Article

JIS

Journal of Information Science
2017, Vol. 43(2) 204–220

DOI: 10.1177/0165551515621005
journals.sagepub.com/home/jis
SAGE

A community-based approach to identify the most influential nodes in social networks

Maryam Hosseini-Pozveh
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

Kamran Zamanifar
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

Ahmad Reza Naghsh-Nilchi
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

Abstract
One of the important issues concerning the spreading process in social networks is the influence maximization. This is the problem of identifying the set of the most influential nodes in order to begin the spreading process based on an information diffusion model in the social networks. In this study, two new methods considering the community structure of the social networks and influence-based closeness centrality measure of the nodes are presented to maximize the spread of influence on the multiplication threshold, minimum threshold and linear threshold information diffusion models. The main objective of this study is to improve the efficiency with respect to the run time while maintaining the accuracy of the final influence spread. Efficiency improvement is obtained by reducing the number of candidate nodes subject to evaluation in order to find the most influential. Experiments consist of two parts: first, the effectiveness of the proposed influence-based closeness centrality measure is established by comparing it with available centrality measures; second, the evaluations are conducted to compare the two proposed community-based methods with well-known benchmarks in the literature on the real datasets, leading to the results demonstrate the efficiency and effectiveness of these methods in maximizing the influence spread in social networks.

Keywords
Community structure; influential nodes; influence maximization; social networks

1. Introduction

Online social networks are an important part of modern daily life; accordingly, they are subject to many studies in different research entities. Considering the effect of the relationships between individuals (i.e. nodes) and their influence on each other, which means the influence spreads in the social networks, one of the major research concerns is the influence maximization problem. Influence maximization, first introduced by Domingos and Richardson [1, 2], is the problem of identifying the set of the most influential nodes in order to begin the spreading process from in a given social network. This issue has important real-world applications such as viral marketing [2, 3], finding proper contamination blockers [4, 5] and recommender systems [6, 7].

In addition to the concept of most influential nodes, known as the target set, and the concept of final expected influenced nodes, known as the influence function, another piece of this problem is the information diffusion model through which the spreading process in the social networks is simulated. In information diffusion models, influence acceptors

Corresponding author:
Kamran Zamanifar, Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan 81746-73441, Iran.
Email: zamanifar@eng.ui.ac.ir

are considered as active and non-influenced nodes are considered as inactive while these models determine the situation where a node switches from inactive to active state as well. The independent cascade model and linear threshold model are two well-known information diffusion models introduced by Kempe et al. [3]. It is demonstrated that the influence maximization problem is an NP-hard optimization problem where the influence function for both the independent cascade and linear threshold models is monotone and submodular [3]. When the optimization objective is to maximize a monotone submodular function, it is proved that a natural greedy hill-climbing algorithm can obtain a near-optimal response within 63% of accuracy in optimal solution. Such a greedy algorithm is introduced to solve influence maximization in the mentioned models by Kempe et al. [3].

There exist other information diffusion models where submodularity does not hold; therefore, the greedy hill climbing approach is not the proper approach for such models since it may not acquire an adequate response. The minimum threshold and multiplication threshold models are two instances of such non-submodular information diffusion models [8]. Accordingly, presenting adequate methods to maximize the influence spread in social networks regarding non-submodular information diffusion models is a challenge in this field. The two algorithms presented in this respect are the Shapley value-based influential nodes (SPIN) algorithm [8] and set-based coding genetic algorithm (SGA) [9]. Although the SGA outperforms the SPIN algorithm with respect to the quality of the final influence, its run time is high, even higher than the run time of the greedy algorithm.

Proposing more scalable methods regarding the mentioned non-submodular information diffusion models is the main objective of this study. Considering that the applied approach here follows a three phase manner – (a) detecting the communities of the social network, (b) identifying the candidate nodes based on the community structure and (c) selecting the target set from the candidate nodes, the main contributions of this study are outlined as follows:

- The concept of the blocked nodes in addition to active and inactive ones for the minimum threshold and multiplication threshold models is introduced.
- Considering the fact that an activated node can activate or block other inactive nodes, a method is presented to select the candidate nodes from the social network by introducing an influence-based closeness centrality measure for minimum and multiplication threshold models.
- Because the linear threshold model does not include the blocked nodes and its diffusion process is different from the two-above mentioned models, modifications including computing the influence-based closeness centrality measure based on live-edges outcomes [3] to reconcile the measure to this model are presented.
- In the target set identification phase, some heuristics are presented with respect to computing the spread of the influence of a given set of nodes in order to improve the run time of the SPIN and SGA.
- Experiments are conducted to evaluate the effectiveness of the influence-based closeness centrality measure where results indicate that this measure outperforms degree centrality and closeness centrality in finding the most influential nodes in the social networks. Additional experiments are conducted to evaluate the presented community-based methods where the results indicate a decrease in the run time in comparison with benchmark methods, while at the same time they produce final influence spread comparable with other methods under the minimum threshold model, multiplication threshold model and linear threshold model.

The rest of this paper is organized as follows. The related information diffusion models are presented in Section 2. The related works are reviewed in Section 3. In Section 4, the proposed methods for maximizing the influence spread considering linear threshold, minimum threshold and multiplication threshold models are described. The details of the experiments on real dataset are presented in Section 5. Finally, the paper is concluded in section 6.

2. Information diffusion model

Linear threshold, minimum threshold and multiplication threshold models are described as follows:

- **Linear threshold model** – In this model, a threshold value θ_v in [0, 1] is assigned to every node, indicating how much a node tends to be influenced by its active neighbours. Where there is a lack of sufficient information about the social network, this value is selected for all of the nodes uniformly at random. A node v is influenced by each of its neighbours w by a weight $b_{v,w}$, where $\sum_{w \text{ neighbours of } v} b_{v,w} \leq 1$ is held. Assuming A_0 to be the initial set of active nodes, the activation process proceeds in discrete steps as follows: in step t in addition to the active nodes of step $t-1$, every node v subject to activation formula $\sum_{w \text{ neighbours of } v} b_{v,w} \geq \theta_v$ will become active. This process continues until no other activation is possible. In this model the influence function is monotone and submodular [3].

- **Minimum threshold model –** In this model, a threshold value θ_v in [0, 1] is assigned to every node. Where there is a lack of sufficient information about the social network, this value is selected for all of the nodes uniformly at random. A node v is influenced by each of its neighbours w by a weight $b_{v,w}$, where, $\sum_{w \text{ neighbours of } v} b_{v,w} \leq 1$ is held. Assuming A_0 as the initial set of active nodes, the activation process proceeds in discrete steps as follows: in step t, in addition to the active nodes of step $t-1$, every node v subject to activation formula $min_{w \text{ active neighbours of } v}\{\alpha_w b_{v,w}\} \geq \theta_v$, where, $\alpha_w \geq 0$, will be active. This process continues until no other activation is possible. In this model influence function is monotone decreasing and non-submodular. In fact it is monotone decreasing supermodular [8].
- **Multiplication threshold model –** In this model, a threshold value θ_v in [0, 1] is assigned to every node. Where there is a lack of sufficient information about the social network, this value is selected uniformly at random for all of the nodes. A node v is influenced by each of its neighbors w by a weight $b_{v,w}$, where, $\sum_{w \text{ neighbours of } v} b_{v,w} \leq 1$ is held. Assuming A_0 to be the initial set of active nodes, the activation process proceeds in discrete steps as follows: in step t, in addition to the active nodes of step $t-1$, every node v subject to activation formula $\prod_{w \text{ active neighbours of } v} b_{v,w} \geq \theta_v$ will be active. This process continues until no other activation is possible. In this model influence function is monotone decreasing and non-submodular. In fact it is monotone decreasing supermodular [8].

3. Related works

The influence maximization problem was introduced by Domingson and Richardson [1, 2] and the preliminaries of this problem considering information diffusion models were first presented by Kempe et al. [3]. Kempe et al. [3] introduced some information diffusion models including the two famous ones: the linear threshold model and the independent cascade model, where they demonstrated the monotonicity and submodularity of the influence function, $\sigma(.)$, for this NP-hard optimization problem. Function $f(.)$ is monotone provided that $f(S \cup \{v\}) \geq f(S)$ is held for all elements v, and is submodular provided that $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$ is held for all elements v and all pairs of sets S and T where $S \subseteq T$. They presented a greedy hill climbing algorithm (GHA), to maximize the influence spread in social networks since this approach is the best one for maximizing the monotone submodular functions. However, the run time of the GHA is high considering that: (a) all the nodes of the social network are evaluated in order to select one of them as the seed node and this procedure is repeated k times where k is the number of the nodes of the target set; and (b) the estimated final influence resulting from a given target set is gained through the Monte Carlo simulation method which is very time consuming [3].

Presenting methods with lower run time in comparison with the run time of the GHA is the focus of many studies in the context of the influence maximization problem. These studies usually present their solutions by focusing on the specific properties of the different information diffusion models in addition to the submodularity property [10–19].

