly. And its value is changed by a random selection of a neighbor from the neighboring neighbors of node *i*.

6)*Objective function*

Given a graph G = (N, E, W) and an integer K <|N|, the selection of k nodes as an initial node set $S=\{s_1,s_2,...,s_k \mid s_i \in N, i=1,2,...,k\}$, so that influence spread σ (S) is maximized (Gong et al., 2016), which can be formulated as :

$$\begin{cases} S = \arg\max\{\sigma(s)\} \\ \quad st: \| S \| = k \end{cases} \tag{5}$$

According to the diffusion models explained in section 3, we use the LIE function to compute the influence spread in the Weighted Cascade model (WC). The difference between the IC and WC models is related to the way the probability of each edge is determined.

In the IC model, the active probability P_{ij} with which node i activates its neighbor j can be computed as follows (Gong et al., 2016):

$$p_{ij} = 1-(1-p)^{w_{ij}} \tag{6}$$

Where P ∈ [0, 1] is the active probability and w_{ij} represents the weight of the edge (i , j). In particular, if the corresponding social network is unweighted the values of active probability for all edges are equal to p (Gong et al., 2016).

In the WC model, the active probability between nodes i and j can be determined by the in-degree number of node j, and it can be computed by equation (7).

$$p_{ij} = \frac{a_{ij}}{d_j^{(in)}} \tag{7}$$

We have $a_{ij}=1$ if the node j is one of the neighbors of node i; and 0 otherwise. $d_j^{(in)}$ is the number of edges referred to as node j in a directed graph G. The $d_j^{(in)}$ in equation (7) is the degree of node j, if graph G is undirected.

A novel metric named expected diffusion value (EDV for short) introduced in the IC model in order to improve the efficiency of a basic greedy algorithm with a view to replacing the expected diffusion simulation (Gong *et al.*, 2016). The EDV metric computes how many neighbors of the selected node set S are expected to be influenced approximates σ (S) as follows:

$$EDV\ (\mathrm{S}) = k + \sum_{i \in N_s^{(1)} \setminus S} (1-(1-\mathrm{p})^{\tau(\mathrm{i})}) \tag{8}$$

where $\mathrm{N}_S^{(1)}$ represents the direct neighbors (i.e. one-hop area) of S, and i ∈ $\mathrm{N}_S^{(1)}$ \S represents that i belongs to $\mathrm{N}_S^{(1)}$ but not to S. p is the preset value for activation probability in the IC model and τ(i) denotes the number of links between node i and the initial node set S (Gong et al., 2016).
The expected influence spread of one-hop area for a node set S can be computed as:

$$\sigma_1(\mathrm{S}) = \sigma_0(\mathrm{S}) + \sigma_1^{*}(\mathrm{S}) = \mathrm{k} + \sum_{i \in N_s^{(1)} \setminus S} (1 - \prod_{(\mathrm{i},\mathrm{j}) \in \mathrm{E}, \mathrm{j} \in \mathrm{S}} (1 - \mathrm{p}_{ij})) \tag{9}$$

Where p_{ij} represents the activation probability of node i activating node j. Based on $\sigma_1(\mathrm{S})$, the LIE function can be formulated as follows:

$$\begin{aligned} LIE(\mathrm{S}) &= \sigma_0(\mathrm{S}) + \sigma_1^{*}(\mathrm{S}) + \tilde{\sigma}_2(\mathrm{S}) \\ &= k + \sigma_1^{*}(\mathrm{S}) + \frac{\sigma_1^{*}(\mathrm{S})}{|\,\mathrm{N}_S^{(1)} \setminus \mathrm{S}\,|} \sum_{u \in N_s^{(2)} \setminus s} p_u^{*} d_u^{*} \\ &= k + \left(1 + \frac{\sigma_1^{*}(\mathrm{S})}{|\,\mathrm{N}_S^{(1)} \setminus \mathrm{S}\,|} \sum_{u \in N_s^{(2)} \setminus s} p_u^{*} d_u^{*}\right) \sigma_1^{*}(\mathrm{S}) \\ &= k + \left(1 + \frac{\sigma_1^{*}(\mathrm{S})}{|\,\mathrm{N}_S^{(1)} \setminus \mathrm{S}\,|} \sum_{u \in N_s^{(2)} \setminus s} p_u^{*} d_u^{*}\right) \sum_{i \in N_s^{(1)} \setminus S} \left(1 - \prod_{(\mathrm{i},\mathrm{j}) \in \mathrm{E}, \mathrm{j} \in \mathrm{S}} (1 - \mathrm{p}_{ij})\right) \end{aligned} \tag{10}$$

Where $\mathrm{N}_S^{(1)}$ and $\mathrm{N}_S^{(2)}$ represents the S's one-hop and two-hop area, respectively. The parameter p_u^{*} is the constant active probability of node i, and it corresponds to $\frac{1}{d_i^{(\mathrm{in})}}$ in the WC model. d_u^{*} is the number of edges of node u within $\mathrm{N}_S^{(1)}$ and $\mathrm{N}_S^{(2)}$, which represents the number of activated probability for node u. Generally, $d_u^{*} \le d_u$ (Gong et al., 2016).

As shown in equation (10), the computational complexity of LIE function is $O(\mathrm{k}\bar{\mathrm{D}}^2)$, where $\bar{\mathrm{D}}$ is the average degree for the given social network. The LIE function can be utilized to estimate the two-hop influence spread in both the IC and WC models (Gong et al., 2016).
In equation (5), the number of seed nodes needs to be determined. Therefore, the model used in this article does not need to determine the number of seed nodes (k nodes) based on the LIE function and is

$$\begin{cases} Max\ LIE(S) \\ st: \|S\| \ge 1 \end{cases} \tag{11}$$

defined as:
However, the model has fewer constraints, the answer space is definitely not smaller and is at least as difficult as the previous model. Therefore, it is impossible to find the maximum effect with precision.

4-Evaluation

In order to clearly compare the performance of our own algorithm with likes, we take into consideration the famous centrality measures. An important issue in this section is the novelty of using the LIE function as an influence measure in a community structure. In the previous contributions to the literature of influencers detection based on community detection, the LIE function was not used as a fitness function, and the diffusion is usually discussed as an iterative process by either the IC or LT model.

Now, we report the results of running our algorithm over several famous datasets in the literature. First, the collection of datasets is introduced, then the results of running our algorithm are reported.

4-1- Dataset

We evaluate the performance of our community-based influential node identification algorithm in four real-world datasets, which are downloaded from the website[1].

Zachary karate club: This is the well-known and much-used Zachary karate club network. The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is undirected. An often-discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers (Zachary, 1977).

Dolphins: This is a directed social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand (spelled fiord in New Zealand). An edge indicates a frequent association. The dolphins were observed between 1994 and 2001 (Lusseau et al., 2003).

American College football: the network of American football games between Division IA colleges during regular season Fall 2000 (Newman, 2002).

Jazz musicians: This is the collaboration network between Jazz musicians. Each node is a Jazz musician and an edge denotes that two musicians have played together in a band. The data was collected in 2003 (Danon, 2003).

Email: This dataset is the e-mails exchanged between the members of the University Rovirai Virgili (URV) (Guimera et al., 2003).

Table 1 shows the structural characteristics of each of these datasets.

