Journal of Physics: Conference Series

Extracting Influential Nodes for Maximization Influence in Social Networks

ECS
The Electrochemical Society
Advancing solid state & electrochemical science & technology
240th ECS Meeting
ORLANDO, FL
Orange County Convention Center
Oct 10-14, 2021
Abstract submission due: April 9
SUBMIT NOW

Extracting Influential Nodes for Maximization Influence in Social Networks

Zainab Naseem Attuah[1]*, Firas Sabar Miften[1], Evan Abdulkareem Huzan[1]

[1]University of Thi-Qar, College of Education for Pure Science, Iraq.

Emails: *zainabnaseem88@gmail.com; firas@utq.edu.iq; evan@utq.edu.iq

Abstract. Influence maximization (IM) is the process focuses on finding active users who make that maximizes the spread of influence into the network. In recent years, community detection has attracted intensive interest especially in the implementation of clustering algorithms in complex networks for community discovery. In this paper the social network was divided into communities using the proposed algorithm which is called (CDBNN) algorithm, CDBNN stands for Community Discovery Based on Nodes Neighbor. The seed nodes(candidate nodes) were extracted using the degree centrality in each community. The propagates model (PSI) was used to information propagates through the network. Finally, using closeness centrality to extract the influential nodes from the network. Experimental results on the real network are efficient for influence propagates, compared with two known proposals.

1. Introduction

The graph is the most significant data structures, and the important models that are highly effective for the purpose of representing social networks in a form G (V,E) where (V) represent (the group of vertex) and (E)) represent the group of edges (links)[1]. The very important thing in social networks is influence, as it is necessary for network analysis applications used for marketing and business, as it is not only from the side of information flow[2]. The relationship that exists between two entities in a particular work is called social influence, where the first entity is called the influencer while the second entity is called the influence, so the first entity affects the second entity[3]. Several methods have been proposed during the recent period, but they are more complex, as they generally include two types. The first is concerned with the structure of the network only. The second is concerned with the social information of the users in addition to the network structure such as interests and trust[4] .The Influence works to clarify and describe the problem of how to find small sub-nodes in the social network that have the potential to increase the spread of the effect, as many algorithms have been proposed to find solutions to this problem[5]. The most important characteristic of social networks is the high possibility of disseminating information at high speed among large groups of individuals and clarifying their opinions[6]. There are two basic models for disseminating an idea across the network, the LT model and the IC model [7].In this research, the focus was on disseminating information in a specific social network, as well as the influencing nodes were extracted, and the research was arranged as follows: retrieval of literature in the second section, introductory definitions in the third section, description of the proposed approach to amplify the spread of influence in the fifth section, providing details of the analysis The experiments applied to a real data set are in the sixth section and finally the conclusion of the research.

2. Literature Review

Influence nodes are the nodes that have the greatest impact on proliferation in complex networks. The process of identifying the nodes with the greatest ability to influence and arranging them according to the amount of their influence on the spread is of great importance .In the field of social networks, there are many research problems in addition to the problem of amplifying the impact, including finding influential nodes and discovering the community .

Kanna Gharib Al-Falahi (2014) in order to find out which nodes have a strong influence in the network it is very important to discover the community, this is based on its installation site in order to start its impact campaigns[8] .

Aftab Farooq et al (2018) depending on the network metrics, including the clustering coefficient, the degree centrality, and Betweenness, Page Rank centrality, Closeness Centrality they proposed a scheme to know the most influential nodes so that these features can be used for the purpose of predicting social networks and thus enables them to know and identify the nodes affecting the networks [9].

Kaiqi Zhang et al (2017) in this research, the LT model was adopted as a propagation model, and the genetic algorithm was the proposed one where they determined the effect of seed group propagation on the basis of a fitness function, and optimal solutions were achieved by development and competition. Also tried to update some research methods in SA in order to be more efficient and more accurate[10].

Muluneh Mekonnen Tulu et al (2018) during recent years, the process of obtaining a powerful leader of the community to play the role of rapidly spreading information through the complex network is a very tiring and worrying issue. In this research, a (community mediator) CbM has been proposed for the purpose of identifying the nodes affecting large and complex networks. They proposed the entropy of a random walk for all nodes to each community. The CbM is in the process of explaining how to link two or more communities by the node in the network [11].

