Unsupervised Learning-based New Seed-Expanding Approach using Influential Nodes for Community Detection in Social Networks

Khaoula AIT RAI[1], Mustapha MACHKOUR[1], Jilali ANTARI[2]
Computer System and Vision Laboratory-Faculty of Sciences, Ibn Zohr University, Agadir BP8106, Morocco[1]
Laboratory of Computer Systems Engineering-Mathematics and Applications-Polydisciplinary Faculty of Taroudant,
Ibn Zohr University, Morocco[2]

***Abstract*—Several recent studies focus on community structure due to its importance in analyzing and understanding complex networks. Communities are groups of nodes highly connected with themselves and not much connected to the rest of the network. Community detection helps us to understand the properties of the dynamic process within a network. In this paper, we propose a novel seed-centric approach based on TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) and k-means algorithm to find communities in a social network. TOPSIS is used to find the seeds within this network using the benefits of multiple measure centralities. The use of a single centrality to determine seeds within a network, like in classical algorithms of community detection, doesn't succeed in the majority of cases to reach the best selection of seeds. Therefore, we consider all centrality metrics as a multi-attribute of TOPSIS and we rank nodes based on the TOPSIS' relative closeness. The Top-K nodes extracted from TOPSIS will be considered as seeds in the proposed approach. Afterwards, we apply the k-means algorithm using these seeds as starting centroids to detect and construct communities within a social network. The proposed approach is tested on Facebook ego network and validated on the famous dataset having the ground-truth community structure Zachary karate club. Experimental results on Facebook ego network show that the dynamic k-means provides reasonable communities in terms of distribution of nodes. These results are confirmed using Zachary karate club. Two detected communities are detected with higher normalized mutual information NMI and Adjusted Rand Index ARI compared to other seed centric algorithms such as Yasca, LICOD, etc. The proposed method is effective, feasible, and provides better results than other available state-of-the-art community detection algorithms.**

Keywords—Complex network; community detection; TOPSIS; seed-centric approach; ground-truth; k-means

I. Introduction

Many complex real-world systems can be represented and studied as networks. Complex networks cover diverse networks as the Internet, metabolic networks, social networks, and many others. Studies conducted on the physical significance and mathematical properties of complex networks have found that these networks share macroscopic properties. Among these properties, we cite prototype properties such as the small-world effect [1] and the free-scale [2], dynamic properties such as diffusion [3][4] and structural properties such as community structure. The community structure property appears to be common to many complex networks and helps to understand the relationship between a single node in microscopy and groups in macroscopy. Communities are defined as parts of the network with numerous internal connections but few exterior connections. They are closely related to the functional components of real-world networks, such as metabolic networks cycles and pathways and protein complexes in protein-protein interaction networks. They can have very different topological properties than the whole network and then affect the dynamic of the network. Therefore, community structure discovery has been the focus of several recent efforts. Numerous approaches have been suggested to detect community structures in networks, some are based on similarity measures, and others rely on network dynamics such as random walk dynamics [5] and label propagation [6]. Other approaches rely on statistical models end on the optimization of quality functions. For instance, the well-known Newman-Girvan modularity [7] can be used as a method of community discovery and as an objective technique to measure the quality of community partitions.

In this paper, we propose a new approach to discover communities within a social network. The proposed approach uses sequentially two techniques. At the beginning, we apply TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) methodology [8] which aggregates centrality measures (degree centrality [9], betweenness centrality [10], closeness centrality [9], and eigenvector centrality [11]) as multi-attribute to rank nodes in a complex network, and then we run the k-means algorithm which is an unsupervised learning algorithm widely used for data clustering. To take advantage of this algorithm of clustering, we use the output of TOPSIS as initial centroids for k-means to get communities that may be in a complex network. To do so, we need to initialize the value of k by the desired output number of clusters. We did this in two ways; classical k-means and dynamic k-means using Elbow [12] and silhouette [13] methods. TOPSIS is a famous technique of multi criteria decision analysis. It's used in different fields such as Supplier selection [14], selecting the best wood type [15], selecting techniques for future avian influenza surveillance in Canada [16], personnel selection [17], etc.

The main contributions of this work are as follows:

- Taking advantage of different indexes of centrality gathered in TOPSIS to identify influential nodes.
- Based on top-K important nodes extracted from TOPSIS we propose a seed centric algorithm based on k-means to discover communities in complex networks. Experiments are realized using adjacency matrix as centrality measures of top-K influential nodes.
- To determine the optimal K we use Elbow and Silhouette methods and then applying the proposed approach with the optimal value of k.

This paper is then organized as following. Section II presents a general overview on seed centric algorithms. Section III details the general context and related concepts. Section IV explains the proposed approach. Section V presents the analysis and results of the experiments. The paper is concluded in Section VI.

II. Seed Centric Algorithms: General Outlines and Related Works

The algorithm of seed centric approaches is based on three important phases that are: the calculation of seeds, the calculation of seed local community, and community calculation from the previous step. In community detection field, there are several seed-based algorithms proposed by researchers. These algorithms are classified according to several factors. For the first phase for instance, we can find approaches that use single seed or group of seeds, linked or not [18]. Apart from the nature of the seeds (single or group)[19] [20], the number of seeds is also a factor that differs from an algorithm to another. There are some algorithms that have the number of seeds as an input, like the application of the classical k-means algorithm. Other algorithms use heuristic approaches to compute adequate seeds, for example, the approaches proposed by D. Shah et al. [18], [21] and Kanawati [18] based on leaders. D.Shah et al. [22] proposed two algorithms; Leader Follower Algorithm (LFA) and Fast Leader Follower Algorithm (FLFA). These algorithms find leaders as seeds of the community, and then they search its other members. During this process of leaders' search, they consider leaders as nodes with lower degree than their followers. Although the FLFA is fast, LFA can detect more communities in some networks. LICOD algorithm proposed by Kanawati [18] is also based on identifying leaders in the network then affecting the remaining nodes to these leaders to build communities. LICOD calculates the number of communities to detect automatically.

The selection of seeds can be random or informed. Random selection means choosing randomly adequate seeds with repetitive process while informed selection consists of choosing a set of nodes or subgraphs. LICOD algorithm proceeds by informed selection. It selects nodes with higher centrality.

For calculating seed local community, two approaches are applied by researchers in this area; expanding approach and agglomerative approach.