Table 1. The Summary of five real-world datasets

Dataset	Zachary karate club	Dolphins	American College football	Jazz musicians	Email
# Nodes	34	62	115	198	1133
# Edges	156	159	613	2742	5451
Type of network	Directed-Weighted	Undirected-Unweighted	Undirected-Unweighted	Directed-Weighted	Directed-Unweighted
Density	0.139	0.084	0.094	0.07	0.086
Clustering coefficient	0.571	0.303	0.403	0.309	0.487

4-2- Experimental results

In this subsection, we investigate the results of running our algorithm over several samples. At the first step, using a non-parametric ANOVA, the differences among several seed selection methods are assessed. Then a pairwise comparison between the methods is done and the best method of seed selection is determined. Finally, the overall performance of our selected method is discussed in terms of famous centrality measures.

[1] http://konect.uni-koblenz.de/

Analysis of the effect of seed node selection methods: We test the difference of fast local influence estimation (LIE) in terms of different methods of selecting seed nodes. For this purpose, a non-parametric Kruskal-Wallis rank sum test is used. The results of the Kruskal-Wallis test for the 20-time execution of the algorithm are summarized in table 2. The table elements are x^2 and corresponding p-value respectively, showing the result of a hypothesis test for each dataset. According to the values of table 2, the equality of average local influence estimation (LIE) for the four methods of selecting seed nodes (random, maximum degree, maximum constraint coefficient and minimum constraint coefficient) is clearly rejected. Therefore, there is a significant difference between the methods mentioned above for the selection of seed nodes.

On the other hand, according to the box plots presented by figures 2, 3, 4, and 5 for each dataset, the adoption of a node by the random method, as the seed node of a community for diffusion, produces a higher LIE value than the seed node selection based on the minimum constraint coefficient. Moreover, a seed node with a maximum degree induces a greater LIE value than the selection of a random node as a seed node. And finally selecting the seed node with the maximum constraint coefficient value is more effective than selecting the node with the maximum degree.

Comparison of different methods of selecting the seed node: In order to ensure the significance of the aforementioned, Paired t-test is used. The results of implementing the above said nonparametric test on the results of the 20-time execution of the algorithm are summarized in table 3. Given the values of table 3 at the significant level of 5%, the zero assumption of all tests, except for karate, is rejected.

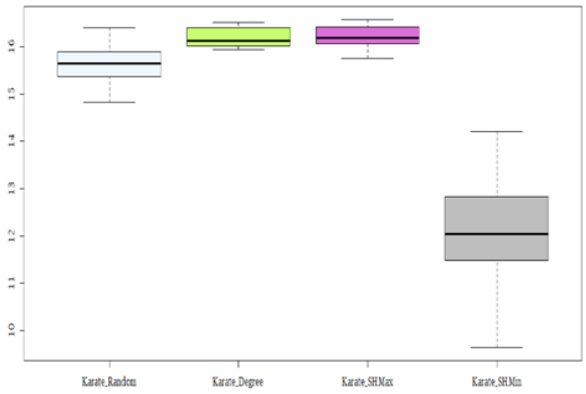


Fig 2. The boxplot of Karate dataset

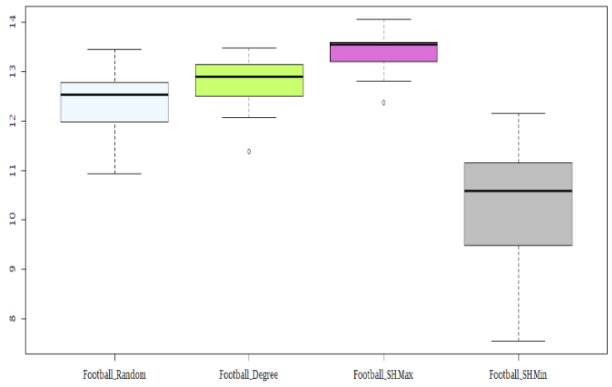


Fig 3. The boxplot of Dolphin dataset

Therefore, there is no reason to reject the superiority of the random method of selecting the seed node over the minimum constraint coefficient method, reject the superiority of the maximum constraint coefficient method over the random selection of the seed node method, and reject the superiority of the maximum constraint coefficient method over the maximum degree selection of the seed node method. So, the order of the effect of the maximum constraint coefficient, maximum degree centrality, random and minimum constraint coefficient is acceptable. This means seed node selection by the maximum constraint coefficient method has a greater effect on the diffusion process than the seed node selection with the maximum degree method. Also, selection of the seed node by the maximum degree method in the propagation process is more effective than selecting the seed node with a random method. And selecting the seed node with the random method also has a significant effect in the diffusion of the node selection with the minimum constraint coefficient. In the case of karate dataset, it should be noted that the result of p-value indicates that the adoption of seed node by the maximum degree method with the maximum constraint coefficient method will not be significantly different. However, this result does not reject the overall order.

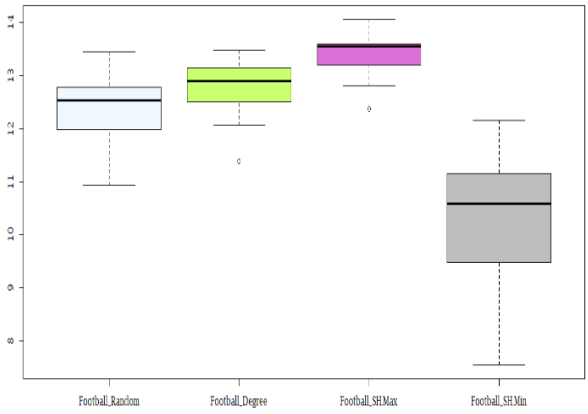


Fig 4. The boxplot of Football dataset

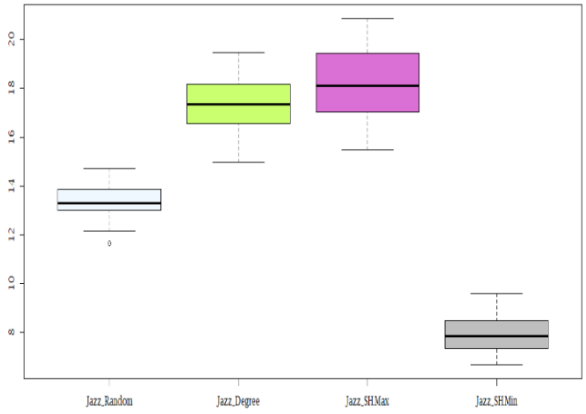


Fig 5. The boxplot of Jazz dataset

Table 2. Summary of Kruskal-Wallis rank sum test results for the 20-time execution of the algorithm

Kruskal – Wallis rank sum test:

$$\begin{cases} \boldsymbol{H_0}: \boldsymbol{\mu_{random}^{Sample} = \mu_{argmax\{Degree\}}^{Sample} = \mu_{argmax\{C_i\}}^{Sample} = \mu_{argmin\{C_i\}}^{sample} = \mu_0^{Sample}} \\ \boldsymbol{H_1}: \exists\ \boldsymbol{s \in S, \mu_s^{Sample} \neq \mu_0^{Sample}, where\ S = \{random, argmax\{Degree\}, argmax\{C_i\}, argmin\{C_i\}\}} \end{cases}$$

Sample Name	$\langle \chi^2_{Kruskal-Wallis} \vert p-value \rangle$
Karate	$\langle 57.686 \vert 1.834e-12 \rangle$
Football	$\langle 57.54 \vert 1.971e-12 \rangle$
Dolphin	$\langle 66.125 \vert 2.882e-14 \rangle$
Jazz	$\langle 67.603 \vert 1.391e-14 \rangle$