Presenting solutions that consider the community structure of the social networks is another approach in order to decrease the time complexity in this context. Almost all of the studies in this group follow a three-phase approach implicitly or explicitly: (a) applying a specific algorithm to find the community structure of the social networks; (b) reducing the number of nodes subject to evaluation as target ones from the number of the all nodes of the social network to the number of elements of a small set of candidate nodes which are selected from different communities based on specific metrics applied in the scope of that communities; and (c) designing a method to select the final target nodes from the candidate nodes set. Scripps et al. [20] introduced community-based node roles in order to select the most influential nodes based on these roles. Cao et al. [21] presented their solution considering community structure where the communities of the social networks are supposed to be disconnected. The set of candidate nodes included k nodes from each one of the communities which are selected by applying the degree centrality measure. Then, the target set is identified under the IC model by mapping the influence maximization problem to a resource allocation problem which is solved by a proposed dynamic programming algorithm. Yan et al. [22] first applied kernel k-means algorithm to partition the social network into k clusters and then selected the most influential node from each cluster by applying the greedy algorithm under the multiple spread model. Chen et al. [23] proposed methods that decrease the number of candidate nodes subject to be influential based on the community structure considering the heat diffusion model. Zhang et al. [24] proposed that the k most influential nodes under the IC model are selected as the k centres of the k clusters resulting from applying the k-medoid clustering algorithm on the transfer probability matrix. Each element of this $n \times n$ matrix, where n is the number of the nodes of the social network, is the information transfer probability between two nodes from all the paths between them computed under the IC model by applying bond percolation process. Lv et al. [25] considered each community of the social network as a player of a cooperative game and defined a measure based on the number of nodes and weight density of each community as the Shapley value of that community. The number of the nodes that is

selected as target nodes from each community is proportional to its Shapley value. Then, in each community, the nodes are selected from two different groups of nodes: (a) bridge nodes; and (b) influential nodes obtained through applying the MixedGreedy [11]. Cong et al. [26], regarding mobile social networks, introduce an information diffusion-based community detection approach under the IC model. Then, a dynamic programming algorithm is proposed to choose the community which the next influential node belongs to. In addition, information propagation is parallelized to improve the run time further and an effectiveness improvement is presented by considering influence between communities. Chen at al. [27] consider the influence maximization under the heat diffusion model and develop a hierarchical community detection algorithm named H-clustering to detect the communities of the social network effectively. Next, by considering the size of the communities and the connection between them, candidate set generation is conducted through specifying some significant communities and selecting candidate nodes from them. Finally, the target set is selected heuristically where target nodes are identified based on their position in the communities. Rahimkhani et al. [28] proposed a method where the candidate nodes are selected by applying the degree of centrality measure in the communities where the number of selected nodes from each community is proportional to its size. In addition, a method is presented to improve the run time of computing the influence from a given seed set. It is noticeable that in the most of this group of studies, some of the effectiveness of final influence is lost for improving the run time while in some of them [21, 24, 28] the effectiveness is not lost.

The greedy hill climbing approach is a guaranteed method to get an acceptable near-optimal solution for a given monotone submodular function to be optimized, while this approach may not get the proper results on non-submodular functions [8]. The minimum threshold and multiplication threshold models introduced by Narayanam and Narahari [8] are the two instances of information diffusion models with non-submodular influence function. Considering these two models, the Shapley value-based influential nodes (SPIN) algorithm is presented by Narayanam and Narahari [8] and the set-based coding genetic algorithm (SGA) is presented by Wang et al. [9] to maximize the spread of influence. SPIN selects k nodes with top Shapley values as the target set where information diffusion in the social network is considered as a cooperative game [8] and SGA searches the space including all the k-element subsets of the set of the nodes in order to find the optimal target set [9]. In both studies, the evaluations are conducted for both submodular (linear threshold model) and non-submodular (minimum threshold model and multiplication threshold model) information diffusion models. The evaluation results indicate that the effectiveness of the SGA is better than that of the SPIN in both kinds of submodular and non-submodular models, better than greedy algorithm in non-submodular information diffusion models, and as good as the CELF version of the greedy algorithm in the linear threshold model [8, 9]. As reported in Wang et al. [9], the effectiveness of the SPIN is not as good as the greedy algorithm and SGA in both kinds of models. The advantage of the SPIN is its efficiency with respect to the run time, which is illustrated by comparing it with the CELF [10] in the linear threshold model on various datasets, while the time complexity of the SGA is even greater than in the greedy algorithm [8, 9].

Considering the non-submodular information diffusion models (i.e. minimum threshold model and multiplication threshold model), here a community structure-aware approach is presented in order to improve the efficiency with respect to the run time. Then, two methods, C-SPIN (community-aware SPIN) and C-SGA (community-aware SGA), are presented based on this approach by improving the SPIN and SGA algorithms, respectively. These proposed methods maintain the effectiveness of the results as well. These methods are applied on the linear threshold model in order to evaluate their generality for both submodular and non-submodular information diffusion models as well.

4. The proposed methods

The proposed methods consist of three main phases: community detection, candidate seeds selection and target set identification (Figure 1). In the first phase, any available community detection algorithm based on the link structure which is proper for the applied datasets considering the run time can be adopted. In the candidate seeds selection phase, in brief, the concept of blocked nodes is introduced in minimum and multiplication threshold models. Next, based on the number and distance of the other nodes which are activated or blocked by the nodes in the extent of the communities, a centrality measure is presented to rank the nodes and select the candidate ones. This measure is also applied with some modifications to the linear threshold model, where blocked nodes do not exist. Finally, in the target set identification phase, a heuristic approach is presented to speed up the influence spread computation method.

4.1. Community detection

There are various methods to detect the communities of a target social network in the literature. In this study, the community finding is conducted based on the link structure. After applying the community finding procedure, which is selected

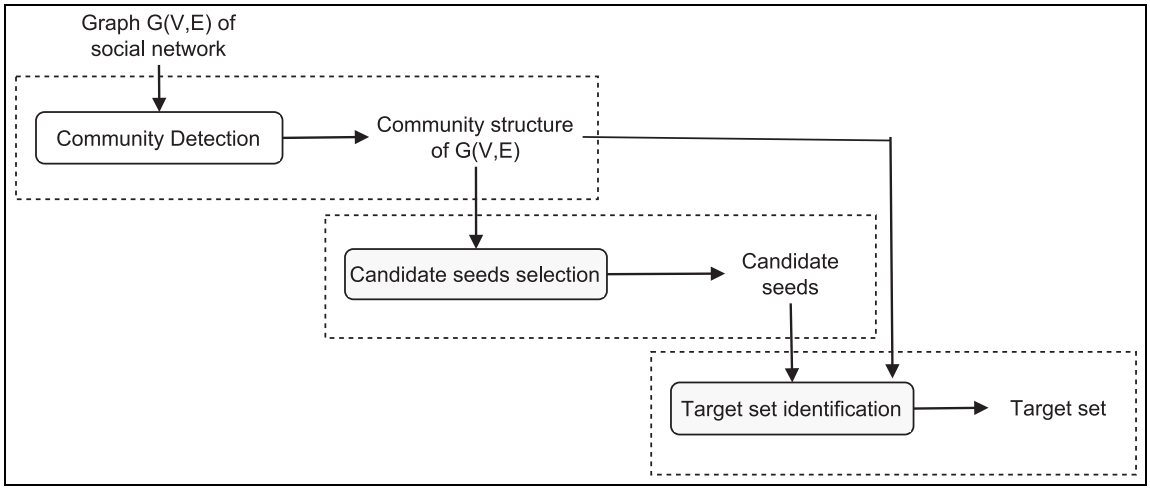

Figure 1. Three phases of the approach: (a) community detection; (b) candidate seeds selection; and (c) target set identification.

from the available methods in the literature, the disconnected partitions in addition to the connected ones are determined as well. A disconnected partition can be composed of one community or more than one community. Therefore, two types of communities are determined: disconnected and connected. Also, it is determined which community belongs to which disconnected partition.

4.2. Candidate seeds selection

4.2.1. The nodes: inactive, active, blocked. In both the minimum threshold and multiplication threshold models, the inactive state, addressed in these models, can itself be considered as two different states: (a) a node is in the first state when none of its neighbours has tried to activate it yet – this state is named as the 'inactive' as before; and (b) when a node is inactive and never will be activated in the future – this is explained as when one of the neighbours of a node (or more than one of them simultaneously) has tried to activate it but has failed. This state is named a 'blocked' state.

