DOI: http://doi.org/10.32792/utq.jceps.10.02.019

Extracting Influential Nodes in Social Networks Based On Community Structure

Zainab Naseem[1] ,Firas Sabar Miften[1] , Evan Abdulkareem Huzam[3]

[1,2,3] University of Thi-Qar, College of Education for Pure Science, Iraq

Received 27/09/2020 Accepted 16/11/2020 Published 30/11/2020

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 detection. In this paper the social network was divided into communities using the proposed algorithm which is called (CDBNN) algorithm, CDBNN stands for Community Detection Based on Nodes Neighbor. The key 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 three known proposals.

Keywords: - social networks, community detection, influential nodes, influence propagation, influence maximization.

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. Within the scope of Big Data, the process of community mapping complex networks has taken great interest. It has become both practical and theoretical importance for the purpose of analyzing the topology as well as analyzing tasks and evaluating the behavior of complex networks.

Di Jin et al (2011):

depending on the modification and improvement of the model units, Lovain's algorithm was proposed, and its most important purpose was to update the entire network model by constantly dividing the community. Within groups there is high interconnection, which is rare is external communications only [8].

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 [9].

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 [10].

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 [11].

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 [12].

Arastoo Bozorgi et al (2017):

an attempt has been made to find solutions to the problem of maximization of the competitive influence .by finding a propagation model that can give a high ability to control the nodes regarding the spread of the effect and represent the extension of the linear threshold. In addition, an algorithm has been proposed with a high potential in identifying the influence nodes through a graph. Social through which it is possible to calculate the spread of each node locally within its community, as well as benefit from the proposed propagation model that takes advantage of the community structure in the first place [13].

Bo Zhang et al (2019):

here, they found a new method for identifying effective nodes is a trust-based most influential node discovery (TMID) method for discovering influential nodes in a social network. This was done through four important stages to discover the effective nodes and determine the degree of their influence, as follows:1- The process of spreading influence. 2- a trust evaluation method. 3- The stage in which the effect is evaluated, which calculates the clear mutual effect between users, which is called (active influence), and the potential mutual effect between users, which is called (inactive effect), as well as the single effect that is called (node effect). 4- Discovering the most influential nodes using a set of algorithms, from two stages, the first phase, the greed phase, and the second, the heuristic phase [14]

Yunming Wang et al (2019):

method has been proposed that depends mainly on the integral k-shell, the benefit of which is to identify the influential nodes in the control and command networks. In this modern method, it takes into account all local and global information for all nodes, presenting the k-shell history and the amount of juxtaposition in a binary order and correcting the analysis process k-shell in the network [15].

Maria-Evgenia G et al (2018):

here in this paper, Matrix Influence (MATI) has been proposed, which is a highly efficient algorithm that can be used in two models, the first is the linear propagation model and the second is the independent Cascade model, as (MATI) works based on making use of simple paths in the areas close to the node for Impact calculation in advance. On the same methodology of the greedy algorithm in all rounds (MATI), the node with the largest marginal influence is selected consecutively and individually in the network, as they can calculate the marginal gain with high efficiency and choose a candidate node [16].

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 $v_1, v_2 \in V$ [17].

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 [18].

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(” [19].

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 (" [20].

Diffusion model ("Also known as propagation model, describes the whole diffusion process and determines how the influence propagates through the network") [21].

Euclidean distance (" is the most popular distance measure, Let i=(x_{i1} ,x_{i2}, …,x_{ip}), j= (x_{j1},x_{j2},…,x_{jp}) be two objects described by p numeric attributes. The Euclidean distance between objects i and j ") [22] is defined as

$$d(i,j)=\sqrt{\sum_{p=1}^{n}(\ xip-xjp)\ 2} \qquad (1)$$

4. Methods:

The proposed method (IMCS) stands for influence maximization based on centrality measures and structural aspect. It consisting of three parts: community's detection, finding key 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 Detection Algorithm Based on Nodes Neighbor. The set of nodes was extracted that playing the role of keys (Candidate nodes) in the second part. After that, the influential nodes are extracted from the active nodes.