Table 3. Performing a nonparametric test on the results of a 20-time implementation of the algorithm

Sample Name	$H_0: \mu^{sample}_{argmin\{C_i\}} \geq \mu^{Sample}_{Random}$	$H_0: \mu^{Sample}_{Random} \geq \mu^{Sample}_{argmax\{Degree\}}$	$H_0: \mu^{Sample}_{argmax\{Degree\}} \geq \mu^{Sample}_{argmax\{C_i\}}$
Karate	3.393e-08	9.578e-06	0.2668
Football	1.973e-09	0.03375	1.246e-05
Dolphin	7.254e-12	1.008e-09	0.02978
Jazz	3.388e-08	3.388e-08	0.02825

Nodes with a minimum constraint coefficient are usually considered to be more effective nodes. But the results show that these nodes do not play a role in the propagation process. The diagram of the process of improving the average impact corresponds to the best results of the 20-time execution of the algorithm for each of the four seed node selection methods for each of the 4 datasets is depicted in figures 6, 7, 8, and 9. In all cases, the above arrangement is clearly visible.

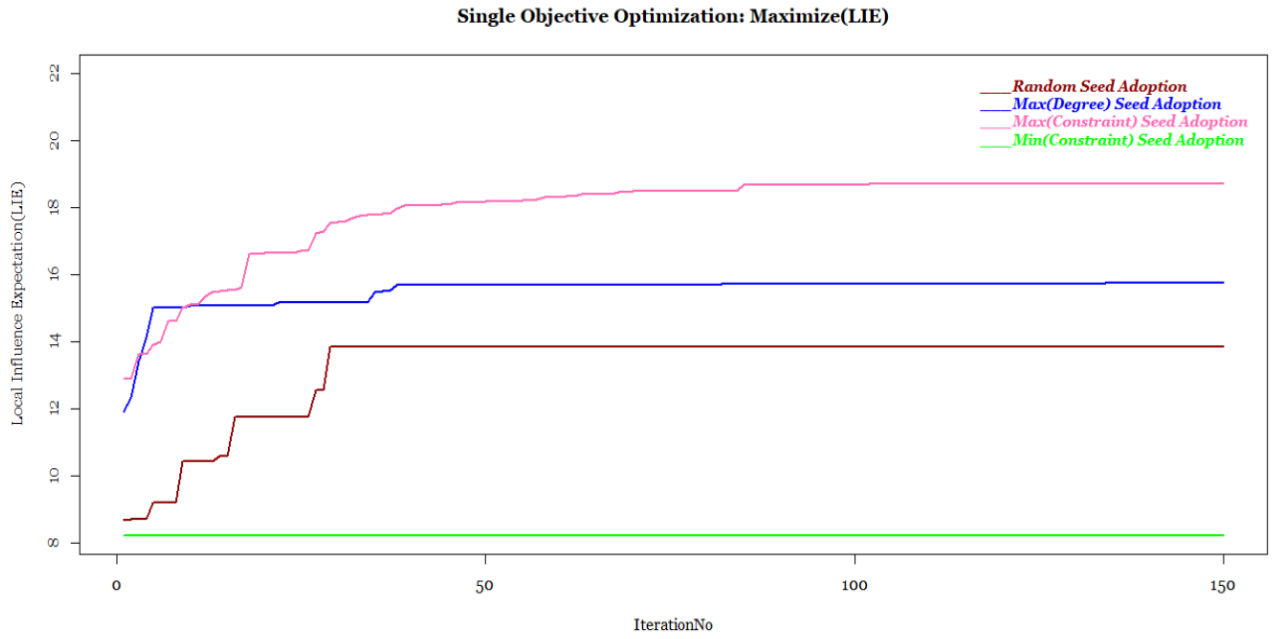


Fig 6. The diagram of the process of improving the average impact Corresponds to the best results of the 20-time execution of the algorithm for Karate dataset

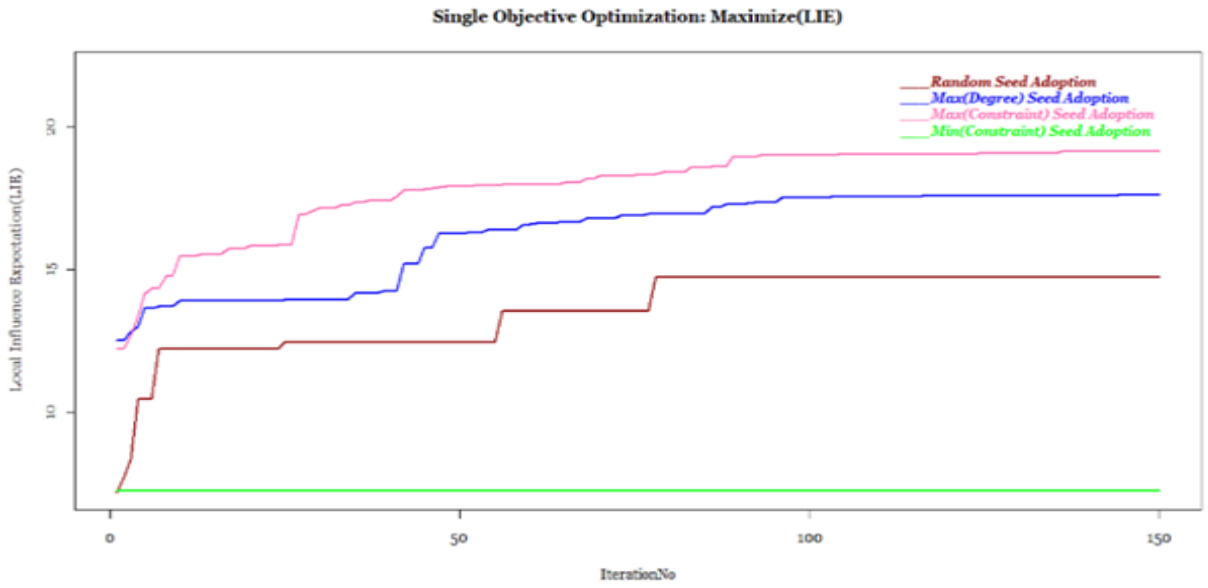


Fig 7. The diagram of the process of improving the average impact Corresponds to the best results of the 20-time execution of the algorithm for Dolphin dataset.