In the minimum threshold model, regardless of the values of α_w, owing to the activation formula $min_{w \text{ active neighbours of } v}\{\alpha_w b_{v,w}\} \geq \theta_v$ regarding node v, if v after receiving the first attempt from its neighbours which try to activate it, is not activated, then the condition $min_{w \text{ active neighbours of } v}\{\alpha_w b_{v,w}\} < \theta_v$ will not change anymore; therefore, v will not be activated in the future steps as well. This process is similar for the multiplication threshold model. Here, owing to the activation formula $\prod_{w \text{ active neighbours of } v} b_{v,w} \geq \theta$, if v after receiving the first attempt from its neighbours which try to activate it, is not activated, then the condition $\prod_{w \text{ active neighbours of } v} b_{v,w} < \theta$ will not change anymore since multiplying a number between 0 and 1 by another number between 0 and 1 will decrease its value. Accordingly, in both of these models, the nodes can be considered in three states: active, inactive and blocked. Moreover, an active node can activate or block other nodes across the social network. For instance, the active node v activates a node w and then w tries to activate x but fails; therefore x will be blocked. This means that v has blocked x in two steps (i.e. after two steps of information diffusion iteration).

4.2.2. The measures for selecting nodes in scope of the communities. Under the minimum and multiplication threshold models, the nodes activated or blocked by any two active nodes may have conflicting overlap with each other. Closeness of the nodes plays an important role here. A subgraph G presenting the nodes that can be influenced by nodes A and B are shown in Figure 2. Considering the minimum threshold model where α_A, α_B and α_C are assumed to be 1, node A activates C and D, and node B activates E but blocks D; therefore, when target set includes both A and B, the node D will not be activated by A because it has been blocked earlier by B.

The degree centrality [29] cannot consider the mentioned issue, since it is only aware of the number of outgoing edges from the nodes. Closeness centrality [29], another well-known centrality measure in the social network analysis domain, considers the closeness of a node to all the other nodes based on the link structure of the social network. Here, it should be noted that the degree centrality can be considered as an instance of the closeness centrality where all of its computations are made at the maximum with a limitation of distance 1 from each node. Obviously, these two measures are not

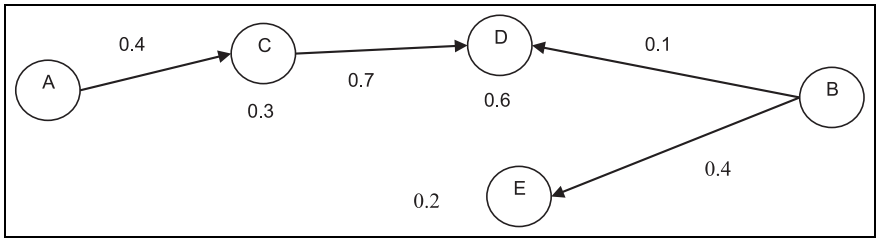

Figure 2. Subgraph G.

aware of the active and blocked nodes and the information diffusion models. In addition, according to the obtained results of this study, none of the above-mentioned measures in relation to each other produces a better result in influence maximization context in all applied datasets. In fact, degree centrality produces better results in most cases. This fact implies that computing the closeness of all the nodes of the social network from a node is not better than those with only one step distance from that node. This occurs because a specific node is not influential on all other nodes of the social network. Heuristically, the influence probability of the path from the source node to the destination node is the important neglected factor here; that is, wherever the length of the path from the source node to the destination node is prolonged, the probability of influencing the source node on the distention node becomes lower and the destination node exits from the influence scope of the source node. Therefore, an influence-based version of closeness centrality is applied here. Influence-based closeness centrality measure in the previous work through the authors of this paper [30] was proposed and applied to design a heuristic mutation operator for improving the performance of the PSO method to maximize the spread of influence in the signed social networks. In this study, this measure is adopted to supersede the closeness centrality as a heuristic centrality measure in order to select the seed nodes in the scope of the communities. In addition, a parameter λ is applied in the computations as an indicator of the main influence scope of the nodes. In this proposed measure, instead of applying distance 1 or maximum distance reachable in the social network, the influence-based closeness distance is considered. An influence-based closeness distance value relates to an influence path where the influence probability of that path is greater than a threshold value, λ. The influence probability of a path is computed by multiplying the influence probabilities (weights) of the available edges of that path. Any such measure that considers the influence flow between nodes beyond the connection between them can determine the influence power of the nodes more accurately. Therefore, the importance of the nodes is determined based on both the link structure (the community detection is conducted based on the link structure) and the flow of the information here.

First, the following measures are computed for each one of the nodes of the social network:

- Influence degree (ID) – considering C_v as the community of the node v where, v is the initial active node (i.e. target set includes only node v), the ID of the v is the number of final activated nodes in C_v where the influence probability of the influence path between v and each one of the activated nodes is greater than λ.
- Active influence-based closeness-aware degree (AICD) – considering C_v as the community of the node v where v is the initial active node (i.e. target set includes only node v), the AICD of the v is the closeness of all the final activated nodes in C_v to v where the influence probability of the influence path between v and each one of the activated nodes is greater than λ.
- Blocked influence-based closeness-aware degree (BICD) – considering C_v as the community of the node v where v is the initial active node (i.e. target set includes only node v), the BICD of the v is the closeness of all the final blocked nodes in C_v to v where the influence probability of the influence path between v and each one of the blocked nodes is greater than λ.

Influence-based closeness centrality in Hosseini-Pozveh et al. [30] was computed for an IC-based information diffusion model by applying the first iteration of the hill climbing greedy algorithm. Here, both the minimum and multiplication threshold models adopt this measure, which can be computed on a similar manner. Accordingly, the algorithm presented in Hosseini-Pozveh et al. [30] is applied here in its modified form in order to compute the three above-mentioned measures under the minimum and multiplication threshold models, considering the parameter λ. The details of the algorithm named InfluenceDegree, are presented in Figure 3. InfluenceDegree works similar for the both minimum and multiplication threshold models where in the minimum threshold model a node v is activated by its neighbour

Input: Graph $G(V,E)$ of a social network, set of communities of the G, Information diffusion model (minimum or multiplication threshold model).
Output: the obtained values of $ID, AICD$ and $BICD$ for all the nodes $v \in V$.
Method:
Initialize R=10,000
for each vertex $v \in V$ do // (v belongs to community C_v)
$ID_v = 0$
$AICD_v = 0$
$BICD_v = 0$
for i=1 to R do // Simulating diffusion process for R times in C_v started from node v.
$dist = 1$;
Do
$ID_v = ID_v + \sigma^A(v)_{step=dist\ \ where\ \ x}$
$AICD_v = AICD_v + ((\sigma^A(v)_{step=\text{dist}\ where\ x}) \times (1/\text{dist}))$
$BICD_v = BICD_v + ((\sigma^B(v)_{step=\text{dist}\ where\ x}) \times (1/\text{dist}))$
$dist++$
while (There exist any inactivated node which is a candidate for activation or blocking in C_v)
end for
$ID_v = ID_v/R$
$AICD_v = AICD_v/R$
$BICD_v = BICD_v/R$
end_for

Figure 3. The pseudocode of the InfluenceDegree algorithm.

w if $\alpha_w b_{v,w} \geq \theta_v$, otherwise it is blocked, and in the multiplication threshold model, a node v is activated by its neighbour w if $b_{v,w} \geq \theta_v$, otherwise it is blocked.

In the InfluenceDegree algorithm, assuming v to be the initial active node, $\sigma^{\mathrm{A}}(v)$ is the number of nodes activated by v and $\sigma^{\mathrm{B}}(v)$ is the number of blocked nodes by v. Here, condition x is (influence probability of the influence path from v to activated nodes) $> \lambda$ in term $\sigma^{\mathrm{A}}(v)_{\text{step} = \text{dist where } x}$ and is (influence probability of the influence path from v to blocked nodes) $> \lambda$ in term $\sigma^{\mathrm{B}}(v)_{\text{step} = \text{dist where } x}$. Therefore, $\sigma^{\mathrm{A}}(v)_{\text{step} = \text{dist where } x}$ is the number of activated nodes by v where their activation distance from v is equal to the dist value and the influence probability of their path from v is greater than the λ value. Moreover, $\sigma^{\mathrm{B}}(v)_{\text{step} = \text{dist where } x}$ is the number of blocked nodes by v where their blocking distance from v is equal to the dist value and the influence probability of their path from v is greater than the λ value. It is noticeable that computing the values of the two functions $\sigma^{\mathrm{A}}(v)_{\text{step} = \text{dist where } x}$ and $\sigma^{\mathrm{B}}(v)_{\text{step} = \text{dist where } x}$ means proceeding one more step (which is determined with parameter dist) in spreading process simulation under the given information diffusion model (here, the minimum or multiplication threshold model).

After computing the measures by the InfluenceDegree algorithm, conducted in the scope of the communities, the following relative measure is computed in the scope of the communities as well in order to rank the nodes in relation to each other:

- Relative influence-based closeness-aware degree RICD – the value of this measure for node v is determined in the extent of the community C_v through the following equation:

$$\mathrm{RICD}_v = \frac{1}{2}(\mathrm{RAICD}_v + \mathrm{RABICD}_v) \tag{1}$$

where RICD_v is the relative active influence-based closeness-aware degree and is computed for node v in the extent of the community C_v through the following equation,

$$\mathrm{RAICD}_v = \frac{\mathrm{AICD}_v}{\sum_{u \in C_v} \mathrm{AICD}_u} \tag{2}$$

and RABICD_v is the relative active to blocked influence-based closeness-aware degree and is computed for node v in the extent of the community C_v through the following equation,

$$\mathrm{RABICD}_v = \frac{\mathrm{AICD}_v/\mathrm{BICD}_v}{\sum_{u \in C_v} (\mathrm{AICD}_u/\mathrm{BICD}_u)} \tag{3}$$

Here, for nodes u and v, provided that the value of BICD is zero, the value of AICD is considered instead of AICD/BICD.