Expanding approach [23] relies on ego-centred algorithms for community detection. The limit of this approach is that it does not take into account all of the network's nodes. To overcome this problem, Whang et al. [24] proposed to attach outliers to the nearest community. The second approach for calculating seed local community is the agglomerative approach, where nodes are agglomerated into communities with nearest identified seed [18] [21].

The last common phase of all community detection algorithms is community calculation which leads to the final communities based on local seeds communities calculated in the precedent step. Fig. 1 illustrates the link between these approaches and algorithms.

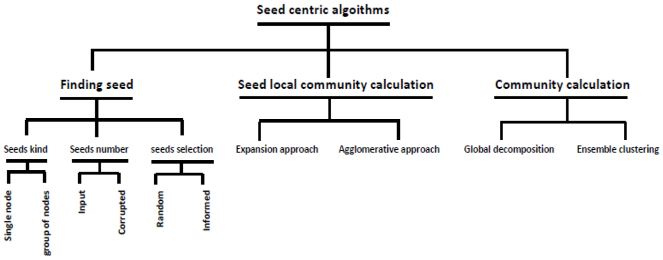


Fig. 1. Summary of seed centric algorithms' characteristics

In the last decade, several seed centric algorithms for community discovery have been proposed in the scientific literature. Each one relies on different seeds finding and different approaches for community calculation. Kanawati et al. [18] proposed Licod algorithm based on two nodes in the community: leaders and followers. Leaders are nodes with high centrality, whatever the centrality is. Followers are founded by computing community membership. Weskida [25] also search in this field how to select seeds using some evolutionary algorithms. The problem in this kind of algorithms is when the evaluation functions are noisy, they will not work. Yufeng Wang et al. [26] proposed an algorithm that selects seeds using page rank-like algorithm. Kanawati proposed also the Yasca algorithm that computes partitions in graphs using local community identification [27]. This algorithm applies an ensemble clustering to the local communities detected for each seed. Akrzewska and Bader proposed also a seed set expansion algorithm of dynamic greedy [28]. In each iteration, the algorithm updates the local communities from an initial partition that is obtained via a set seed expansion method. Table I highlights several seed-centric algorithms' properties.

TABLE I. Properties of Some Seed Centric Algorithms [29]

Algorithm	**Seed Nature**	**Seed number**	**Seed selection**	**Local community**
Licod [18]	Set	Computed	Informed	Agglomerative
Yasca [27]	Single	Computed	Informed	Expansion
[30]	Subgraph	Computed	Informed	Expansion
[24]	Single	Computed	Informed	Expansion
[31]	Subgraph	Computed	Informed	Expansion
[32]	Set	Computed	Automatic	Expansion

III. General Context

In this section, we present fundamental concepts that constitute the general context of the proposed approach. These concepts complete each other to emphasize the proposed idea.

A. *TOPSIS Methodology*

The Technique for Order Preference by Similarity to the Ideal Solution (TOPSIS) [8] emerged in the 1980s as a decision-making method based on several criteria. It chooses the alternative of the shortest Euclidean distance from the ideal solution and the longest distance from the negative ideal solution. TOPSIS demonstrates its power to solve this MCDM (Multi-Criteria Decision Making) problem in different fields such as supply chain management, engineering, health, design, etc. TOPSIS has gained considerable interest from the scientific community due to its success in various areas. TOPSIS was applied for supplier selection [33]. Different factors are considered by Chen et al [33] like quality and technological power. Yong used TOPSIS also to choose the plant location [34]. In the field of human resource management, Kelemenis and Askounis [17] used the fuzzy TOPSIS to select the best management member in an IT department. Wang and Elhag [35] used Fuzzy TOPSIS and nonlinear programming for selecting a system analysis engineer. Kaya et Kahraman [36] applied fuzzy TOPSIS combined with fuzzy AHP to Select the best energy technology alternative.

The process of TOPSIS consists of the following main steps [8]:

Step 1: construction of the normalized decision matrix

Step 2: construction of the weighted normalized decision matrix

Step 3: determination of the positive and the negative ideal solutions

Step 4: calculation of the separation of each alternative

Step 5: the calculation of the relative closeness to the ideal solution.

How each step is modeled and used is explained in the Section III.

B. *Centrality Measures*

The computation of centrality measures has been an important issue in the field of social network analysis for several decades [37]. Centrality is a notion that makes it possible to account the popularity or the visibility of an element within a group. Freeman's article "Centrality in social networks: Conceptual clarification [9]" represents arguably one of the most important contributions in the field of social network analysis. In his article, Freeman proposes three formal definitions of the concept of centrality that we present below. We also present a fourth centrality measure introduced by Bonacich.

1) Degree centrality: It represents the basic and the most concise way of the notion of centrality. It is based on the concept that a person's influence within a group depends on the total number of individuals he knows or has direct contact with [9]. According to this measurement, indicating the value of a node in a graph depends on the number of its neighboring vertices, i.e. the number of its incident links. In graph theory, Degree Centrality originally comes from this number, which is known as the node's degree.

Because it simply considers a node's immediate neighborhood and ignores the overall structure of the network, degree centrality is also known as local centrality measure [38]. While degree centrality is pertinent in some contexts, it is ineffective in others, such as the analysis of web page graphs [39].

2) Closeness centrality: It is a measure of global centrality based on the assumption that a node occupies a strategic (or favorable) position in a graph if it is globally nearby to the other nodes of this graph [9]. For example, in a social network, this metric refers to the fact that an actor is influential if he can quickly get in touch with a lot of other actors while exerting the least amount of effort (the effort here is relative to the size paths).

3) Betweenness centrality: It is another metric of global centrality that Freeman has presented [9]. The idea behind this measurement is that a node in a graph is important if a maximum number of other nodes cross it. More precisely, a node with a strong betweenness centrality is a node through which passes a large number of geodesic paths (i.e. shortest paths) in the graph. In a social network, an actor with a strong betweenness centrality is a node on which depends a large number of interactions between non-adjacent nodes [40]. In a communication network, betweenness centrality of a node can be considered as the probability that information transmitted between two nodes passes through this intermediate node [40].

4) Eigenvector centrality: It’s a measure suggested by Bonacich[11] based the idea that the centrality of a node is determined by the centrality of the nodes to which it is connected. In a social network, this refers to the concept that an actor is influential when he is linked to other influential actors. In fact, it is an extension of degree centrality in which the same weight is not given to the neighboring nodes. Practically talking, Bonacich proposes to consider the centrality of a node as being dependent on the linear combination of the centralities of its neighboring nodes[11].