3. Preliminaries

Directed graph("A graph G = (V, E) where V is the set of nodes or actors (say “n”) and E is the set of edges or connections (say “m”) is directed)", if E is a set of ordered pairs meaning that $[v_1 , v_2] \neq [v_2 , v_1]$ where $[v_1 , v_2] \in E$ and $v_1, v_2 \in V$ [12].

Undirected Edges ("Here the edges do not have any particular direction from one vertex to another; there is no difference between the two vertices connected via one undirected edge").A straight line can be represented[13].

Degree Centrality ("It can be defined as the most important and simpler method used to find the most effective nodes in any network. For node i, It was found that the influence is directly affected by its degree, that is called degree centrality")[14].

Closeness centrality ("Closeness centrality indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network ")[15].

Diffusion model ("Also known as propagation model, describes the whole diffusion process and determines how the influence propagates through the network") [16].

4. Methodology

The proposed method (IMCS) stands for influence maximization based on centrality measures and structural aspect. It consisting of three parts: communities discovery, finding seed nodes and finding of influential nodes. The proposed method is presented in Figure 1. In the first part, the social network was divided into communities using the Community Discovery Algorithm Based on Nodes Neighbor. The set of nodes was extracted that playing the role of seed (Candidate nodes)in the second part. After that, the influential nodes are extracted from the active nodes.

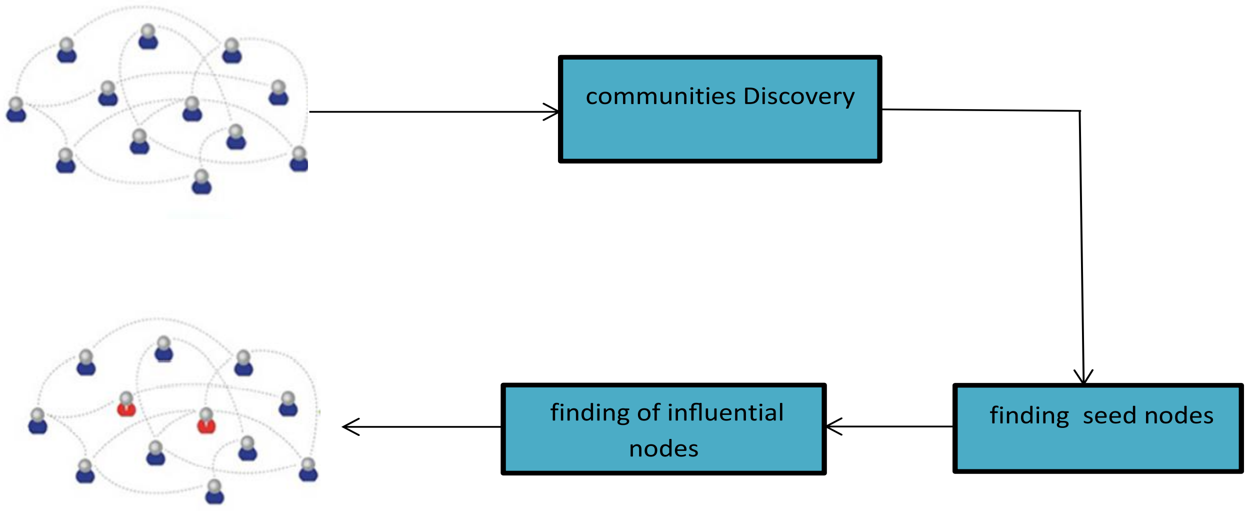

Figure1: Block diagram of proposed method for Influence maximization.

4.1 Communities Discovery

Which step communities discovery , the proposed algorithm is Community Discovery Algorithm Based on Nodes Neighbor(CDBNN) works on clustering data based on relation among nodes(neighbors) .The distance matrix(D) calculates distance between nodes according of the proposed algorithm(1),Where D(i, j) represents the distance between node i and node j of G = (V,E) is the v × v, where v is a number of nodes in dataset . To illustrate this algorithm, input for the algorithm is an adjacency matrix with v node. In step (2) all nodes enter into (for)looping ,to finding neighbors nodes, and for every node of neighbors finding their neighbors as well, in step (3). After that, calculating the intersection and union of nodes resulting from the previous two steps.
Then computing the distance for each node by calculating the product of the division of the intersection of the nodes by their union.