Input: Graph $G(V,E)$ of a social network, set of communities of the G.
Output: the obtained values of ID and $AICD$ for all the nodes $v \in V$.
Method:
Initialize R=10,000
for each vertex $v \in V$ do
    $ID_v = 0$
    $AICD_v = 0$
end_for
for r=1 to R do
    Generate live_edges $Outcome_r$ from G
    for each vertex $v \in V$ do // (v belongs to community C_v)
        $dist = 1$;
        Do
            $ID_v = ID_v + \sigma^A(v)_{step=dist \ \ where\ x}$
            $AICD_v = AICD_v + ((\sigma^A(v)_{step=\text{dist} \ \ where\ x}) \times (1/\text{dist}))$
            $dist++$
        while(There exist any inactivated node which is a candidate for activation in community C_v of $Outcome_r$)
    end_for
end_for
for each vertex $v \in V$ do
    $ID_v = ID_v/R$
    $AICD_v = AICD_v/R$
end_for

Figure 4. The pseudocode of the InfluenceDegree2 algorithm.

4.2.3. Measures for the linear threshold model. In this model, there are no blocked nodes; therefore, only ID, AICD and RAICD measures are computed. The computational details are as follows.

In the linear threshold model, the influence degree of a node cannot be obtained through the first iteration of the greedy hill climbing. In this respect, a different approach is presented based on constructing the outcomes of the live edges [3]. Each outcome is considered as equivalent to a round of the information diffusion process simulation. For the linear threshold model, this outcome is generated in a manner that each node v of the social network selects at most one of its income edges at random (among the neighbours of the node v, w, one is selected with probability $b_{v,w}$ and no edges is selected with probability $(1 - \sum_{w \text{ neighbours of } v} b_{v,w})$). The selected edge is live and the other ones are blocked. In such a subgraph of the social network which consists of the live edges, a node activates all the other nodes that can have access to them through an available path. Therefore, to compute the influence degree of the node v under the linear threshold model, an instance of the linear threshold model where all the weights of the edges and all the thresholds of the nodes are set to value 1 is run on the outcome subgraph where v is the initial active node (i.e. target set includes only node v). Here, the equivalent weight of every live edge in the main social network is applied in computing the influence probability of the paths. The details of the algorithm named InfluenceDegree2 to compute ID, and AICD are presented in Figure 4, where the measures are considered in the extent of the communities.

In the InfluenceDegree2 algorithm, when v is the initial active node, $\sigma^A(v)$ is the number of nodes activated by v and condition x is (influence probability of the influence path from v to activated nodes) $> \lambda$; therefore, $\sigma^A(v)_{\text{step} = \text{dist where } x}$ is the number of activated nodes by v where their activation distance from v is equal to the dist value and the influence probability of their path from v is greater than the λ value .

After computing ID and AICD, the relative measure RAICD is computed accordingly; here, RAICD is adopted instead of RICD.

4.2.4. The candidate seeds selection. The higher the number of the communities, the smaller the average size of the communities, and vice versa, which can affect the strategy of selecting the candidate nodes. Here, when k is the size of the target set, the number of selected nodes from each one of the communities changes from 1 to k, that is, a threshold $\theta_S = ck$ is adopted, where c is a constant value, and (a) k nodes are selected from each one of the communities the size of

which is more than or equal with *ck*, (b) $\lceil N_i/(max - Min)) \times \beta \rceil + \alpha$ nodes [28] are selected from each one of the communities with sizes less than *ck*, where, *max* and *min* are the number of the nodes of the largest and smallest communities and N_i is the number of the nodes of the *i*th community, and (c) when the number of the detected communities is high, the probability of the existence of very small communities (with sizes < 10 nodes or with very small sizes in comparison with *k*) becomes higher. The number of selected nodes from this kind of communities is between 1 and 2. Finally, if after applying these three rules, the number of selected candidate nodes is less than $1.5 \times k$, an appropriate number of communities matches with rule (b) are selected one by one randomly and the selected number of their nodes is increased to *k* until number of candidate nodes becomes $1.5 \times k$.

4.3. Target set identification

Here, the objective is to select the final target set from the candidate nodes. Under the minimum threshold and multiplication threshold models, the problem with a smaller-size set of nodes remains NP-hard and the influence function remains non-monotone and non-submodular, likewise. Therefore, algorithms such as SPIN or SGA can be applied here. Moreover, in order to decrease the run time of computing the final influence of a given target set in the social networks, the concept of fitness approximation (i.e. function approximation) [31] from the evolutionary computation domain is applied. Functional approximation is one of the approaches in this respect applied when fitness computation is very time consuming. In this approach, in each iteration, a portion of population is evaluated with its real objective function and another portion of population is evaluated with its approximate objective function [31]. This concept is applied here in both the SPIN and SGA. The details are presented as follows:

- Here, the objective function is approximated by computing the influence in a shorter distance from a target node instead of computing it in the whole social network domain. The three heuristic distances that are designated and applied instead of diameter of the social network are: (a) computing the influence of a target node up to the half of the maximum influence distance of the community which it belongs to; (b) computing the influence of a target node in the extent of the community which it belongs to; and (c) computing the influence of a target node one step further in the neighbouring communities in addition to the extent of the community which it belongs to.
- The identified community structure of the social network includes some essential information. The connected communities and disconnected ones in addition to the connected and disconnected partitions are identified based on the community structure. Moreover, in a heuristic sense, the large size of a community and its high influence on other communities constitute a more influential community. Communities with high influence range, for instance can be identified proportional to their bridge nodes. To compute this issue more accurately, the closeness of the communities in the social network is identified. For this purpose, a new graph where its nodes are the communities and its edges are the intra-community edges, is constructed. The closeness centrality of each community C_i (CC_{C_i}), is determined in addition to its size S_{C_i}. Next, the values are normalized as $CC_{C_i}/\sum_j CC_{C_j}$ and $S_{C_i}/\sum_j S_{C_j}$ and finally, the communities are sorted based on the value of $(1/2) \times ((CC_{C_i}/\sum_j CC_{C_j}) + (S_{C_i}/\sum_j S_{C_j}))$. The communities for which their corresponding sorted value is bigger than a threshold are identified as well.
- Considering SPIN, target nodes are selected based on the rank of their Shapley value where the players of the game are the candidate nodes instead of all the nodes of the social network. This newly introduced method is named C-SPIN for community-aware SPIN. In order to compute the contribution of the candidate nodes, a set of permutations of the candidate nodes,Ω, where its cardinality is polynomial in *cn* (*cn*is the number of the candidate nodes), is sampled on the random basis. Accordingly, a fixed number t_1 is selected and for each permutation, t_1 numbers of the places of the permutation are selected at random and the contribution of the nodes of these places is computed approximately. In this respect, if each one of these node has not been activated yet by any of the previous nodes in the current permutation, then: (a) each computation for a node is conducted in the disconnected part which that node belongs to (this issue is also considered for the nodes of other group); (b) for the nodes which belong to communities with low related sorted value, the contribution of a node is identified at random in order to be computed up to the half maximum or full maximum distance of its community; and (c) for the nodes that belong to communities with high relative sorted values, the contribution of a node is identified at random in order to be computed up to the half maximum, full maximum distance of its community, or in one step further in the neighbouring communities.
- Considering SGA, target set is selected through searching in the space which includes all the *k*-element subsets of the candidate node sets. In the original SGA, there exists the set of all the nodes of the social network instead of the set of candidate nodes. This newly introduced method is named C-SGA for community-aware SGA. When

the number of the population is p, a fixed number p_1 is selected and, in each iteration, the p_1 number of the chromosomes is selected on a random basis and the contribution of them is computed approximately. In this respect, the influence is computed separately in each community for summation. When only one node from a community exists in a chromosome, its ID value is used instead of computing the influence again.

- The methods are the same for minimum threshold, multiplication threshold and linear threshold models.