C. *K-means Algorithm*

K-means is an unsupervised algorithm widely used in data clustering. It proceeds by analyzing a set of data characterized by a set of descriptors, in order to group “similar” data into groups (or clusters).

To split a dataset into k distinct clusters, k-means algorithm needs a way to compare the degree of similarity between all the observations. Usually, the distance between two elements to compute their similarity is used.

Thus, two similar data will have a small dissimilarity distance, while two different data will have a big separation distance. The famous used metric to measure such similarity is the Euclidean distance [41], and it is the one used in this paper.

1) Euclidian distance: It's a geometric distance that considers a matrix X with n quantitative variables in the vector space E^n. The Euclidean distance d between two observations x_1 and x_2 is calculated as follows:

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n}(x_{1n} - x_{2n})^2} \quad (1)$$

Algorithm 1 shows the principal of k-means algorithm. The beginning is the choice of k elements chosen arbitrarily from the dataset as centroids. Then the distance between all the elements and each one of these centroids is computed. Each element belongs then to the cluster whose centroid is closest to it. As a second step, new centroids are computed as the mean of all the elements of each cluster. These steps are repeated until there is no new centroid.

The purpose of clustering algorithms in general and K-Means specifically is to form clusters of similar elements. As long as they are stored in a data matrix, these items can be any kind of thing.

2) K-means principal: The sum of the distances between each item and the centroid is minimized by the iterative algorithm k-means. The outcome depends on the original selection of centroids.

Adopting a cloud of a given set of points, K-Means updates the members of each cluster until the sum can no longer reduce. Depending to picking the right value K for the number of clusters, the outcome is a collection of compact and separate clusters.

Algorithm 1: K-means algorithm

```
INPUT
        K: number of clusters to construct
        The training set
BEGIN
      Randomly choose K points from the dataset that will be
      the centroids of the starting clusters
      REPEAT
      Assign each point (element of the dataset) to the cluster
      to which it is closest
      Update the centroid of each cluster by the mean of its
      points.
      UNTIL convergence OR stabilization of total population
      inertia
END
```

The k-means algorithm may converge under one of the following cases:

- When the number of iterations is fixed in advance, K-means will run its iterations and then end, regardless of how the compound clusters are shaped.
- Stablization of cluster centers (centroids no longer move during iterations).

The main challenge of k-means algorithm is the value of k. Indeed, it is not always evident to choose k as the number of clusters; particularly, if the dataset is sizable, and we don't have a priori assumptions on the data. A big value of k can generate too fragmented partitioning data. This will prevent discovering interesting patterns in the data. That is on one hand; on the other hand, too small value of k will potentially generate general clusters containing a lot of data. In this case, there will be no "fine" patterns to discover. We need then to know the optimal value of k. The most used methods for this purpose are the Elbow and the Silhouette methods and the technique is called dynamic k-means.

D. Dynamic k-means

The challenge of any clustering algorithm is to determine the optimal number of clusters in which the data can be grouped. For the k-means, Elbow and Silhouette methods are the most popular methods for determining this optimal k value.

1) Elbow method: The Elbow Method [12] is one of the most popular methods that helps to find the optimal number of groups to which the k-means algorithm splits the dataset. The idea starts by varying the number of clusters k from 1 to N, assuming that the data has already been divided into k clusters by a clustering method. For each value of k, the WCSS (Within-Cluster Sum of Square) which is the sum of the squared distance between each point and the centroid of a cluster is calculated. The plot generated then in the visualization looks like an elbow. That's why it's nomination. Fig. 2 shows an example of the relation between WCSS values and clusters' number. While the WCSS value is decreasing, the value of k is increasing. The X-axis value where the curve appears as if it starts to bend represents the optimal value of k. This deformation value represents the elbow of the curve.

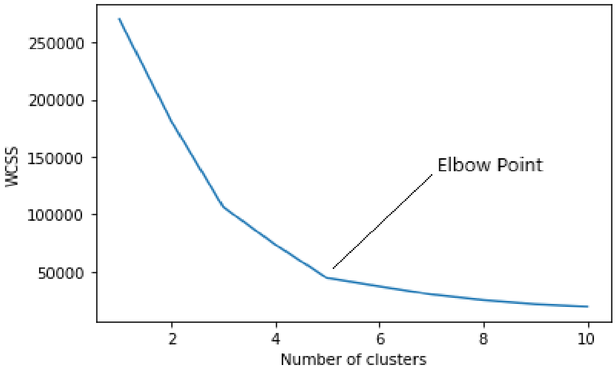


Fig. 2. Example of WCSS plot

2) Silhouette method: The silhouette algorithm [13] is usually used to find the optimal number of clusters for an unsupervised learning technique. In the Silhouette algorithm, we suppose that data has already been split into k groups by a clustering technique (typically k-means algorithm). It proceeds by computing silhouette coefficient for each data point to measure how well this point is assigned to the most appropriate cluster. Its value varies between -1 and 1. A value close to 1 means that the sample is assigned to the best cluster and vice versa. To compute the silhouette coefficient, the following values is needed:

$C(i)$: the cluster assigned to the ith data point.

$|C(i)|$: the number of data points in $C(i)$

$a\ (i)$: gives a measure of the quality of the assignment of the ith data point to its cluster.

$$a\ (i) = \frac{1}{|\ C(i)|-1}\sum_{C(i),i\neq j} d(i,j) \qquad (2)$$

$b(i)$: It is defined as the average dissimilarity with the nearest cluster.

$$b(i) = \ min_{i\neq j}\left(\frac{1}{C(j)}\sum_{j\in C(j)} d(i,j)\right) \qquad (3)$$

The silhouette coefficient $s(i)$ is then given by:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (4)$$

Equation (4) gives the silhouette coefficient for each value of k, and the k having the maximum value of $S(i)$ is taken as the optimal number of clusters for the unsupervised learning algorithm. Fig. 3 illustrates the plot corresponding to the silhouette coefficient's value for each value of k. From that plot we notice that the optimal value of k is 4.