Proposed algorithm (1) compute distance

```
Input: A(v, v)  \\ adjacency matrix with v node.
Output :  D(v, v) \\ Distance matrix.
1-Begin
2-        For all nodes i  ∈ (1,…, v) do
3-          Ni= find all neighbors of node i
 4-         For all j : j ∈ Ni
                Nj = find all neighbors of node j
5  -            Ci = intersect between Ni and Nj .

6-              Cu=union between Ni and Nj
7-              Compute D( i,j) =size of ci / size of cu
8-        End for
18-   End for
19-End algorithm
```

The proposed algorithm(CDBNN) can be used in large data set by giving it just any data set, it returns a number of communities and labels for each node. To illustrate this algorithm, input for the algorithm is a distance matrix between node i and node j. In short, the nodes are arranged according to the highest neighbors, and then the node is tested if it does not belong to a community, it is placed in the first community and its neighbors are extracted. The distance between the node i and one of its neighbors(node j) is tested if it is greater than the threshold and does not belong to a community in advance, it is placed within the community of the node i, but if it is greater than the threshold and belongs In advance to a community and the distance between the node i and the node j is the largest among its neighbors, it remains in the same community, but if the condition is not met, this node remains outside the node i community .

Proposed algorithm (2) Communities discovery

```
Input: D(v, v):distance of v node
Output : C : Communities ,L :label
Begin
  C=0,  L(1,….,n)=0 \\ Label is vector content cluster of node.
  Sn = sort all nodes according to height neighbors.
  For each node i in Sn
      If L(i)=0      \\ the node is not assign to any community.
          C=C+1      \\ increment community number
          L(i)=C       \\  assign node  i to community C.
      End if
      Th=(min(D(i,:))+max(D(i,:)))/2*0.16;  \\ calculate adaptive     threshold.
      Ni = find all neighbors of node i
      For each j ∈ Ni
          If D(i,j)≥ th
              If L(j)=0
                  L(j)= L(i)       \\ assign node j to same community of node i.
          Else if
                  D(i,j) is max distance for all neighbors of node j
                  L(j)= L(i)
              End if
          End if
      End if
      End for
  End for
  End algorithm
```

4.2 Finding seed nodes

After dividing the network into communities using the proposed algorithm Community Discovery Algorithm Based on Nodes Neighbor. The seed nodes(Candidate nodes) were selected from each community. A node is considered ,key if it is degree centrality is greater or equal then its neighbors. These nodes are considered the candidate nodes for spreading influence in the social network Recall that a node's degree is simply a count of how many social connections (i.e., edges) it has. To calculate the degree centrality of a node v of a given graph G =(V, E) used the following equation:

$$\mathbf{DC(v)} = \frac{degree(v)}{|V|} \qquad \mathbf{(1)}$$

Where degree(v) is the number of connections (edges) a node (vertex) has in the network and (|V|) is the nodes number of G .The (Algorithm 1) describe the part of key nodes.
The second part of this section is the process of spreading the impact by applying the influence propagation model to the seed nodes obtained from each community. Linear Threshold (LT) and the Independent Cascade (IC) are models the two popularly used models. These models are used to spread influence in the social network. It identifies inactive nodes that can turn into active nodes and vice versa. The semantic insert in a graph G through connected among nodes by link with the weight .their information which is given in the following equation defined by the semantic similarity of their information.

$$\mathbf{Sem(u ,v)} = \frac{\boldsymbol{common}}{\boldsymbol{leng(u)+leng(v)}} \quad \mathbf{(2)}$$

Where (u,v) represent the common attributes among two nodes (u) and (v) . leng (u) is the length of the attributes vector of a node (u) and leng (v) is the length of the attributes vector of a node (v) [17]. The model used in the impact deployment process in the suggested method is probabilistic social influence model. The model of PSI , can adopt following model, in a part p=(c1,c2,….cr) the node v is related with a threshold value θc(v). Threshold values are the generated according to the degree of each node between 0.3 and 0.9. If the summation of the aggregate weight of its active neighbors , exceeds the value θc(v) is a node (v) becomes active. The equation used to calculate diffusion model as follows::

$$\sum_{\boldsymbol{u\in Av}} \boldsymbol{w(u,v)} > \boldsymbol{\theta c(v)} \quad \mathbf{(3)}$$

Where, θc(v) is represents the threshold value. Av is represents the group of active neighbors of (v) and w(v, u) is represents an activation probability that show the summation of similarity among two nodes (v) and (u) active. This process is repeated until there are no inactive nodes that can be activated[18].