4.4. Time complexity

The running time of both the C-SPIN and C-SGA includes four segments: (a) community finding; (b) computing the value of measures for all the nodes (ID, AICD, BICD and RICD); (c) candidate nodes set identification; and (d) target set identification. In the first segment, the time complexity depends on the community detection algorithm which is applied, that is, the LabelRank algorithm [32] with the time complexity of $O(m)$ for undirected social networks and generalized LabelRank algorithm [33] with the time complexity of $O(m)$ for directed social networks where m is the number of edges of the social network in both the algorithms. The time complexity of the second segment consists of $O(R\sum_{i=1}^{N} n_i R m_i)$ which is an equivalent of $O(R\sum_{i=1}^{N} n_i m_i)$ for running the InfluenceDegree algorithm in addition to $O(n)$ for finding the final relative measures for all the nodes where N is the number of the detected communities, R is the number of Monte Carlo simulation iterations, n_i is the number of the nodes in community c_i, m_i is the number of the edges in community c_i and n is the number of the nodes in the social network. It should be noted that, because of using the parameter λ in InfluenceDegree algorithm, the run time can be considerably less than $O(R\sum_{i=1}^{N} n_i m_i)$. The time complexity of the third segment includes sorting the nodes in each community based on the identified measure and selecting at the most k top nodes from each one. In the fourth segment, the time complexities of C-SPIN and C-SGA vary. For C-SPIN, the running time is $O(T(p+m)R + plog(p) + kp + kRm)$ in comparison with the running time of the original SPIN $O(T(n+m)R + nlog(n) + kn + kRm)$, where T is the number of the permutations, p is the number of candidate nodes, n is the number of the nodes of the social network, R is the number of Monte Carlo simulation iterations, and m is the number of the edges of the social network. In addition, the value of p is at the most kN where N is the number of the nodes of the communities. The m here is the upper bound while owing to applying the distance oriented improvements, the real time is less than in some parts of the program. For C-SGA, the running time is $O(p_2 kmRT_2)$ in comparison with the running time of the original SGA $O(p_1 kmRT_1)$, where, T_1 and T_2 are the number of the iterations of the genetic algorithm, p_1 and p_2 are the number of chromosome, R is the number of Monte Carlo simulation iterations, and m is the number of the edges of the social network. The m here is the upper bound while, owing to applying the distance oriented improvements, the real time is reduced in some parts of the algorithm. Also, $p_2 < p_1$ and $T_2 < T_1$ since the dimensions of the problem is decreased from n to $O(kN)$.

5. Experimental evaluations

Datasets applied here are described in Section 5.1. The obtained community structure of the datasets is presented in Section 5.1 as well. Experiments are run to address the following two questions:

Can the influence-based closeness centrality measure better identify the influential nodes in a given social network in comparison with available centrality measures?
Can the proposed community-based methods outperform the benchmark methods with respect to both the efficiency (run time) and effectiveness (quality of response)?

The experiments are conducted on two types of information diffusion models: non-submodular (multiplication and minimum threshold models) and submodular (linear threshold model). In all of the mentioned models, the weight of the directed edge from node x to node y is $l(x, y)/d_y$ and the weight of the directed edge from node y to node x is $l(x, y)/d_x$, where d_x is the degree of the node x and d_y is the degree of the node y and $l(x, y)$ is the number of parallel edges between the nodes x and y [8]. The details of the experiments and the results are presented in the next three sections (5.2–5.4).

All of the implementations are performed in Java and run on a Dual-Core of Intel PC with a 2 GHz CPU and 4 GB memory.

5.1. Datasets

In order to follow the settings of two previous related studies [8, 9], the datasets used there are applied here. Real-world social network datasets, Political Books, Celegans, Netscience co-authorships, Zachary's Karate Club and Adjacency of

Table 1. Summary of the datasets used in the experiments

Dataset	Number of nodes	Number of edges	Directed
Political Books	105	441	No
Celegans	297	2359	Yes
Netscience	1589	2742	No
Zachary's Karate Club	34	78	No
Adjacency of nouns	112	425	No
Delicious	1867	7668	No

Table 2. Obtained community structure of datasets

Dataset	Applied community detection algorithm	Number of communities
Political Books	LabelRank [32]	3
Celegans	Generalized LabelRank [33]	12
Netscience	LabelRank [32]	281
Zachary's Karate Club	LabelRank [32]	2
Adjacency of nouns	LabelRank [32]	6
Delicious	LabelRank [32]	89

nouns [8, 9] (http://www-personal.umich.edu/~mejn/netdata/), are applied in this study. In addition, another real dataset, delicious bookmarks (http://grouplens.org/datasets/hetrec-2011/), is also used in order to evaluate the behaviour of the proposed methods on more real datasets. A summary of all the datasets is tabulated in Table 1.

The results of applying community detection algorithms are tabulated in Table 2. As mentioned before, LabelRank algorithm [32] is applied for undirected social networks and generalized LabelRank algorithm [33] is applied for directed social networks here. These algorithms are label propagation-based with $O(m)$ time complexity.

5.2. *Effectiveness of the influence-based closeness centrality measure*

Here, the experiments are conducted to evaluate whether the proposed influence-based closeness centrality (ICC) measure, considering the λ parameter, performs better than the original closeness centrality (CC) measure. In addition, degree centrality (DC) is applied here since it is a highly adopted measure in the influence maximization domain. The k most influential nodes are identified through each one of the methods and the number of final activated nodes gained through them, is computed.

The value of the λ parameter is set to 0.005 for all the datasets here. This value is selected heuristically based on the fact observed in the experiments on the datasets where selecting this value leads to counting almost all the nodes with distance 2 and most of the nodes with distance 3 from the initial active node. In larger datasets, some nodes with distance 4 and a few with distances 5 and 6 are included. Since the nodes are counted, provided that they are activated or blocked, this value covers the main scope influence of the nodes in these datasets.

It is noticeable that, when applying λ parameter in other datasets with sizes varying from average to very large, its value should be determined based on the specific characteristics of the dataset, such as the graph diameter of the social network.

The results of final activated nodes on the minimum threshold model and the multiplication threshold model are illustrated in Figures 5 and 6, respectively. The result of final activated nodes under the linear threshold model is illustrated in Figure 7.

It can be seen that, in some cases, the DC performs better than the CC measure and vice versa, while ICC outperforms both of them in all cases. The degree centrality method does not consider the influence overlap among the selected nodes where high degree nodes may be closely connected to each other. The closeness centrality measure does not consider the real influence paths or prefer the important ones to the others. Computing the closeness based on the information diffusion model leads to more accurate results, as confirmed by the results. The latter measure considers only influence paths for which their influence probability is as influential as a threshold value, λ.

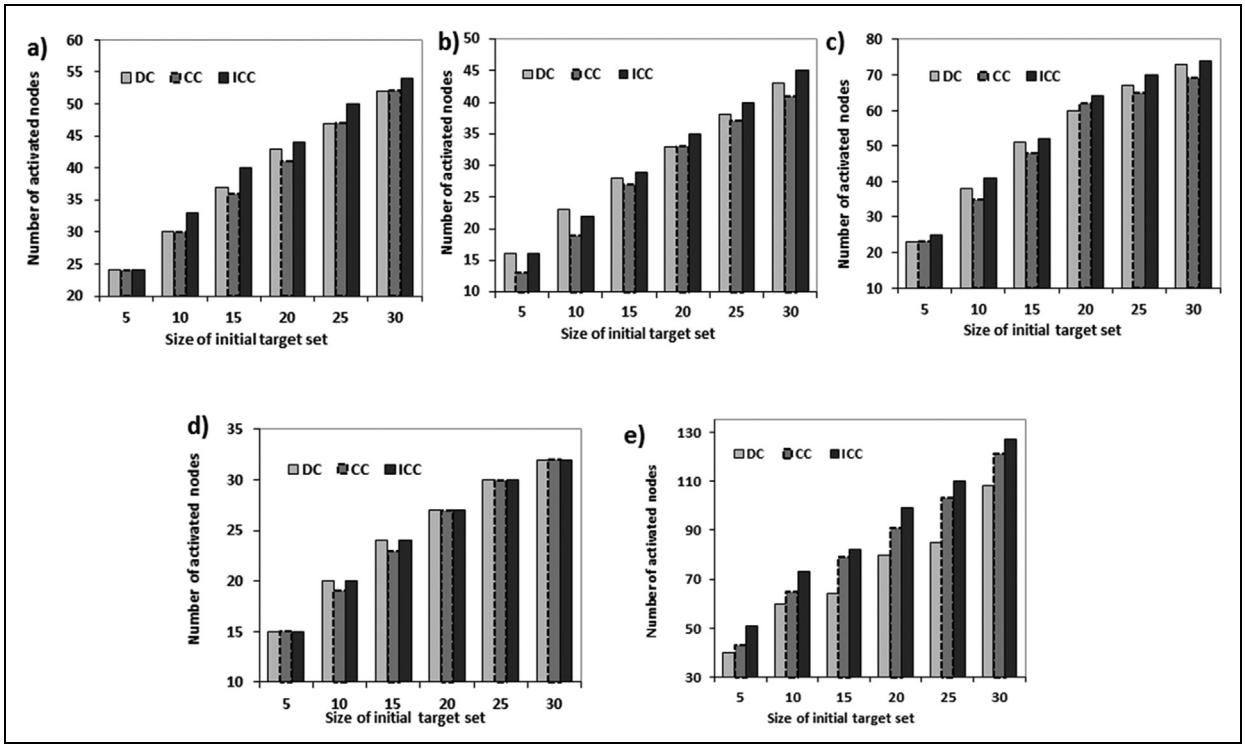

Figure 5. Number of active nodes vs the size of the initial target set under the minimum threshed model: (a) Adjacency of nouns dataset; (b) Political Books dataset; (c) Celegans dataset; (d) Karate dataset; and (e) Delicious Dataset.