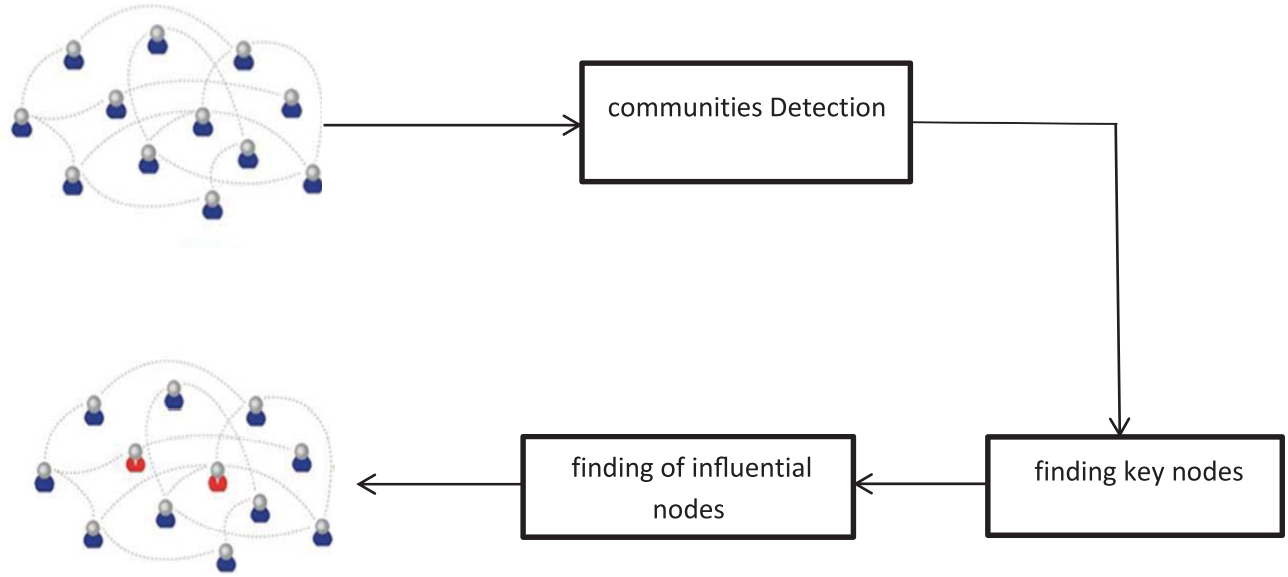

Figure1: Block diagram of proposed method for Influence maximization

4.1 Communities detection:

Which step communities detection , the proposed algorithm is community detection based on nodes neighbor(CDBNN) works on clustering data based on relation among nodes .The distance matrix(D) calculates distance between nodes According of equation 1,Where Dij 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 .This algorithm 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.

Proposed algorithm(CDBNN) (1):

```
Input: A network G = (V, E).
 V=set of nodes
 E= set of edge
Output:
Label \\ contains the label for every nodes
Numc \\ the number of communities.
Begin
  Compute D (v * v) is distance matrix among nodes by using the
    Euclidean distance    equation 1.
  Labels=zeros(v)
  Numc=0;
  for i=1 to v
      if Label(i)==0 then
          Numc=Numc+1;
          Label(i)=Numc;
  End if
  Thr= (min (D (i, :))+max (D (i, :)))/2*0.16; \\calculate adaptive
        threshold.
  for j=i+1:m
      if D (i, j) <= Thr then
        if Label(j)==0
            Label(j)=Label(i);
  End if
  End if
  End for
  End for
End algorithm
```

4.2 Finding Key nodes:

After dividing the network into communities using the proposed algorithm Community Detection Algorithm Based on Nodes Neighbor. The key 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:

$$\mathbf{DC(v)} = \frac{degree(v)}{|V|} \qquad \mathbf{(2)}$$

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 key 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.

$$\text{Sem}(u, v) = \frac{common}{leng(u)+leng(v)} \quad (3)$$

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) [23]. 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:

$$\sum_{u\in Av} w(u,v) > \theta c(v) \quad (4)$$

Where, θc(v) is represents the threshold value. Av is representing 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 [24].