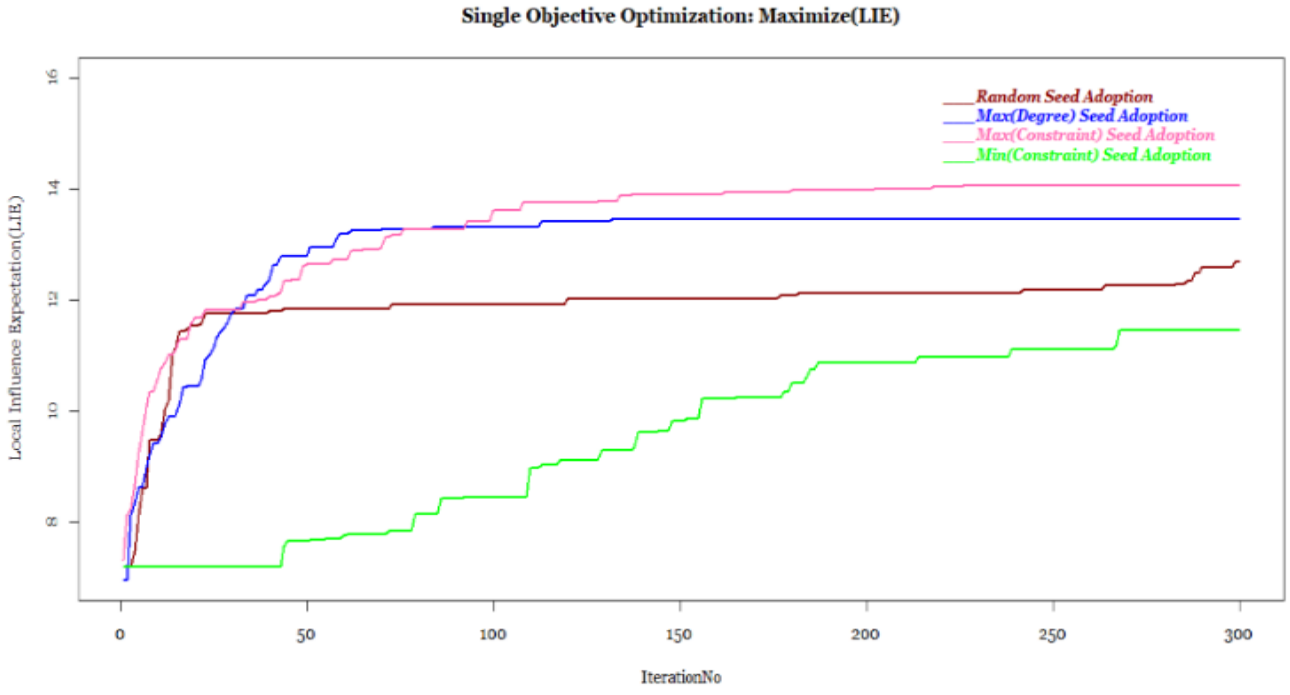


Fig 8. The diagram of the process of improving the average impact Corresponds to the best results of the 20-time execution of the algorithm for Football dataset

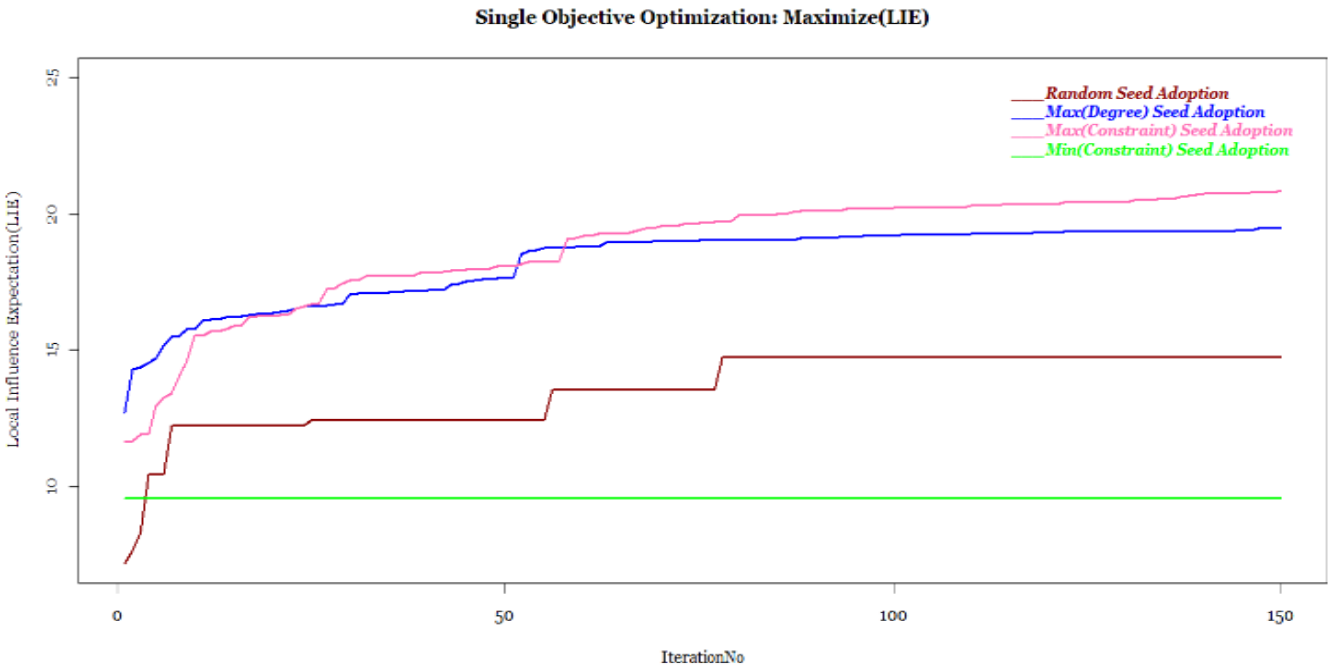


Fig 9. The diagram of the process of improving the average impact Corresponds to the best results of the 20-time execution of the algorithm for Jazz dataset

Our community-based measure vs centrality measures: we make a comparison between our community-based influencers detection and the famous centrality measures. For this purpose, we compare our method with page-rank, degree, betweenness, and closeness centrality measures. The comparison method was conducted in the following manner. First, we run our algorithm on each dataset, but to obtain clearer results we also run the algorithm on the email dataset. Then, we find the seed node set (k). The seed node set (k) is equal to the number of nodes that are affected by the SH_{max} method, which have the maximum information diffusion. We sort each centrality measure in a descending order up to the number of nodes in the seed set and for each number of seed nodes up to the seed set cardinality where the LIE function is calculated. For example, in the first row of table 4, we calculate the LIE function for the seed set including the best centrality measure, and the second row is the LIE function of seed set of the two best nodes in terms of each centrality, ...etc. In this manner, accumulatively, the influence of seed set, which is selected in terms of centrality measures, is computed. Therefore, the comparison between our method and the centrality measures is feasible and fair. Then the comparison of our methods with the centrality measures for karate, dolphin, football, and jazz and email datasets was arranged respectively in tables 4, 5, 6, 7, and 8. According to the trends of the LIE function exhibited by figures 10, 11, 12, 13, and 14, it is clear that our community-based method for influencers detection is not dominated by any centrality measure and in all datasets the LIE trend produced by our method (selecting the influential nodes with SH_{max} method) is superior to the LIE yielded by centrality measures.