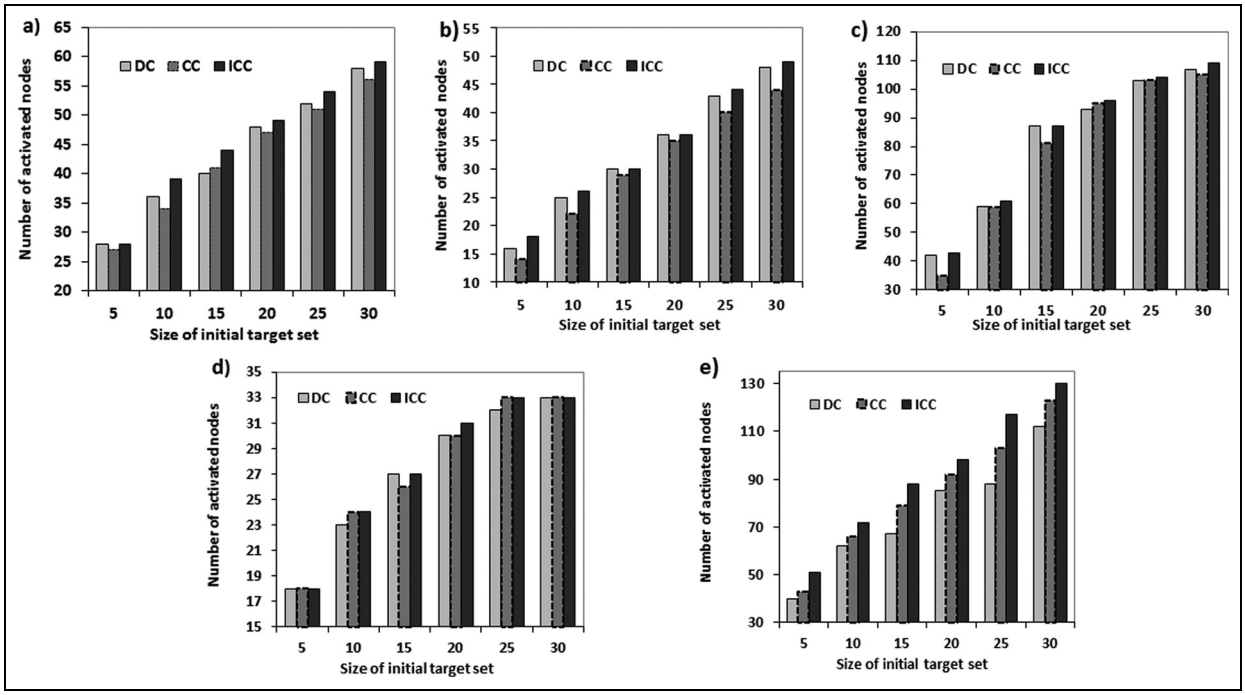

Figure 6. Number of active nodes vs the size of the initial target set under the multiplication threshed model: (a) Adjacency of nouns dataset; (b) Political Books dataset; (c) Celegans dataset; (d) Karate dataset; and (e) Delicious Dataset.

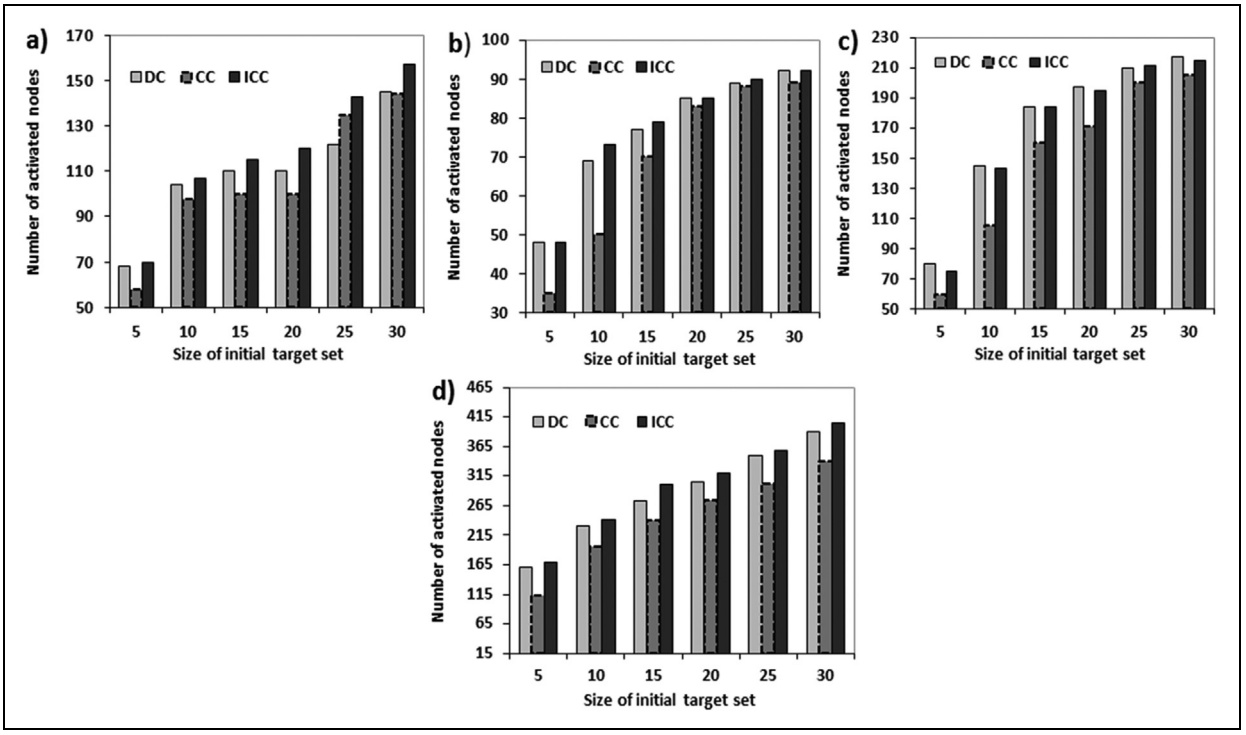

Figure 7. Number of active nodes vs the size of the initial target set under the Linear threshed model: (a) Netscience Curatorship dataset; (b) Political Books dataset; (c) Celegans dataset; and (d) Delicious dataset.

5.3. Evaluation of C-SPIN and C-SGA: results for non-submodular information diffusion models

The proposed methods, C-SPIN and C-SGA, here are compared with SPIN, SGA and GHA. Since the information diffusion models are not submodular, the improved versions of GHA where the submodularity property is applied with respect to run time cannot be used here. The experiments are conducted considering both the quality of response and the run time. The result of the final activated nodes for minimum threshold model is illustrated in Figure 8 and the result of the final activated nodes for multiplication threshold model is illustrated in Figure 9 on various datasets.

The result of running time for minimum threshold model is illustrated in Table 3 and the result of running time for multiplication threshold model is illustrated in Table 4 on different datasets.

As indicated by the results, C-SPIN and C-SGA have low run times while they produce results as well as SGA for all the datasets. The greedy hill climbing method does not produce acceptable results in all the cases and has a higher run time in comparison with C-SPIN and C-SGA. SPIN does not produce acceptable results in some cases while its community aware version produces acceptable results in all cases and has the lowest run time among all of the applied algorithms.

5.4. Evaluation of C-SPIN and C-SGA: results for submodular information diffusion model

Here, the proposed methods, C-SPIN and C-SGA, are compared with the original SPIN, the original SGA and CELF++ (the improvement version of GHA). The experiments are conducted considering both quality of response and run time. The result of final activated nodes for the linear threshold model on different datasets is illustrated in Figure 10. The result of running time for linear threshold model on different datasets is illustrated in Table 5.

In the linear threshold model, C-SPIN produces the best results regarding to both the run time and the final influence considering all of the applied datasets. The CELF++ and SGA have acceptable final influence as well, while their run time is high in comparison with the community aware SPIN method.