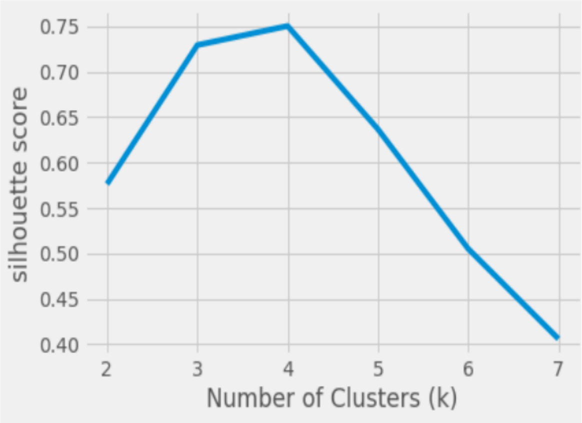


Fig. 3. Silhouette coefficient's plot (maximum value of $S(i)$ is 4)

IV. Proposed Approach

The main goal of the proposed approach is to detect communities in a social network. It's a task of clustering the dataset that requires the application of an appropriate algorithm. The famous algorithm used for clustering is k-means. Since this algorithm requires starting with a value of k, the main idea here is to use the top- K influential nodes within the used dataset. These nodes are detected using TOPSIS methodology and then are used as centroids to run k-means algorithm. These steps are synthesized in Algorithm 2 and Fig. 4.

```
Algorithm 2: Steps of the proposed algorithm
Input:  $G(V,E)$ : A social network,
        $K$ : the number of influential nodes (number of communities),
        $k = \{k_1, k_2, \dots, k_n\}$ : a set of centrality measures,
        $w$ : weight given centralities.
Output: detected communities
Begin
    1. Select the top-K influential nodes $P$ using TOPSIS:
           $P \leftarrow$ TOPSIS($G, k, w, K$)
    2. Use each influential nodes $v \in P$ as a centroid of a community
           $CP_{centroid} \leftarrow \{\{v\} |\ v \in P\ \}$
    3. Enlarge communities into $CP_{centroid}$ :
           $CP \leftarrow$Enlarging $(G, CP_{centroid})$
    4. Return communities detected $CP$
End
```

The main two steps in the proposed method are influential nodes detection using TOPSIS and then the community detection using k-means. These two steps are respectively highlighted in Algorithm 2 and Algorithm 3.

To get nodes' influence ranking using TOPSIS methodology involves four centrality measures as multi-attribute criteria. These measures are degree centrality (DC), betweenness centrality (BC), closeness centrality (CC) and eigen-vector centrality (EC) that have already been introduced in the subsection III.B. Algorithm 3 details technically the five steps highlighted in the subsection III.A.

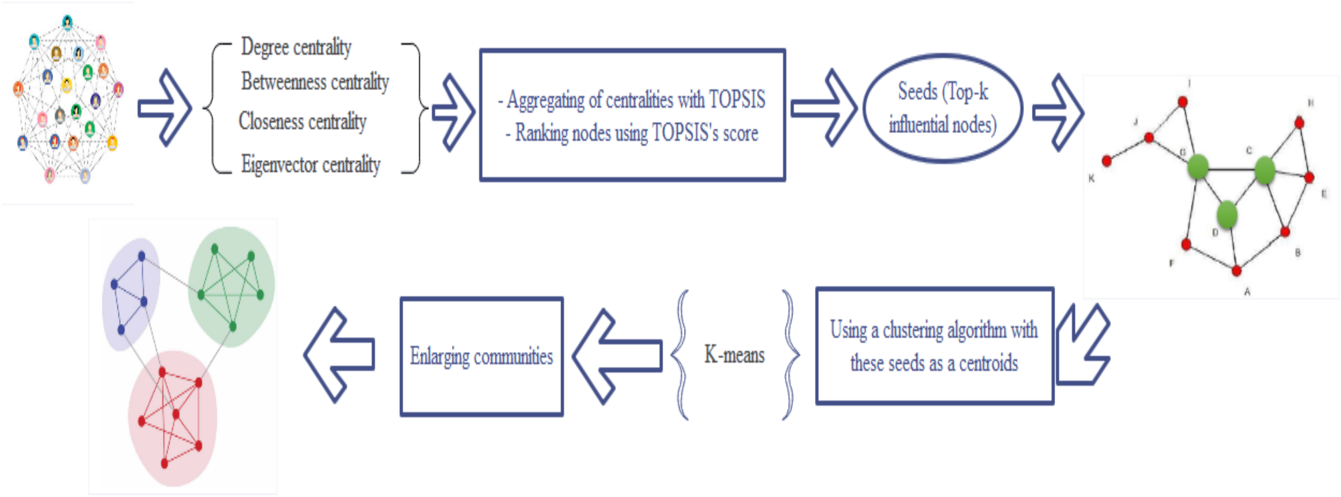

Degree centrality
Betweenness centrality
Closeness centrality
Eigenvector centrality
- Aggregating of centralities with TOPSIS
- Ranking nodes using TOPSIS's score
Seeds (Top-k influential nodes)
Using a clustering algorithm with these seeds as a centroids
K-means
Enlarging communities

Fig. 4. The research methodology for seeds selection and expansion

Algorithm 3: **S**election of top-K influential nodes as centroids using TOPSIS' scores.

```
Input: $G(V,E)$ : A social network,
       $K$ : the number of influential nodes (number of communities),
       $k = \{k_1, k_2, \dots, k_n\}$ : a set of centrality measures,
       $w = \{w_1, w_2, \dots, w_n\}$: a set of weights assigned to centralities.
Output: $P$ top-K influential nodes (a set of centroids)
Begin
    1. Create a decision matrix D based on k measures of centrality where $i = 1,2, \dots n$ and $j = 1,2, \dots k$ and n is the number of nodes
       $D = \begin{bmatrix} d_{11} & \cdots & d_{1k} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nk} \end{bmatrix}$
    2. Normalize the decision matrix D:
       For each $d_{ij}$ $in\ D\ do\ t_{ij} = \frac{d_{ij}}{\sqrt{\sum_{i=1}^{k} d^2{}_{ij}}}$
       $\mathrm{N} = \begin{bmatrix} t_{11} & \cdots & t_{1k} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nk} \end{bmatrix}$
    3. Make the decision matrix normalized and weighted to obtain the matrix R
       For each $t_{ij}$ in N do $r_{ij} = t_{ij} * w_j$
       $\mathrm{R} = \begin{bmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nk} \end{bmatrix}$
    4. Calculate the ideal best solution $A^+$ and negative ideal solution $A^-$ (centrality measures are taken as benefit attributes)
       $A^+ = \max\left(r_{ij}\right) and\ A^- = \min(\ \left(r_{ij}\right)$
    5. Calculate separation from ideal and negative solutions
       $Sep_p = \sqrt{\sum_{j=1}^{k}(r_{ij} - A^+)^2}$ and
       $Sep_n = \sqrt{\sum_{j=1}^{k}(r_{ij} - A^-)^2}$
    6. Calculate the relative closeness score of each node
       For each $r_{ij}$ $in\ R$ do $C_i = \frac{Sep_n}{Sep_n + Sep_p}$
    7. Rank nodes based on $C_i$ the higher score means important node
    8. Select K nodes using top-K $C_i$ : $P \leftarrow top\_k(V, C_i, K)$
End
```