Table 4. The Influence of seed nodes for Karate datasets

k	LIE				
	Degree	Betweenness	Closeness	PageRank	SH_{max}
1	3.00	3.00	2.98	3.00	3.02
2	7.25	7.13	7.01	7.11	7.57
3	8.96	8.65	8.44	9.18	9.80

Table 5. The Influence of seed nodes for Dolphin datasets

k	LIE				
	Degree	Betweenness	Closeness	PageRank	SH_{max}
1	3.31	3.46	3.06	3.16	3.64
2	7.35	6.98	6.82	5.90	8.50
3	10.02	9.36	9.04	9.33	10.59
4	13.47	13.14	13.31	13.09	13.77
5	16.18	15.79	15.45	14.73	16.54
6	17.53	17.63	17.86	17.62	18.17
7	20.46	20.53	19.44	19.12	20.99
8	21.70	21.89	21.40	21.85	22.39
9	24.04	23.45	24.76	24.68	25.91
10	27.99	28.75	28.94	28.26	29.24
11	30.78	28.89	29.06	28.08	31.36

Table 6. The Influence of seed nodes for Football datasets

k	LIE				
	Degree	Betweenness	Closeness	PageRank	SH_{max}
1	3.03	3.03	3.01	3.01	3.02
2	6.01	6.03	6.00	5.57	5.95
3	8.98	8.60	8.42	9.04	8.98
4	11.94	11.95	11.92	11.47	12.00
5	14.95	13.95	14.49	14.87	14.94
6	17.51	17.15	17.48	17.13	17.54
7	19.75	20.46	19.99	20.10	20.65
8	23.13	21.60	22.97	22.47	22.99
9	25.62	25.78	24.81	24.92	25.93
10	28.78	29.00	27.51	28.59	28.84
11	28.57	30.29	31.23	28.94	31.70
12	33.07	32.42	32.49	32.67	33.79

Table 7. The Influence of seed nodes for Jazz datasets

k	LIE				
	Degree	Betweenness	Closeness	PageRank	SH_{max}
1	3.03	3.03	3.01	3.01	3.02
2	6.01	6.03	6.00	5.57	5.95
3	8.98	8.60	8.42	9.04	8.98
4	11.94	11.95	11.92	11.47	12.00
5	14.95	13.95	14.49	14.87	14.94
6	17.51	17.15	17.48	17.13	17.54
7	19.75	20.46	19.99	20.10	20.65
8	23.13	21.60	22.97	22.47	22.99
9	25.62	25.78	24.81	24.92	25.93
10	28.78	29.00	27.51	28.59	28.84
11	28.57	30.29	31.23	28.94	31.70
12	33.07	32.42	32.49	32.67	33.79
13	34.46	35.95	34.09	35.03	37.99

The Influence of seed nodes for Email datasets is shown in appendix A.