Algorithm (3): Generate the seed nodes

```
input: social network G (V,E) , V set of nodes, E set of edges, a community C;
output: A SN set of seed nodes;
    1- KN ← Ø;
    2- Begin
    3-   Degree Centrality is computed of each node of the graph by
         equation 1
    4 -    for each v ∈ V of a community C do
    5 -        if is seed (v) then
    6 -            SN ←S N ∪ {v};
    7 –        End if
    8-    End for
    9-End algorithm
```

4.3 Finding of influential nodes

In this part, the influential nodes are extracted from the active nodes in each community, by calculating the degree of influence for each community , where it is equal to the result of dividing the active nodes number in each community (N active) by all number of nodes in the graph G (V) .

$$R(Cr)=Nactive / N \qquad (4)$$

Through the highest degree of influence for each community, the influencing nodes are determined. Closeness centrality is to define each active nodes. Used the flowing equation to calculate the closeness centrality.

$$\mathbf{CC(vi)}=\frac{\mathbf{1}}{\sum_{\boldsymbol{vj}\in\boldsymbol{V}}|\boldsymbol{shortpath}(\boldsymbol{vi},\boldsymbol{vj})|} \qquad \mathbf{(\ 5)}$$

The shortest paths between two nodes represents (ShortPath(vi, vj)) where vi within to the group of seed nodes and vj within to the group of active nodes. The closeness centrality are calculated of each seed node, These following nodes are ranked in ascending order. The node is the influential node if it is the greatest closeness centrality .The Algorithm 3 describes the part of detecting Influential nodes. Inspired by CGA[19].

Algorithm (4): Finding Influential Nodes

```
Input: social network G = (V,E), NActive;
NActive set of active nodes
Output: A set InfN of influential nodes;
InfN ← ∅,SN← ∅, NActive ← ∅;
Begin
    InfN ← ∅;
    C ← Discovery community (Algorithm 1)
    numC = |C|        // number of communities
    while numC ≠ ∅ do
        the degree of influence of all communities was computed by
        equation 5. Mining about influence nodes inside community
        has the value of the highest degree of influence. COmax
        represents the community has the value of the highest degree of influence;
        ACOmax ← { Active nodes in COmax };
    while (COmax = ∅)et (ACOmax = ∅) do
            using the equation (5 )compute the closeness centrality of
            each node
        assort the nodes in ascending order of
        closeness centrality;
        vmax ← the node having the maximum closeness centrality;
        InfN ← InfN ∪ {vmax};
```

```
14 –end while
15- end while
16 - return InfN
17- End algorithm
```

Algorithm proposed (5): Influence Maximization Algorithm(IMCS)

```
input: social network G = (V,E)
, a set of key nodes KN , a set of active nodes NActive;
output: A set InfN of influential nodes;
1- InfN ← ∅;
2-K N ← ∅;
3- NA ← ∅;
4- Begin
5-     Communities Detection C of graph(Algorithm 2);
6-     for each community C do
7-         Generate seed nodes(C, G);( Algorithm 3);
8 -        Apply the propagates model to determine the inactive nodes
           that can be activated . by apply the equation 2 and the
           equation 3
9-     end for
10-    Finding influential nodes (NActive , G); // (Algorithm 4);
11-    return Ninf ;
12- End algorithm
```

The proposed method is described in algorithm5, it first to detect the communities that implement algorithm 1 (line 5). After this, the algorithm does a very important job, which is to create a group of nodes that take the role of seed nodes, as a result, candidate nodes will be formed for the purpose of propagation a new idea (Algorithm 3). This diffusion model is applied for the purpose of identifying active nodes in most parts of the network (line 7). Finally, all the influential nodes can identify (Algorithm 4).

5 . Experimentation

The proposed algorithm assesses its efficiency practically on real networks, by linking each node with sites of great importance, all applications are executed on Matlab and run on an Intel Core i5-3360M processor, and the CPU was 2.80 GHz and 4 GB memory. In the experiments, the algorithm IMCS compare with present algorithms that play the role of influence maximization: C-SGA where method considering influence-based closeness centrality measure of the nodes are presented to maximize the spread of influence and the community structure of the social networks[20]. Coreness centrality is used in determining the ability of the node to spread through Coreness the neighbors of each node, the basic idea on which this method relies is that it is a strong diffusion that has many connections to the central nodes in the network [21].