The output of this algorithm is the first k influential nodes. A node's influence depends on its score. The higher the score is, the more influential the node is. The idea of the proposed approach is to use the top-K influential nodes as initial centroids to start k-means algorithm in order to detect communities in the network. The number of these nodes can be specified by the user or found automatically using the elbow method or silhouette method.

The process of community detection in the proposed approach is given in Algorithm 3. Communities within a network can be modeled as groups of individuals that have a kind of similarity. In this paper this similarity is computed by the Euclidean distance between these individuals. The value given to k will corresponds to the number of the output communities. Each top-K influential node resulted from Algorithm 2 is considered as a reference to all the nodes of the network. Euclidean distances are computed between each centroid and all the other nodes. Each centroid gathers around it the nodes that are closest to it, and the new centroid of each cluster is computed by the mean of all its elements. These steps are repeated until there are no new computed centroids. The last clusters are the detected communities in the network and their centroids are the resulted seeds.

Algorithm 4: Computing communities using k-means

```
Input: $(V, E)$ : A social network,
       $CP_{centroid}$ : Top-K influential nodes (the output of
       Algorithm 2)
Output: $CP$ detected communities
Begin
    1. Initialize the centroids with Top-k influential nodes
       /*number of communities to be found/
    2. For each community $CP_j$ , REPEAT
        2.1 Assign each node $v$ to the communities which
            has the closest mean based on their centrality
            measures
                $c^{(i)} = \arg\min_j \left\| v^{(i)} - CP^i_{centroid} \right\|^2$
        2.2 Compute new centroids for each community
                $CP^i_{centroid} = \frac{\sum_{i=1}^{m} 1\{c_{(i)}=j\}v^i}{\sum_{i=1}^{m} 1\{c_{(i)}=j\}}$
    3. UNTIL convergence criteria is reached
    4. END FOR
    5. Return detected communities $CP$
End
```

V. Experimental Setup

In this section a series of experiments are managed to show the efficiency of the proposed approach. In the beginning network with ground truth is used in order to validate the proposed approach referring to the reference of each network. Thereafter, this approach is applied to a larger network without ground truth to find different communities within it. These series of experiments have been performed using Python (3.7.14) as a tool of implementation. The use of Python is argued by its richness by a large number of useful libraries that make the data analysis and computing with visualization easier and simplest. For instance, in these experimentations we took advantage of the librairies; NetworkX for the manipulation of the complex network, Numpy for the scientific computing, Pandas for data analysis, Matplotlib for visualizations, Scikit-learn for clustering, CDlib for evaluating communities.

A. Networks Presentation

1) Facebook ego network: This dataset consists of 'circles' (or 'friends' lists') from Facebook. It includes node features (profiles), circles, and ego networks. It's downloadable from the Stanford large network dataset collection [42]. Fig. 5 presents this dataset. It has 4039 nodes and 88234 edges. Table II presents some characteristics of the dataset.

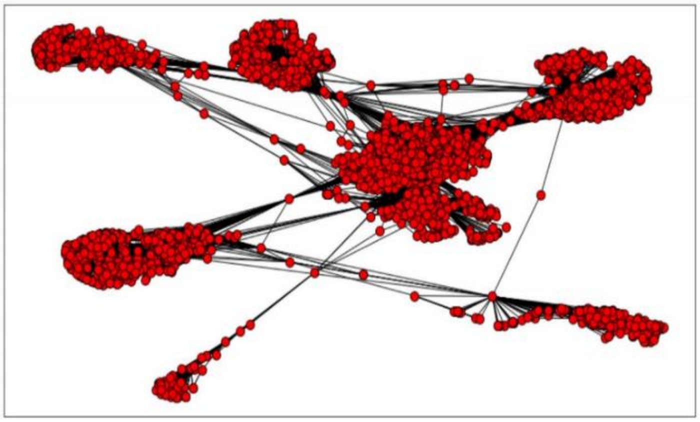


Fig. 5. Facebook ego network

TABLE II. CHARACTERISTICS OF THE DATASET

NODES	4039
EDGES	88234
NODES IN LARGEST WCC	4039 (1.000)
EDGES IN LARGEST WCC	88234 (1.000)
NODES IN LARGEST SCC	4039 (1.000)
EDGES IN LARGEST SCC	88234 (1.000)
AVERAGE CLUSTERING COEFFICIENT	0.6055
NUMBER OF TRIANGLES	1612010
FRACTION OF CLOSED TRIANGLES	0.2647
DIAMETER (LONGEST SHORTEST PATH)	8
90-PERCENTILE EFFECTIVE DIAMETER	4.7

2) Zachary karate club network: It's a real-world network that is well-known in the domain of community structure [43]. The data was collected from a university of karate club in 1977 by Zachary. It represents relationships between members of the karate club. The network is split into two groups after a dispute between the two masters John A and Mr. Hi. Table III presents the characteristics of this dataset.

TABLE III. CHARACTERISTICS OF ZACHARY DATASET

Nodes $\lvert N \rvert$	Edges	Max degree	Average degree	Diameter
34	78	17	4.588 24	5

B. Results and Discussion

This subsection details all the steps of the proposed approach with the two versions of k-means applied on the datasets presented above. We start the experimentation on Facebook dataset and then we validate it on Zachary network that has a ground truth of the detected communities within it.