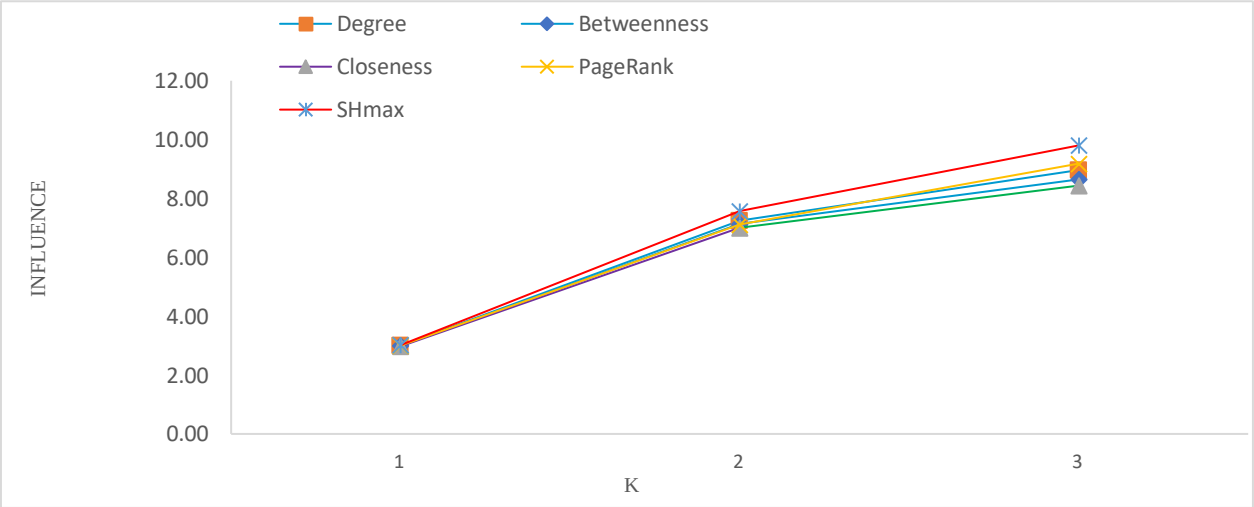


Fig 10. The influence spreading of different algorithms on Karate dataset

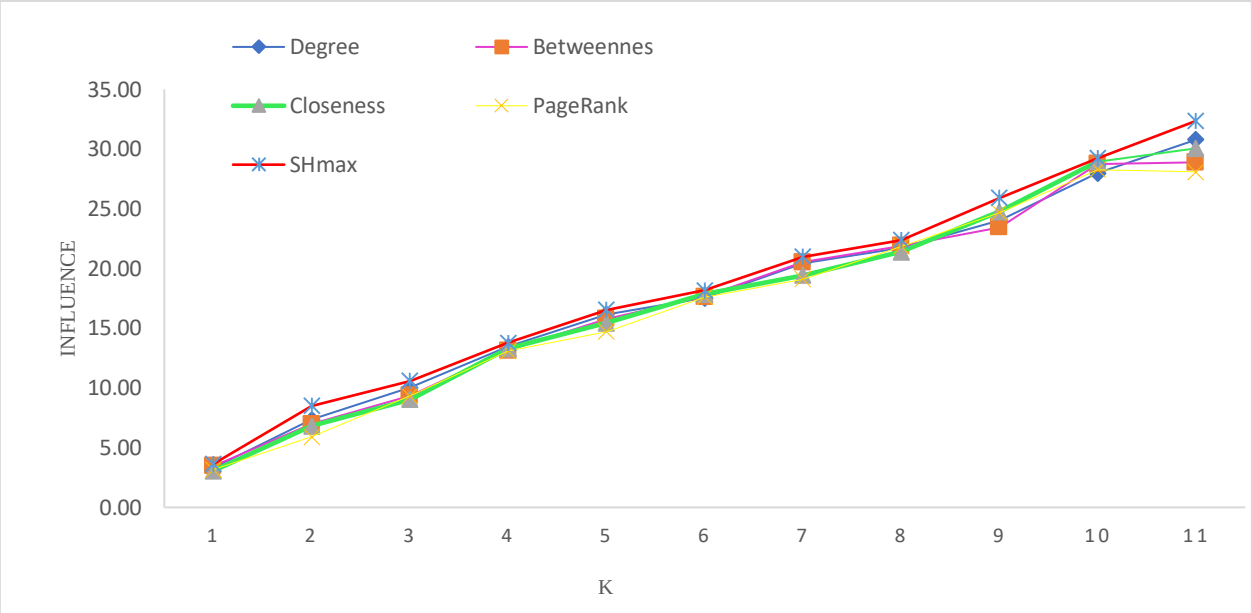


Fig 11. The influence spreading of different algorithms on Dolphin dataset

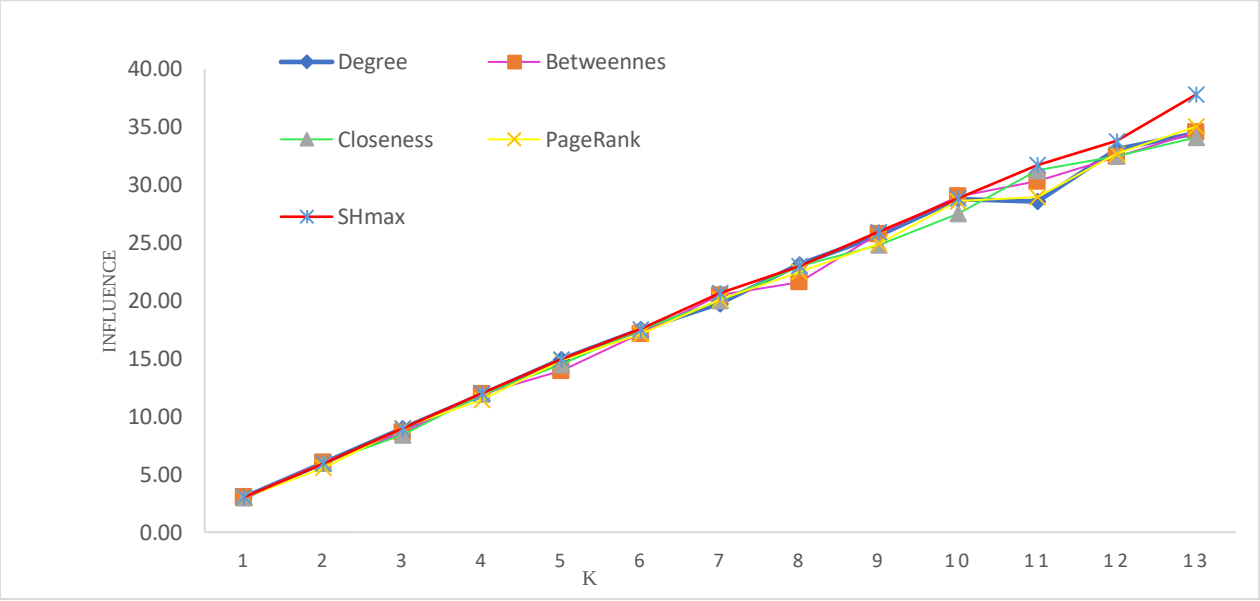


Fig13. The influence spreading of different algorithms on football dataset

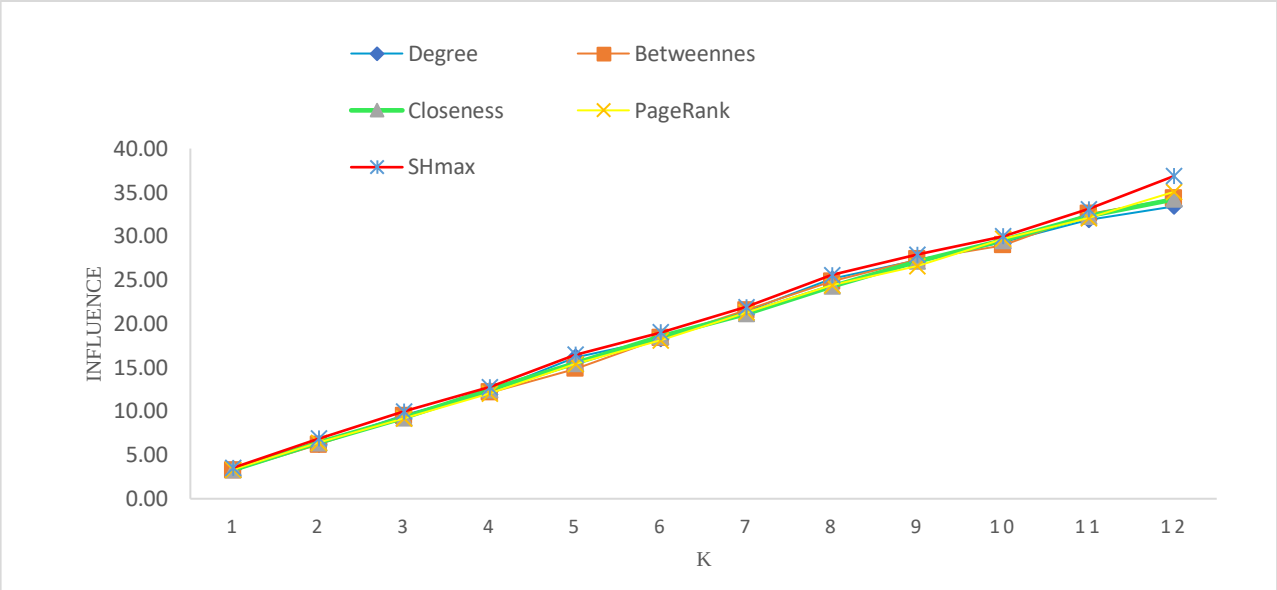


Fig 12. The influence spreading of different algorithms on jazz dataset

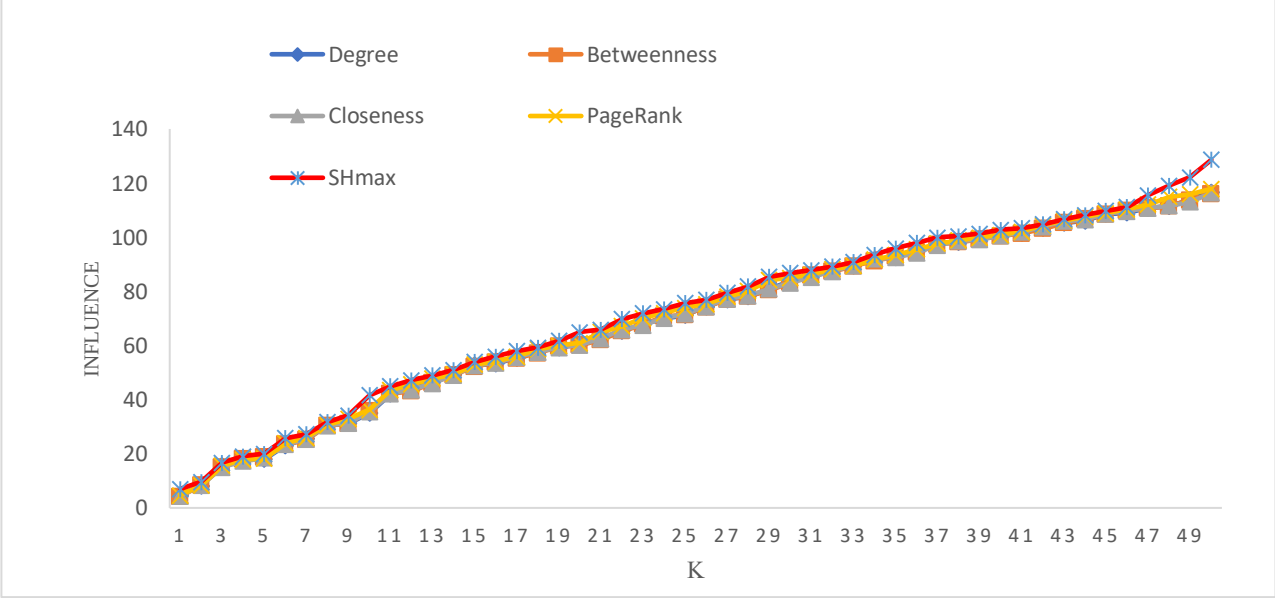


Fig 14. The influence spreading of different algorithms on Email dataset

5-Conclusion

Identifying the influential nodes in social network structures with the aim of influence maximization is one of the most commonly used methods in the field of social network analysis. This method is especially noticeable in the field of social commerce due to its application in the marketing and advertising of goods and services. Companies try to benefit from the effects of influencers for sale and advertising. Therefore,

the ability to identify influencers in online social networks has become valuable to e-commerce marketing companies. Owing to the channel of these networks, marketing information can be released faster and therefore promoted more effectively by influencers on online social networks through recommendations to their followers and peers. In this paper, we proposed a new algorithm based on a genetic algorithm for identifying influential nodes with the aim of influence maximization in complex networks with community structure. To identify the influential nodes in the community and compute their effect on influence propagation, we used three methods (degree centrality, random and structural hole). The results showed that seed node selection using the maximum constraint coefficient method has a greater effect on the diffusion process than using the maximum degree method. Furthermore, the selection of seed node by the maximum degree method in the propagation process is more effective than with a random method. And selecting the seed node with a random method also has a significant effect in the diffusion of the node selection with the minimum constraint coefficient. And we showed that a node with minimum constraint coefficient, unlike the nodes that have maximum constraint coefficient, has no important role in the propagation process. Our community-based influencers detection makes it possible to find more influential nodes with maximum diffusion than those suggested by page-rank and other centrality measures. Our work can be extended in the future by applying other metaheuristics to identify communities and review other centrality measures and their impact on the propagation process.