Influence-based closeness centrality (ICC) measure it is a highly adopted measure in the influence maximization domain. the number of final activated nodes gained through them, is computed and through this method ,the k most influential nodes are identified .The results of final activated nodes is computed on the multiplication threshold model and minimum threshold model[20].The influence propagation is used metric to evaluate the performance of IMCS .

5.1 Experimental Results

The datasets used for evaluate the performance this proposed scheme of the different algorithms such as Political Books, Zachary's Karate Club and Adjacency of nouns. A summary of all the datasets in Table(1)

Table (1): The Real Social Networks in the experiments

Data set	Number of Nodes	Number of edges
Zachary's Karate Club	34	78
Political Books	105	441
Adjacency of nouns	112	425

Varying the influence spread with the size of the network. The algorithm proposed (IMCS) covers a large number of activated nodes. According to the results, when we increase the network size, IMCS is still able to have a better value of influence spread. In figure(4.7) Varying the influence spread with the size of the network.

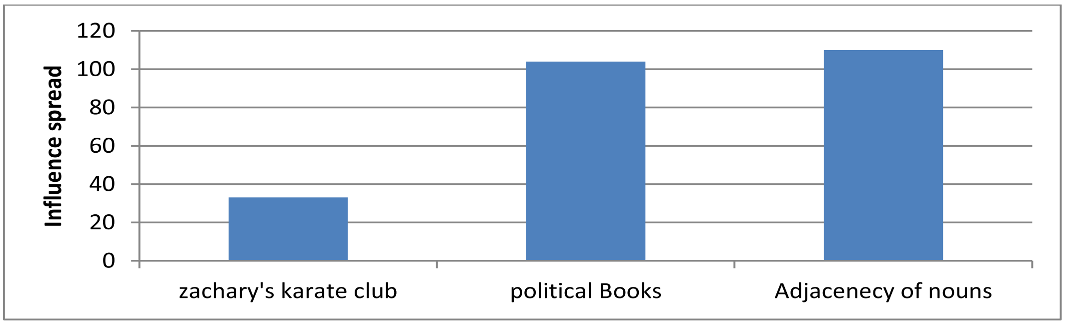

Figure 2. The results of the proposed algorithm (IMCS) experiments for influence spread on social networks

IMCS maintains influence propagation values for three networks ,Where this values close to cc . The influence propagation of IMCS is always better than the methods C-SGA , CC and ICC. The approach covers a large number of activated nodes, With a small number of initially active nodes. According to the results ,the Influence Spread change with the size of network , observe in figure 3.