1) Top-K influential nodes detection: Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. For each ego-network, files for circles, edges, egofeat, feat and featnames are provided. As we are going to cluster people only by their friendship, we only consider the edges file. The first step to implement TOPSIS is the construction of the evaluation matrix. Centrality measures are calculated beforehand and concatenated in a decision matrix. The measures DC, BC, CC and EC constitute the columns of this matrix, and each node in the network constitutes one of its rows. This matrix will be normalized and weighted in the next step to be ready for the following steps, see Algorithm 3.

After implementing and running TOPSIS methodology, we consider k=10 to return the top-10 influential nodes presented by descending sort in Table IV. These nodes are numbered in the column named “Node” and their scores are given in the column named “TOPSIS”. The ranking of these nodes is as following: 107>1684>1912>3437>0>1085>698>567>58>428, with “>” means more influential.

TABLE IV. TOP-10 INFLUENTIAL NODES USING TOPSIS

Node	SCORE
107	0,913277
1684	0,695566
1912	0,496063
3437	0,488865
0	0,304321
1085	0,297379
698	0,231106
567	0,193646
58	0,169464
428	0,132923

The Susceptible-Infected (SI) model is used to look at the spreading effect of top-K influential nodes in order to assess the effectiveness of the ranking model. The SI model is frequently used to study the dynamics of epidemics on networks. Every node in the SI model has two distinctive states:

a) Susceptible S(t) indicates the number of people who are susceptible to the disease but have not yet obtained it;

b) Infected I(t) reflects the number of people who have contracted the disease and are able to disseminate it to susceptible people. For each contaminated node, one randomly sensitive neighbor contracts the disease with probability at each step (here, $\lambda = 0.3$ for uniformity).

For this epidemic model, λ indicates the range across which a node can have an effect on epidemic spreading on networks. Through the intermediaries, an infected node can spread the infection not just to its immediate neighbors but also to its higher order neighbors. In this mode, F(t) stands for the number of contaminated nodes at time t. Using different initially infected nodes, the number of infected nodes should be equal to the overall number of nodes in networks. The average number of infected nodes at each iteration or the spreading rate is the indicator to assess the influence of the initial infected node. The proposed technique is compared with degree, closeness, betweenness, and eigenvector centrality using the SI model. Each implementation identifies the top 10 nodes to infect, and then the SI model is used to determine how the information spreads throughout the network. The influence of the nodes that either presents in the top-10 rankings by the proposed model and the four metrics of centrality are studied. In order to increase the precision of the results, the algorithm will be repeated 10000 times. Thus, the variation (standard deviation) and the mean of each iteration will be calculated. Fig. 6 presents the results. F(t) is the cumulative infected nodes The simulations are on F(t) as a function of time for the proposed network. F(t) increases with time and finally reaches the steady value. According to Fig. 6, the proposed method outperforms DC, CC and EC. The results between BC and the proposed approach are close. Their lines almost overlap as shown in the Fig. 6 and the members of their top-10 lists are the same.

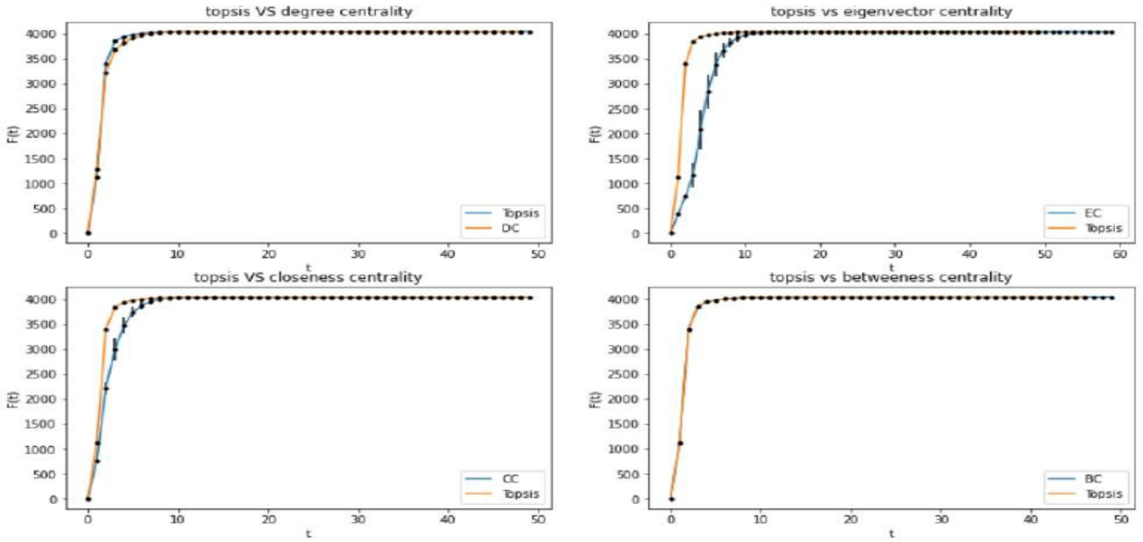


Fig. 6. The cumulative number of infected nodes as a time function, with the initially infected are the top-10 list by the proposed method, DC, CC, BC, and EC. Results are obtained by averaging over 100000 implements (β=0.3).

2) Community detection using static k-means: The second step of the proposed approach is to run k-means algorithm, with k=10. The starting centroids for the k-means are the top-10 influential nodes computed by TOPSIS methodology and given in Table IV. Concretely, theses nodes are:

$$CP^{0}_{centroid} = node_{107},\ CP^{1}_{centroid} node_{1684}, \ldots, CP^{9}_{centroid} = node\_428.$$

The k-means algorithm is applied in two ways; one with the centrality measures and the other with the k-means algorithm is applied in two ways; one with the centrality measures and the other with the adjacency matrix.

a) K-means with centrality measures: In this first way of application of k-means, the centrality measures of top-10 influential nodes are calculated as a training data. Fig. 7 demonstrates the clustering using centrality measures, the big nodes with different colors present the top-10 influential nodes.