References

Ahajjam, S. and Badir, H. (2018) 'Identification of influential spreaders in complex networks using HybridRank algorithm', *Scientific Reports*. Springer US, (July), pp. 1–10. doi: 10.1038/s41598-018-30310-2.

Azaouzi, M. and Romdhane, L. Ben (2018) 'An Efficient Two-Phase Model for Computing Influential Nodes in Social Networks Using Social Actions', 33(2), pp. 286–304. doi: 10.1007/s11390-018-1820-9.

B, S. S. S. *et al.* (2019) *CoIM : Community-Based Influence Maximization in Social Networks*. Springer Singapore. doi: 10.1007/978-981-13-3143-5.

Bao, Z. K. *et al.* (2017) 'Identification of influential nodes in complex networks: Method from spreading probability viewpoint', *Physica A: Statistical Mechanics and its Applications*, 468, pp. 391–397. doi: 10.1016/j.physa.2016.10.086.

Bian, T. and Deng, Y. (2017) 'A new evidential methodology of identifying influential nodes in complex networks', *Chaos, Solitons and Fractals*. Elsevier Ltd, 103, pp. 101–110. doi: 10.1016/j.chaos.2017.05.040.

Burt, R. (1992) *Structural holes: the social structure of of competition*. Harvard University Press. Available at: https://books.google.com/books/about/Structural_Holes.html?id=FAhiz9FWDzMC (Accessed: 20 January 2018).

Cai, D. *et al.* (2017) 'A new method for identifying influential nodes based on D-S evidence theory', in *Proceedings of the 29th Chinese Control and Decision Conference, CCDC 2017*, pp. 4603–4609. doi: 10.1109/CCDC.2017.7979310.

D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. D. (2003) 'The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations Behav. Ecol. and Sociobiol., 54:396–405'.

Danon, P. M. G. and L. (2003) 'Community structure in jazz. Advances in Complex Systems Advances in Complex Systems, 6(4):565–573'.

Fei, L. *et al.* (2017) 'A new method to identify influential nodes based on combining of existing centrality measures', *Modern Physics Letters B*, 175024317. doi: 10.1142/S0217984917502438.

Goldenberg, J. *et al.* (no date) 'Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular', *search.proquest.com*. Available at: http://search.proquest.com/openview/613a0da9fb3ec5e34ca37c07f8585e77/1?pq-origsite=gscholar&cbl=25818 (Accessed: 28 January 2018).

Gong, M. *et al.* (2016) 'Influence maximization in social networks based on discrete particle swarm optimization', *Information Sciences*. Elsevier, 367–368, pp. 600–614. doi: 10.1016/j.ins.2016.07.012.

Goyal, A., Lu, W. and Lakshmanan, L. V. S. (2011) 'CELF++', in *Proceedings of the 20th international conference companion on World wide web - WWW '11*, p. 47. doi: 10.1145/1963192.1963217.

Halappanavar, M., Sathanur, A. V and Nandi, A. K. (2016) 'Accelerating the mining of influential nodes in complex networks through community detection', in *Proceedings of the ACM International Conference on Computing Frontiers - CF '16*, pp. 64–71. doi: 10.1145/2903150.2903181.

Handl, J. (2007) 'An evolutionary approach to multiobjective clustering', *ieeexplore.ieee.org*. Available at: http://ieeexplore.ieee.org/abstract/document/4079614/ (Accessed: 27 January 2018).

He, J. S., Kaur, H. and Talluri, M. (2016) 'Positive opinion influential node set selection for social networks: Considering both positive and negative relationships', in *Lecture Notes in Electrical Engineering*, pp. 935–948. doi: 10.1007/978-81-322-2580-5_85.

Hosseini-Pozveh, M., Zamanifar, K. and Naghsh-Nilchi, A. R. (2017) 'A community-based approach to identify the most influential nodes in social networks', *Journal of Information Science*, 43(2), pp. 204–220. doi: 10.1177/0165551515621005.

Hu, P. and Mei, T. (2018) 'Ranking influential nodes in complex networks with structural holes', *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 490, pp. 624–631. doi: 10.1016/j.physa.2017.08.049.

Huang, H. *et al.* (2019) 'Community-based influence maximization for viral marketing'. Applied Intelligence, pp. 9–12.

Jaouadi, M. and Ben Romdhane, L. (2017) 'DIN: An efficient algorithm for detecting influential nodes in social graphs using network structure and attributes', in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*. doi: 10.1109/AICCSA.2016.7945698.

Kaur, H., Talluri, M. and He, J. S. (2015) 'Blocking negative influential node set in social networks: From host perspective', in *2015 International Conference on Collaboration Technologies and Systems, CTS 2015*, pp. 472–473. doi: 10.1109/CTS.2015.7210470.

Kempe, D. and Kleinberg, J. (2003) 'Maximizing the Spread of Influence through a Social Network Categories and Subject Descriptors', *Science*. New York, New York, USA: ACM Press, pages, pp. 137–146. doi: 10.1145/956750.956769.

Leskovec, J. *et al.* (2007) 'Cost-effective Outbreak Detection in Networks'.

Liu, D., Jing, Y. and Chang, B. (2016) 'Identifying influential nodes in complex networks based on expansion factor', *International Journal of Modern Physics C*. IEEE, 27(9), p. 1650105. doi: 10.1142/S0129183116501059.

Liu, S. *et al.* (2015) 'Identifying effective influencers based on trust for electronic word-of-mouth marketing: A domain-aware approach', *Information Sciences*. Elsevier, 306, pp. 34–52. doi: 10.1016/j.ins.2015.01.034.

Newman, ichelle G. and M. E. J. (2002) 'Community structure in social and biological networks Proc. Natl. Acad. Sci. U.S.A., 99(12):7821–7826'.

Park, YoungJa, and M. S. (1998) ‘A genetic algorithm for clustering problems’.

Pei, S. *et al.* (2016) ‘Searching for superspreaders of information in real- world social media’, *nature.com*, pp. 1–34. doi: 10.1038/srep05547.

R. Guimera, L. Danon, A. Diaz-Guilera, F. G. and A. A. (2003) ‘Physical Review E’.

Reihanian, A., Feizi-Derakhshi, M. R. and Aghdasi, H. S. (2017) ‘Community detection in social networks with node attributes based on multi-objective biogeography based optimization’, *Engineering Applications of Artificial Intelligence*. Pergamon, 62, pp. 51–67. doi: 10.1016/j.engappai.2017.03.007.