Algorithm (2): Generate the key nodes:

```
input: social network G (V, E), V set of nodes, E set of edges, a community C;
output: A KN set of key nodes;
    1- KN ← ∅;
    2- Begin
    3-  Degree Centrality is computed of each node of the graph by
        equation 2
    4 -   for each v ∈ V of a community C do
    5 -       if is key (v) then
    6 -           KN ←K 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 (5)

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

$$CC(vi)=\frac{1}{\sum_{vj\in V}|shortpath(vi,vj)|} \quad (6)$$

The shortest paths between two nodes represents (ShortPath (vi, vj)) where vi within to the group of key nodes and vj within to the group of active nodes. The closeness centrality is calculated of each key 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 [25].

Algorithm (3): Finding Influential Nodes:

```
Input: social network G = (V, E), NActive;
NActive set of active nodes
Output: A set InfN of influential nodes;
InfN ← ∅, KN← ∅, NActive ← ∅;
Begin
    InfN ← ∅;
    C ← detect 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 (6) 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};
        end while
    end while
    return InfN
End algorithm
```

Algorithm proposed (4): 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;
  InfN ← ∅;
  K N ← ∅;
  NA ← ∅;
  Begin
      Communities Detection C of graph (Algorithm 1);
      for each community C do
          Generate key nodes (C, G) ;( Algorithm 2);
```

```
8 -     Apply the propagates model to determine the inactive nodes
        that can be activated. by apply the equation 3 and the
        equation 4
9-    end for
10-   Finding influential nodes (NActive, G); // (Algorithm 3);
11-   return Ninf;
12- End algorithm
```

The proposed method is described in algorithm4, 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 keys nodes, as a result, candidate nodes will be formed for the purpose of propagation a new idea (Algorithm 2). 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 3).

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 [26]. 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 [27].

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 are computed on the multiplication threshold model and minimum threshold model [26].

the influence propagation is used metric to evaluate the performance of IMCS. Also evaluate the effect of nodes number can be influenced by parameter set of K influential nodes.

A large number of influential nodes are formed without depending on the increment of k by IMCS.

5.1 Experimental Results:

The datasets used for evaluate the performance this proposed scheme of the different algorithms such as Political Books, Celegans, Netscience co-authorships, 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
Celegans	297	2359
Netscience	1589	2742

The number of initially active nodes are increase with K (leader nodes) from 5 to 30 for the three methods on the Netscience dataset which contains 1589 nodes. in this experiment, the influence propagation changes with the increase of K.

According to the Figure 2, and by varying the number of initially active nodes we observe that the influence propagation of C-SGA is still very far from those of IMCS and CC. While the results of CC are close to IMCS. As for IMCS its results independent of the increment of K. The influence propagation is constant with every change for K.

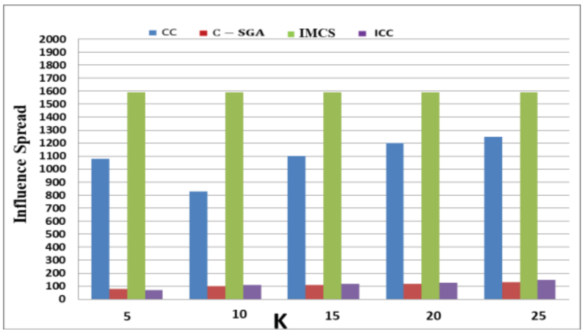

Figure 2. change the Influence Spread with K

According to the results, the Influence Spread change with the size of network, observe in Figure 3.

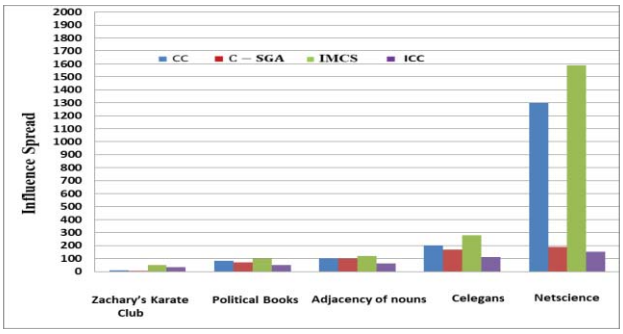

Figure 3. change the Influence Spread with the size of dataset

IMCS maintains influence propagation values for the first three networks, where this values close to cc. when increase the network size like Netscience 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. The following table (2) shows the results of some social networks after extracting the leader nodes and applying the diffusion model PSI.

Table 2. Experimental results of the proposed method.