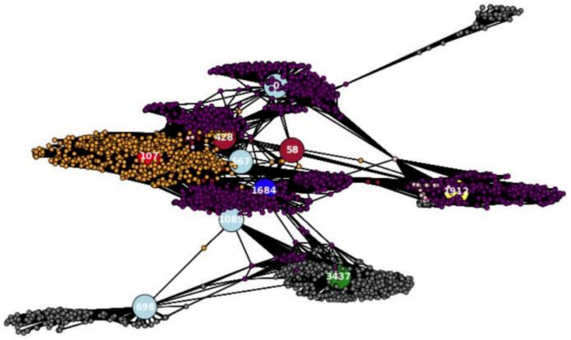


Fig. 7. Clustering based on k-means using centrality measures with initialization with top-10 influential nodes

b) K-means with adjacency matrix: We're going straight in this step to construct the adjacency matrix of the network, then the matrix of the top-10 nodes. An adjacency matrix is a way of representing a graph as a matrix of Booleans; 0 when two nodes are not connected, and 1 when two nodes are connected.

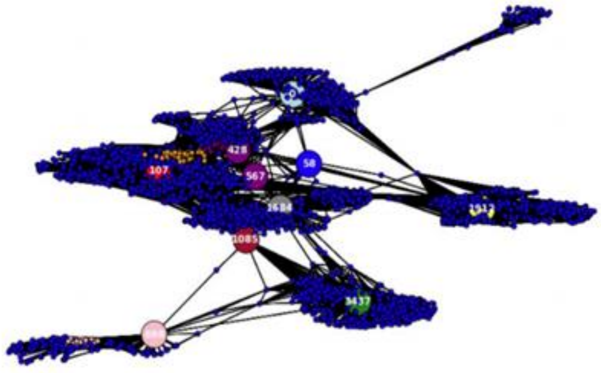
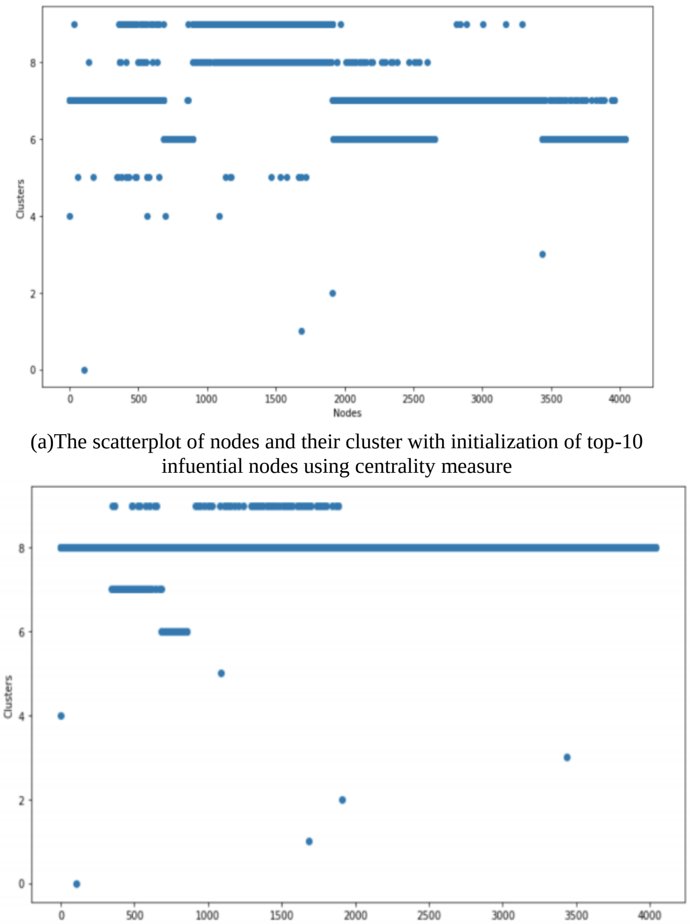


Fig. 8. Clustering based on k-means using adjacency matrix with initialization with top-K influential nodes

According to Fig. 8, it appears that implementing k-means using centrality measures gives good results in terms of distribution of nodes than using adjacency matrix. Although the algorithm is initialized with Top-10 influential nodes, one dominant cluster (blue cluster) is obtained which explains that the majority of nodes are grouped in this cluster. To confirm this result we resort to the scatterplot of these two approaches. The scatterplot below in Fig. 9(a) shows the distribution of nodes by their cluster.



(a)The scatterplot of nodes and their cluster with initialization of top-10 infuential nodes using centrality measure

(b)The scatterplot of nodes and their cluster with initialization of top-10 infuential nodes using adjacency matrix

Fig. 9. Distribution of nodes in their clusters using centrality measures and adjacency matrix

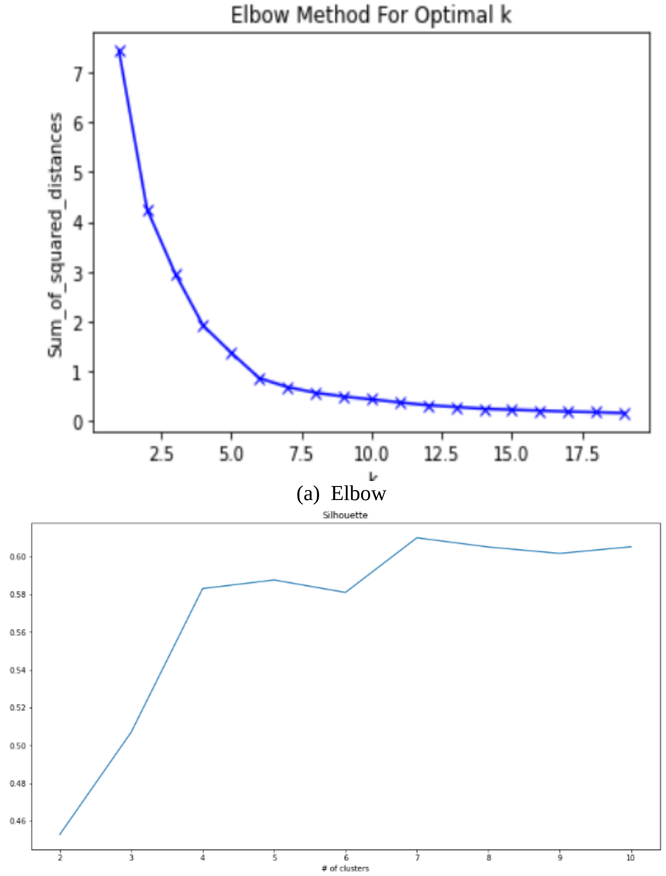
From Fig. 9(b) the cluster 8 contains the majority of the network's nodes; the other nodes are distributed to the cluster 9, the cluster 7 and the cluster 6. Other clusters contain just one node. Fig. 9(a) demonstrates the distribution of nodes using centrality measures as node features for k-means training data. Nodes in this clustering are not also well distributed and the top-10 nodes have been isolated in 6 separate clusters as Table V shows.