Scripps, J. (2013) ‘Discovering Influential Nodes in Social Networks through Community Finding.’, *WEBIST*. Available at: https://pdfs.semanticscholar.org/fe1a/e2b2854e66c8e70eb32c0a6c08b9869e25e2.pdf (Accessed: 19 January 2018).

Shang, J. *et al.* (2017) ‘CoFIM: A community-based framework for influence maximization on large-scale networks’, *Knowledge-Based Systems*. Elsevier B.V., 117, pp. 88–100. doi: 10.1016/j.knosys.2016.09.029.

Singh, S. S. *et al.* (2019) ‘C2IM : Community based context-aware influence maximization in social networks’, *Physica A*. Elsevier B.V., 514, pp. 796–818. doi: 10.1016/j.physa.2018.09.142.

Song, G. *et al.* (2017) ‘Influential Node Tracking on Dynamic Social Network: An Interchange Greedy Approach’, *IEEE Transactions on Knowledge and Data Engineering*, 29(2), pp. 359–372. doi: 10.1109/TKDE.2016.2620141.

Ullah, F. and Lee, S. (2017) ‘Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization’, *Physica A: Statistical Mechanics and its Applications*, 486, pp. 968–985. doi: 10.1016/j.physa.2017.05.089.

Walker, S. K. (2011) ‘Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives’, *Journal of Family Theory & Review*, 3(3), pp. 220–224. doi: 10.1111/j.1756-2589.2011.00097.x.

Wang, Y. *et al.* (2010) ‘Community-based greedy algorithm for mining top-K influential nodes in mobile social networks’, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10*, p. 1039. doi: 10.1145/1835804.1835935.

Watts, D. (2002) ‘A simple model of global cascades on random networks’, *National Acad Sciences*. Available at: http://www.pnas.org/content/99/9/5766.short (Accessed: 28 January 2018).

Wei Li and Jianbin Huang (2015) ‘Mining top- k influential nodes in social networks via community detection Wei Li and Jianbin Huang * Shuzhen Wang’, 14, pp. 172–184.

Wenfeng Kang, Guangming Tang, Yifeng Sun, S. W. (2017) ‘Identifying Influential Nodes in Complex Networks Based on Weighted Formal Concept Analysis’, *IEEE Access*, 5, pp. 3777–3789. doi: 10.1109/ACCESS.2017.2679038.

YEH, J. and LIN, W. (2007) ‘Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department’, *Expert Systems with Applications*, 32(4), pp. 1073–1083. doi: 10.1016/j.eswa.2006.02.017.

Zachary, W. (1977) ‘An information flow model for conflict and fission in small groups’.

Zhang, X. *et al.* (2013) ‘Identifying influential nodes in complex networks with community structure’, *Knowledge-Based Systems*. Elsevier B.V., 42(4), pp. 74–84. doi: 10.1016/j.knosys.2013.01.017.

Zhao, J. *et al.* (2015) ‘Identifying influential nodes based on graph signal processing in complex networks’, *Chinese Physics B*, 24(5). doi: 10.1088/1674-1056/24/5/058904.

Zhao, X. *et al.* (2017) 'Evaluating Influential Nodes in Social Networks by Local Centrality with a Coefficient', *ISPRS International Journal of Geo-Information*, 6(2), p. 35. doi: 10.3390/ijgi6020035.

Zhao, Y., Li, S. and Jin, F. (2016) 'Identification of influential nodes in social networks with community structure based on label propagation', *Neurocomputing*. Elsevier, 210, pp. 34–44. doi: 10.1016/j.neucom.2015.11.125.

Zhou, C. *et al.* (2014) 'An Upper Bound based Greedy Algorithm for Mining Top-k Influential Nodes in Social Networks'. doi: 10.1145/2567948.2577336.

Zhou, J., Zhang, Y. and Cheng, J. (2014) 'Preference-based mining of top-K influential nodes in social networks', *Future Generation Computer Systems*. North-Holland, 31, pp. 40–47. doi: 10.1016/J.FUTURE.2012.06.011.

Zhu, J., Liu, Y. and Yin, X. (2017) 'A New Structure-Hole-Based Algorithm For Influence Maximization in Large Online Social Networks', *IEEE Access*, 5(c), pp. 23405–23412. doi: 10.1109/ACCESS.2017.2758353.

Appendix A: The Influence of seed nodes for Email datasets

k	LIE				
	Degree	Betweenness	Closeness	PageRank	SH_{max}
1	4.69	4.13	4.45	4.71	7.01
2	8.18	8.41	8.35	8.37	9.57
3	15.68	15.10	15.07	15.68	16.70
4	17.64	17.97	17.25	17.90	19.10
5	18.06	18.87	18.43	18.22	20.00
6	23.11	23.49	23.43	23.91	25.78
7	25.68	25.16	25.42	25.98	27.27
8	30.54	30.40	30.37	30.63	31.75
9	31.26	31.16	31.34	32.86	34.16
10	35.08	35.89	35.54	36.32	41.64
11	42.13	42.27	42.25	43.53	44.98
12	43.65	43.09	43.66	45.68	47.16
13	46.86	46.29	46.05	47.81	49.04
14	49.39	49.14	49.10	49.46	50.99
15	52.80	52.22	52.70	53.02	53.90
16	53.26	53.88	53.55	54.09	55.95
17	55.49	55.29	55.68	55.92	57.93
18	57.84	57.17	57.65	58.63	59.30
19	59.26	59.59	59.22	60.02	61.73
20	60.49	60.13	60.26	60.96	64.99
21	62.45	62.09	62.61	65.03	65.95
22	65.42	65.25	65.88	67.14	69.74
23	67.75	67.86	67.53	70.09	71.88
24	70.07	70.76	70.13	71.99	73.46
25	71.23	71.30	71.71	73.93	75.65
26	74.44	74.18	74.46	75.36	76.88
27	77.04	77.02	77.24	78.05	79.53
28	78.06	78.13	78.38	80.21	81.89
29	80.95	80.42	80.90	84.17	85.39
30	83.97	83.27	83.15	85.07	86.79
31	85.61	85.90	85.24	86.26	87.85
32	87.62	87.26	87.41	88.27	89.18
33	89.45	89.27	89.56	89.37	90.89
34	91.63	91.20	91.97	91.80	93.55
35	92.66	92.91	92.52	93.46	95.94
36	94.79	94.44	94.23	95.53	97.97
37	97.60	97.33	97.11	97.72	99.90
38	98.33	98.14	98.68	98.85	100.39
39	99.58	99.18	99.23	99.91	101.38
40	100.78	100.47	100.58	100.92	102.78
41	101.59	101.41	101.93	102.05	103.4
42	103.33	103.17	103.96	104.43	104.8
43	105.29	105.25	105.96	106.15	106.73
44	106.07	106.72	106.60	107.57	108.25
45	108.61	108.44	108.68	108.94	109.75

k	LIE				
	Degree	Betweenness	Closeness	PageRank	$\mathbf{SH_{max}}$
46	109.08	109.89	109.67	110.30	111.15
47	110.81	110.68	110.76	112.10	115.58
48	111.36	111.47	111.68	114.93	119.08
49	113.88	113.60	113.11	115.85	122.2
50	116.66	116.05	116.36	117.90	128.61