6. Conclusion

In this paper, a community-based approach is proposed to improve the run time efficiency in finding the target set in the social networks considering the multiplication threshold and minimum threshold models and the linear threshold

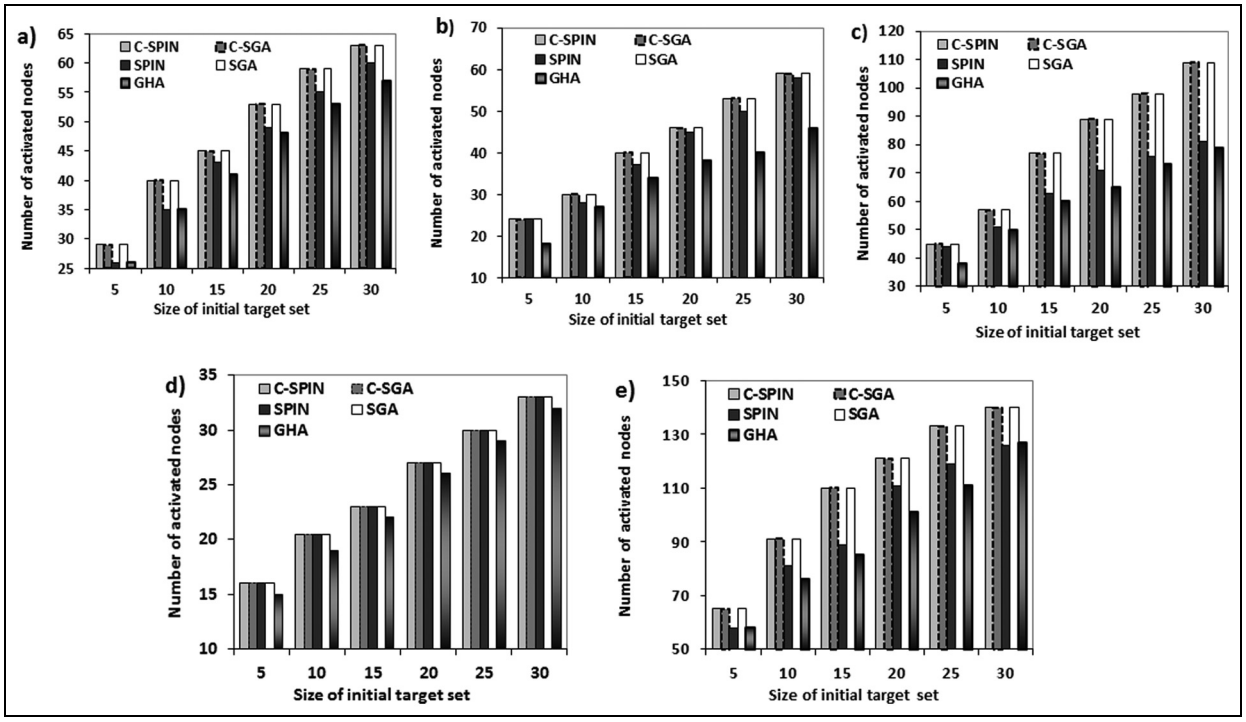

Figure 8. Number of active nodes vs the size of the initial target set under the minimum threshed model: (a) Adjacency of nouns dataset; (b) Political Books dataset; (c) Celegans dataset; (d) Karate dataset; (e) Delicious Dataset.

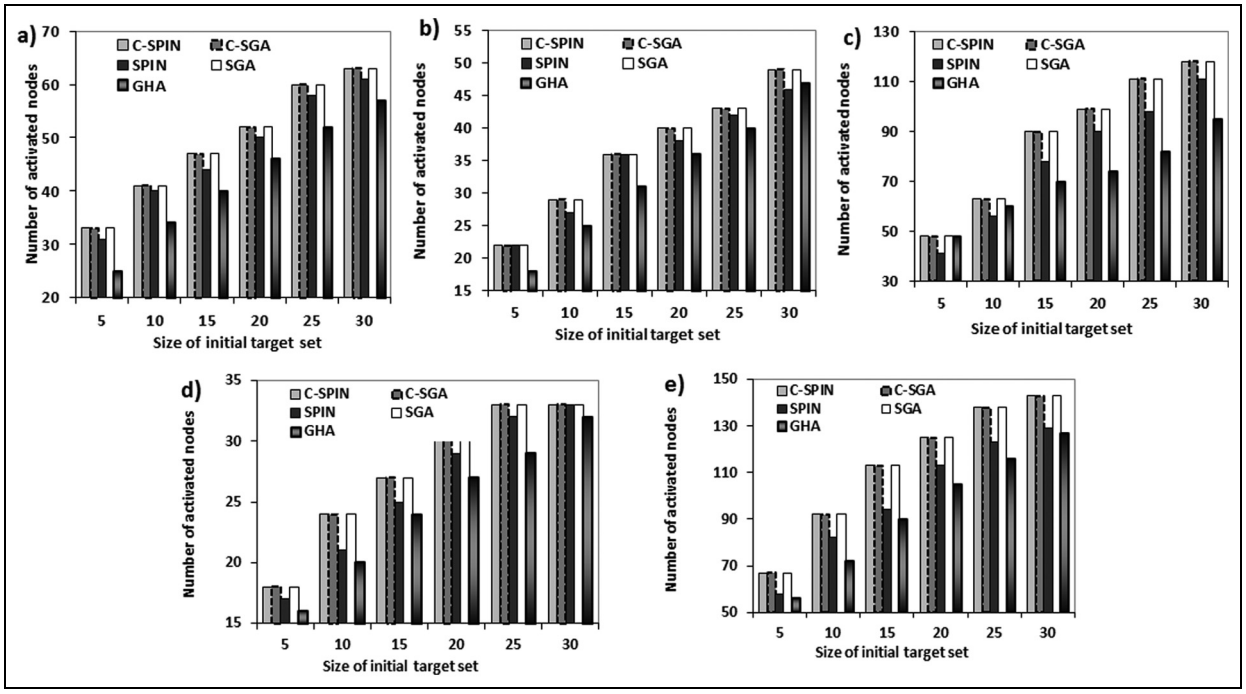

Figure 9. Number of active nodes vs the size of the initial target set under the multiplication threshed model: (a) Adjacency of nouns dataset; (b) Political Books dataset; (c) Celegans dataset; (d) Karate dataset; and (e) Delicious Dataset.

Table 3. Running time comparison of the algorithms on various datasets under the minimum threshold model

Dataset	GHA	SGA	SPIN	C-SGA	C-SPIN
Adjacency of nouns	3.11	10.543	1.074	7.322	0.605
Political Books	1.397	5.218	0.516	2.297	0.189
Celegans	14.10	30.54	3.818	8.865	1.324
Zachary's Karate Club	0.46	0.627	0.087	0.580	0.078
Delicious	100.40	340.54	17.45	32.55	2.169

Table 4. Running time comparison of the algorithms on various datasets under the multiplication threshold model

Dataset	GHA	SGA	SPIN	C-SGA	C-SPIN
Adjacency of nouns	8.11	19.325	2.50	16.655	1.033
Political Books	0.672	2.09	0.185	1.28	0.087
Celegans	30.75	40.10	7.148	13.25	1.98
Zachary's Karate Club	0.477	0.75	0.092	0.645	0.079
Delicious	201.47	610.68	39.385	39.1	5.83

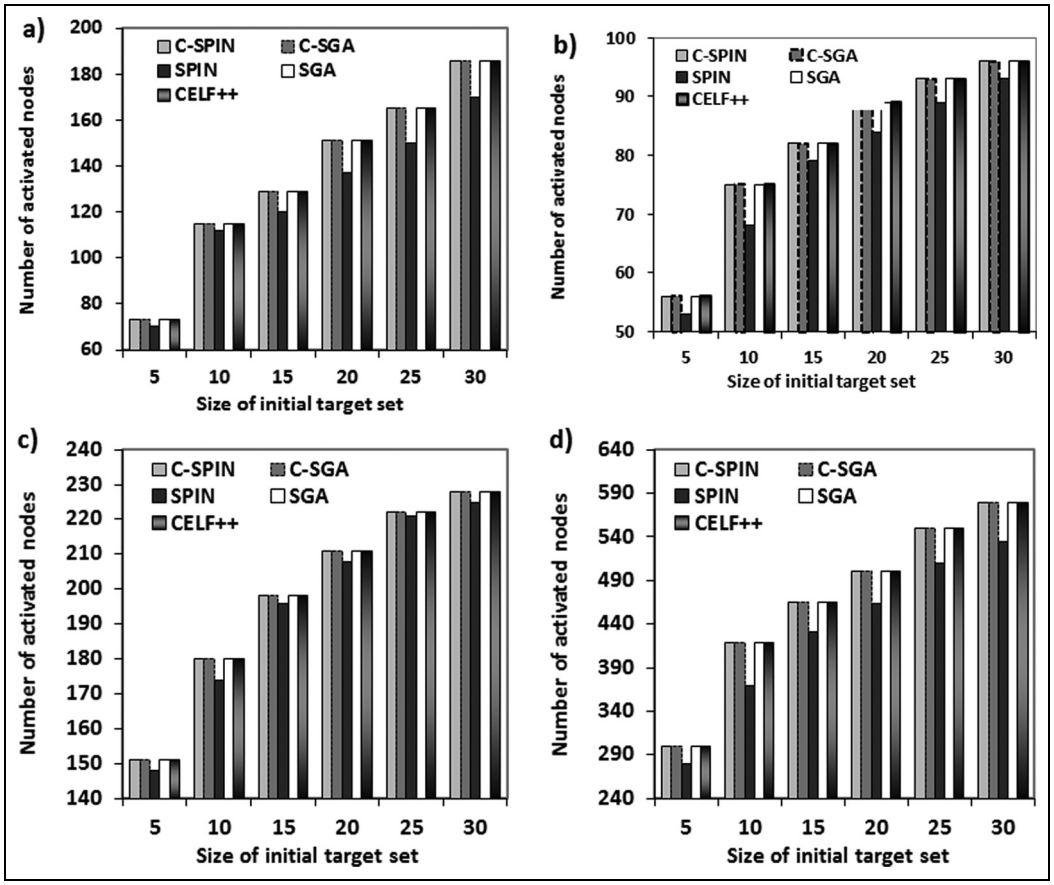

Figure 10. Number of active nodes vs the size of the initial target set under the linear threshed model: (a) Netscience Curatorship dataset; (b) Political Books dataset; (c) Celegans dataset; and (d) Delicious dataset.