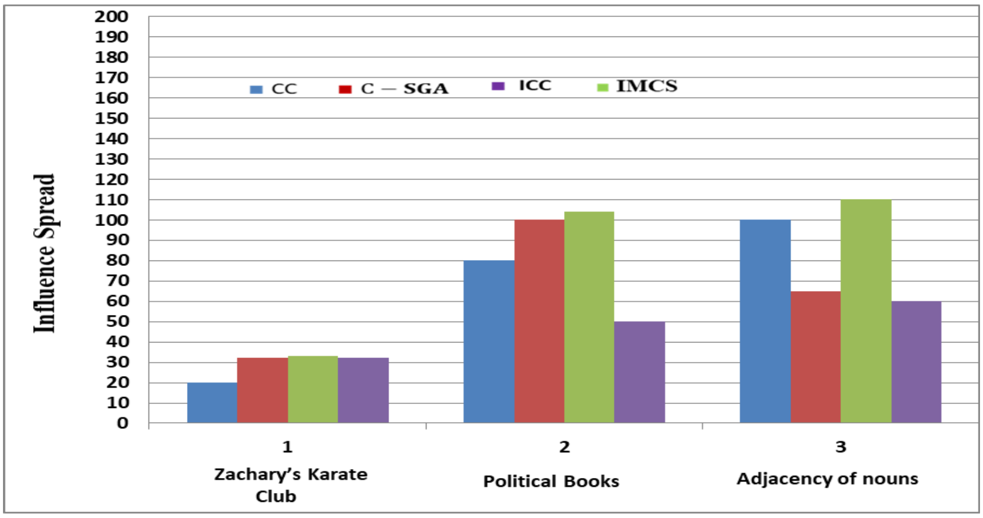

Figure 3. change the Influence Spread with the size of dataset

6. Conclusion

To get the more effective spread of information and Influence Maximization in Social Networks is of vital importance .The main goal is define a set of influential users to maximize their influence to other users in a social network and propagate the influence with a higher quality of seed . Many approaches based on centrality measures such as degree centrality and closeness centrality are proposed for this purpose. In closeness centrality is computed for each node the sum of the shortest path to all other nodes .In this paper, to find influential nodes in information networks, at first communities are detected and by applying degree centrality , key node from each community is extracted and apply propagation model(PSI). This model is applied based on the semantic similarity among nodes to identify the active nodes in the network. Finally, the nodes are extracted that the highest closeness centrality. These nodes are the influential node .In this paper ,the proposed an approach called(IMCS) test with several other influence maximization methods on both real-world networks for comparison. The experimental results explain the efficiency and effectiveness of proposed method.

References

[1] Koutrouli, M., Karatzas, E., Paez-Espino, D., and Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, ***8*****, 34.**

[2] Carolina, D .(2012) .Finding Influencers in Social Networks . *Institute Superior Tecnico*.**45**

[3] Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., and Jia, W. (2018). Influence analysis in social networks: A survey. *Journal of Network and Computer Applications*, *106*, **17-32.**

[4] Zareie, A., Sheikhahmadi, A.,and Jalili, M. (2019). Identification of influential users in social networks based on users' interest. *Information Sciences*, **493**, 217-231.

[5] Wang, F., Zhu, Z., Liu, P., and Wang, P. (2019). Influence Maximization in Social Network Considering Memory Effect and Social Reinforcement Effect. *Future Internet*, **11**(4), 95.

[6] Wang, Q., Jin, Y., Cheng, S., and Yang, T. (2017). ConformRank: A conformity-based rank for finding top-k influential users. *Physica A: Statistical Mechanics and its Applications*, **474**, 39-48.

[7] KESKİN, M. E., and Güler, M. G. (2018). Influence maximization in social networks: an integer programming approach. *Turkish Journal of Electrical Engineering & Computer Sciences*, **26**(6), 3383-3396.

[8] Al-Falahi, K. G.(2014). Community Detection and Influence Maximization in Online Social Networks .58 (2014) :100-115 .

[9] Farooq, A., Joyia, G. J., Uzair, M., and Akram, U. (2018, March). Detection of influential nodes using social networks analysis based on network metrics. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-6). IEEE.

[10] Zhang, K., Du, H., and Feldman, M. W. (2017). Maximizing influence in a social network: Improved results using a genetic algorithm. *Physica A: Statistical Mechanics and its Applications*, **478**, 20-30.

[11] Tulu, M. M., Hou, R., and Younas, T. (2018). Identifying influential nodes based on community structure to speed up the dissemination of information in complex network. *IEEE Access*, **6**, 7390-7401.

[12] Zhan, J., Gurung, S., and Parsa, S. P. K. (2017). Identification of top-k nodes in large networks using katz centrality. *Journal of Big Data*, ***4*(1),** 1-19.

[13] Chakraborty, A., Dutta, T., Mondal, S., and Nath, A. (2018). Application of graph theory in social media. *International Journal of Computer Sciences and Engineering*, **6**, 722-729.

[14] Mao, C., and Xiao, W. (2018). A comprehensive algorithm for evaluating node influences in social networks based on preference analysis and random walk. *Complexity*, *2018*.

[15] Golbeck, Jennifer. (2013).*Chapter 3—Network Structure and Measures. Analyzing the Social Web*. 25-44.

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

[17] Hafiene, N., and Karoui, W. (2017, October). A new structural and semantic approach for identifying influential nodes in social networks. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1338-1345). IEEE.

[18] Doo, M., and Liu, L. (2014). Probabilistic diffusion of social influence with incentives. *IEEE Transactions on Services Computing*, ***7*(3),** 387-400.

[19] Song, G., Zhou, X., Wang, Y., and Xie, K. (2014). Influence maximization on large-scale mobile social network: a divide-and-conquer method. *IEEE Transactions on Parallel and Distributed Systems*, **26(5)**, 1379-1392.

[20] 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)**, 204-220.

[21] Bae, J., and Kim, S. (2014). Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A: Statistical Mechanics and its Applications*, 395, 549-559.