Dataset	Number nodes	Leader nodes(candidate)	Influence spread
Zachary's Karate Club	34	2	33
Political books network	105	2	104
Adjacency of nouns	112	3	111

6. Conclusion:

To get the more effective spread of information and Influence Maximization in Social Networks is of vital importance. The main goal is defining 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, Mikaela, et al. "A Guide to Conquer the Biological Network Era Using Graph Theory." Frontiers in Bioengineering and Biotechnology 8 (2020): 34.

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

[3] Peng, Sancheng, et al. "Influence analysis in social networks: A survey." *Journal of Network and Computer Applications* 106 (2018): 17-32.

[4] Zareie, Ahmad, Amir Sheikhahmadi, and Mahdi Jalili. "Identification of influential users in social networks based on users' interest." Information Sciences 493 (2019): 217-231.

[5] Wang, Fei, et al. "Influence Maximization in Social Network Considering Memory Effect and Social Reinforcement Effect." Future Internet 11.4 (2019): 95.

[6] Wang, Qiyao, et al. "ConformRank: A conformity-based rank for finding top-k influential users." Physica A: Statistical Mechanics and its Applications 474 (2017): 39-48.

[7] KESKİN, MUHAMMED EMRE, and Mehmet Güray Güler. "Influence maximization in social networks: an integer programming approach." Turkish Journal of Electrical Engineering & Computer Sciences 26.6 (2018): 3383-3396.

[8] Jin, Di, et al. "Fast complex network clustering algorithm using local detection." Dianzi Xuebao (Acta Electronica Sinica) 39.11 (2011): 2540-2546.

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

[10] Farooq, Aftab, et al. "Detection of influential nodes using social networks analysis based on network metrics." International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, (2018).

[11] Zhang, Kaiqi, Haifeng Du, and Marcus W. Feldman. "Maximizing influence in a social network: Improved results using a genetic algorithm." Physica A: Statistical Mechanics and its Applications 478 (2017): 20-30.

[12] Tulu, Muluneh Mekonnen, Ronghui Hou, and Talha Younas. "Identifying influential nodes based on community structure to speed up the dissemination of information in complex network." IEEE Access 6 (2018): 7390-7401.

[13] Bozorgi, Arastoo, et al. "Community-based influence maximization in social networks under a competitive linear threshold model." Knowledge-Based Systems 134 (2017): 149-158.

[14] Zhang, Bo, et al. "A most influential node group discovery method for influence maximization in social networks: a trust-based perspective." Data & Knowledge Engineering 121 (2019): 71-87.

[15] Wang, Yunming, et al. "Influential Node Identification in Command and Control Networks Based on Integral k-Shell." Wireless Communications and Mobile Computing 2019 (2019).

[16] Rossi, Maria-Evgenia G., et al. "MATI: An efficient algorithm for influence maximization in social networks." PloS one 13.11 (2018): e0206318.

[17] Zhan, Justin, Sweta Gurung, and Sai Phani Krishna Parsa. "Identification of top-k nodes in large networks using katz centrality." Journal of Big Data 4.1 (2017): 1-19.

[18] Chakraborty, Anwesha, et al. "Application of graph theory in social media." International Journal of Computer Sciences and Engineering 6 (2018): 722-729.

[19] Mao, Chengying, and Weisong Xiao. "A comprehensive algorithm for evaluating node influences in social networks based on preference analysis and random walk." Complexity 2018 (2018).

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

[21] Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. (2003).

[22] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, (2011).

[23] Hafiene, Nesrine, and Wafa Karoui. "A new structural and semantic approach for identifying influential nodes in social networks." IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). IEEE, (2017).

[24] Doo, Myungcheol, and Ling Liu. "Probabilistic diffusion of social influence with incentives." IEEE Transactions on Services Computing 7.3 (2014): 387-400.

[25] Song, Guojie, et al. "Influence maximization on large-scale mobile social network: a divide-and-conquer method." IEEE Transactions on Parallel and Distributed Systems 26.5 (2014): 1379-1392.

[26] Hosseini-Pozveh, Maryam, Kamran Zamanifar, and Ahmad Reza Naghsh-Nilchi. "A community-based approach to identify the most influential nodes in social networks." Journal of Information Science 43.2 (2017): 204-220.

[27] Bae, Joonhyun, and Sangwook Kim. "Identifying and ranking influential spreaders in complex networks by neighborhood coreness." Physica A: Statistical Mechanics and its Applications 395 (2014): 549-559.