Table 5. Running time comparison of the algorithms on various datasets under the linear threshold model

Dataset	CELF++	SGA	SPIN	C-SGA	C-SPIN
Political Books	4.395	39.1	3.55	20.683	1.322
Celegans	35.47	490.45	22.345	130.651	8.764
Netcience	86.14	970.55	40.69	110.211	15.73
Delicious	380.56	3160.13	294.25	340.804	31.15

information diffusion model as a monotone submodular one. Accordingly, the two C-SPIN and C-SGA methods improving the SPIN and SGA are proposed. The improvements are based on considering the community structure of the social network and the newly introduced influence-based closeness centrality measure of the nodes. Computing this measure includes introducing the concept of blocked nodes in the minimum and multiplication threshold models and applying the live-edges view in the linear threshold model. This measure is adopted in order to select the candidate nodes to the extent of the communities of the social network. Moreover, some heuristics are adopted to compute the influence spread in the shorter distance than the diameter of the social network, with respect to the quality of response. Evaluations are conducted first, to evaluate the proposed influence-based closeness centrality measure and second, to compare the proposed community-based methods with well-known benchmarks in the literature with real datasets. The effectiveness of the influence-based closeness centrality is illustrated in comparison with degree and original closeness centralities. Also, the results indicate the efficiency and effectiveness of the proposed methods in maximizing the spread of influence in social networks.

Run time efficiency can be improved by parallelizing the proposed methods. The other future task is to evaluate the impact of the various community detection algorithms on the outcomes.

Funding

This research was supported by University of Isfahan.

References

[1] Domingos P and Richardson M. Mining the network value of customers. In: *Proceedings of the seventh international conference on knowledge discovery and data mining KDD 01*. New York: ACM, 2001, pp. 57–66.

[2] Richardson M and Domingos P. Mining knowledge-sharing sites for viral marketing. In: *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining KDD 02*. New York: ACM, 2002, pp. 61–70.

[3] Kempe D, Kleinberg J and Tardos E. Maximizing the spread of influence through a social network. In: *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining KDD 03*. New York: ACM, 2003, pp. 137–146.

[4] Yu Y, Berger-Wolf T-Y and Saia J. Finding spread blockers in dynamic networks. In: *Advances in social network mining and analysis*. Berlin: Springer, 2010, pp. 55–76.

[5] Kimura M, Saito K and Motoda H. Minimizing the spread of contamination by blocking links in a network. In: *AAAI' 08 Proceedings of the 23rd national conference on artificial intelligence – Volume 2*. Palo Alto, CA: AAAI Press, 2008, pp. 1175–1180.

[6] Song X, Tseng B-L, Lin C-Y and Sun M-T. Personalized recommendation driven by information flow. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: ACM, 2006, pp. 509–516.

[7] Morid M-A, Shajari M and Hashemi A-R, Defending recommender systems by influence analysis. *Information Retrieval* 2004; 17(2): 137–152.

[8] Narayanam R and Narahari Y. A Shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering* 2010; 99: 1–18.

[9] Wang C, Deng L, Zhou G and Jiang M, A global optimization algorithm for target set selection problems. *Information Sciences* 2014; 267: 101–118.

[10] Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J and Glance N. Cost-effective outbreak detection in networks. In: *Proceedings of the 13th ACM SIGKDD conference on knowledge discovery and data mining KDD 07*. New York: ACM, 2007, pp. 420–429.

[11] Chen W, Wang Y and Yang S. Efficient influence maximization in social networks. In: *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining KDD 09*. New York: ACM, 2009, pp. 199–208.

[12] Chen W, Yuan Y and Zhang L. Scalable influence maximization in social networks under the linear threshold model. In: *10th International conference on data mining (ICDM)*. Washington, DC: IEEE Press, 2010, pp. 88–97.
[13] Goyal A, Lu W and Lakshmanan L-V. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In: *11th International conference on data mining (ICDM)*. Washington, DC: IEEE, 2011, pp. 211–220.
[14] Goyal A, Lu W and Lakshmanan LVS. CELF++: Optimizing the greedy algorithm for influence maximization in social networks. In: *Proceedings of the 20th international conference companion on World Wide Web*. New York: ACM, 2011, pp. 47–48.
[15] Yang Y, Chen E, Liu Q, Xiang B, Xu T and Shad S-A, On approximation of real-world influence spread. In: *Proceedings of the 2012 European conference on machine learning and knowledge discovery in databases ECML PKDD'12*. Berlin: Springer, 2012, pp. 548–564.
[16] Liu Q, Xiang B, Zhang L, Chen E, Tan C and Chen J. Linear computation for independent social influence. In: *13th International conference on data mining ICDM*. New York: IEEE, 2013, pp. 468–477.
[17] Kim J, Kim S-k and Yu H. Scalable and parallelizable processing of influence maximization for large-scale social networks. In: *29th International conference on data engineering (ICDE)*. New York: IEEE, 2013, pp. 266–277.
[18] Ohsaka N, Akiba T, Yoshida Y and Kawarabayashi K-I. Fast and accurate influence maximization on large networks with pruned monte-carlo simulations. In: *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*. Palo Alto, CA: AAAI Press, 2014, pp. 138–144.
[19] Zhou C, Zhang P, Guo J and Guo L. An upper bound based greedy algorithm for mining top-k influential nodes in social networks. In: *Proceedings of the companion publication of the 23rd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee*, 2014, pp. 421–422.
[20] Scripps J, Tan P-N and Esfahanian A-H. Exploration of link structure and community-based node roles in network analysis. In: *Seventh IEEE international conference on data mining, ICDM 2007*. New York: IEEE, 2007, pp. 649–654.
[21] Cao T, Wu X, Wang X-S and Hu X. Maximizing influence spread in modular social networks by optimal resource allocation. *Expert Systems with Applications* 2011; 38(10): 13128–13135.
[22] Yan Q, Guo S and Yang D. Influence maximizing and local influenced community detection based on multiple spread model. In: *Advanced Data Mining and Applications*. Berlin: Springer, 2011, pp. 82–95.
[23] Chen Y, Chang S, Chou C, Peng W and Lee S. Exploring community structures for influence maximization in social networks. In: *The 6th workshop on social network mining and analysis held in conjunction with KDD, SNA-KDD*, 2012, pp. 1–6.
[24] Zhang X, Zhu J, Wang Q and Zhao H. Identifying influential nodes in complex networks with community structure. *Knowledge-Based Systems* 2013; 42: 74–84.
[25] Lv J, Guo J and Ren H. A new community-based algorithm for influence maximization in social network. *Journal of Computational Information Systems* 2013; 9(14): 5659–5666.
[26] Cong G, Zhou X, Wang Y and Xie K. Influence maximization on large-scale mobile social network: A divide-and-conquer method. *IEEE Transactions on Parallel and Distributed Systems* 2014; 26(5): 1379–1392.
[27] Chen Y-C, Zhu W-Y, Peng W-C, Lee W-C and Lee S-Y. CIM: Community-based influence maximization in social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2014; 5(2): article no. 25.
[28] Rahimkhani K, Aleahmad A, Rahgozar M and Moeini A. A fast algorithm for finding most influential people based on the linear threshold model. *Expert Systems with Applications* 2015; 42(3): 1353–1361.
[29] Landherr A, Friedl B and Heidemann J. A critical review of centrality measures in social networks. *Business and Information Systems Engineering* 2010; 2(6): 371–385.
[30] Hosseini-Pozveh M, Zamanifar K, Naghshnilchi A and Dolog P, Maximizing the spread of positive influence in signed social networks. *Intelligent Data Analysis* 2016; 20(1).
[31] Jin Y, A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing* 2005; 9(1): 3–12.
[32] Xie J and Szymanski B-K. Labelrank. A stabilized label propagation algorithm for community detection in networks. In: *Proceedings of the IEEE network science workshop*, WestPoint, NY. New York: IEEE, 2013, pp. 138–143.
[33] Xie J, Chen M and Szymanski B-K. LabelrankT: Incremental community detection in dynamic networks via label propagation. In: *Proceedings of the workshop on dynamic networks management and mining*. New York: ACM, 2013, pp. 25–32.