In this clustering some of the top-10 nodes are the lonely nodes in their cluster like 107, 1684, 1912 and 3437. While the nodes 0, 1085, 567 and 698 are gathered in one cluster and the two other nodes of the top-10 nodes are grouped with some few nodes in the same cluster. The k-means algorithm using the centrality measures clusters the nodes based on their influence in the network, that's why we notice that the more the nodes have more influence in the network, the more they are isolated in other clusters. This kind of clustering can be applied to minimize the influence in social networks. The influence minimization can predict the spread of deprecatory rumors, fake news, and spread of disease.

TABLE V. DISTRIBUTIONS OF INFLUENTIAL NODES IN THEIR FINAL CLUSTERS

Node	TOPSIS	Cluster
107	0.913277	0
1684	0.695566	1
1912	0.496063	2
3437	0.488865	3
0	0.304321	4
1085	0.297379	4
698	0.231106	4
567	0.193646	4
58	0.169464	5
428	0.132923	5

3) Community detection using dynamic k-means: Although the proposed approach is simple and fast, yet there are also some limitations. From the scatterplots in Fig. 9, we conclude that the value 10 given to k is big and not precise. We need then to assign an optimal value to k before starting the k-means algorithm. It's hard to know this value from the beginning, reason for what we need to apply the dynamic K-means. As discussed in the section D, we implement Elbow and Silhouette methods for this purpose. The two approaches based on the centrality measures and also the adjacency matrix is also used in this context. Centrality measures exceeds adjacency matrix in terms of defining the optimal-k. The problem of working with the adjacency matrix is the instability of defining the optimal k because we get different values of k for each execution. Hence, we focus on centrality measures to define the optimal value for k. The value generated using the Elbow method and Silhouette method are respectively k=6 and k=7, Fig. 10. From the previous experiments where k=10, we concluded that big values of k don't give good clustering. So we consider the optimal value of k is 6 and then we apply the proposed approach using the first six influential nodes as starting centroids for k-means algorithm. The output of the proposed approach into six clusters is highlighted in Fig. 11 below.



(a) Elbow

(b) Silouhette

Fig. 10. The clustering of Facebook ego network using dynamic k-means

According to the graph in Fig. 11, it seems that we have four dominant communities in the network which is confirmed by the scatterplot in Fig. 12. The distribution of nodes to the detected communities looks reasonable in this results compared to the previous one. This means that when k is well defined, the proposed approach gives better results. These results need to be confirmed on datasets with ground truth.

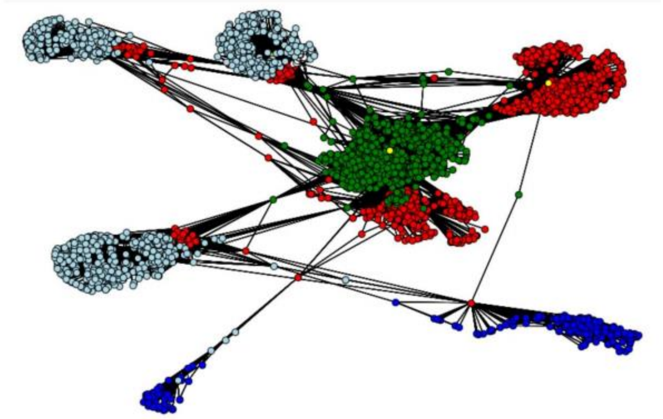


Fig. 11. The clustering of facebook ego network using dynamic k-means

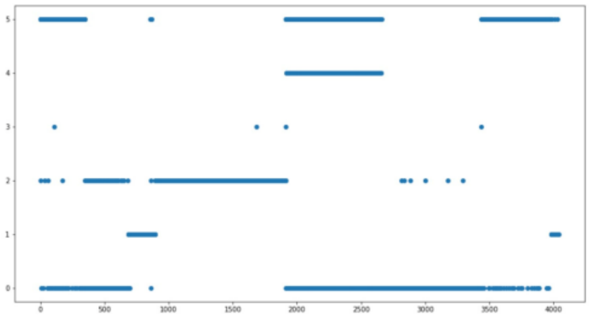


Fig. 12. The scatterplot of Facebook ego network using dynamic k-means

4) Evaluation metrics: It is necessary to use specific performance measures to measure the similarity between the two dataset partitions. In this experimental study, two classical measures are used.

The first one is the Rand index [44]. It is the portion of point pairs $(x1, x2)$ that are organized similarly in both divisions. Either $x1$ and $x2$ are members of the same cluster in both the situations or they are members of different clusters.

The Rand index can be inflated artificially by predicting many clusters. Numerous pairs of points will belong to separate clusters, and the possibility that two points with different labels will be found in two different clusters will be considerable. This effect is remedied by The Adjusted Rand Index (ARI), which normalizes the Rand Index (RI) [45]:

$$ARI = \frac{RI-E(RI)}{\max(RI)-E(RI)} \qquad (5)$$

Where $E(RI)$ is the expectation of the value of the Rand index, in other words, the index obtained by randomly splitting the data. This adjusted index is only close to 1 when the clustering exactly matches the original partition and is close to 0 otherwise.

The second measure used in this experimental study is the normalized mutual information [46]. In probability theory and information theory, the mutual information of two random variables is a quantity measuring of the statistical dependence of these variables. Normalized mutual information NMI is a variant of mutual information. Its value is between 0 and 1. The closer NMI is to 1, the closer the result is to the ground truth. It can be defined as:

$$NMI\ (X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \qquad (6)$$

Where, $I()$ is the mutual information and $H()$ is the entropy.

5) Experiments on dataset with ground-truth: The evaluation of detected communities is still an open issue in the scientific community because of the lack of models and references. Since the availability of a ground truth community structure for large real networks is difficult, we choose to validate the proposed approach on Zachary karate club network.

In this experiment, we rely on the dynamic k-means to get the optimal value of k. In this case the optimal k=2 by the silhouette method as shown in Fig. 13. We test k-means with top-2 influential nodes using centrality measures but the results are not satisfactory reason for which we use dynamic k-means with the adjacency matrix on Zachary network. Fig. 15 shows the obtained results and Fig. 14 gives the ground-truth community structure of Zachary karate club.

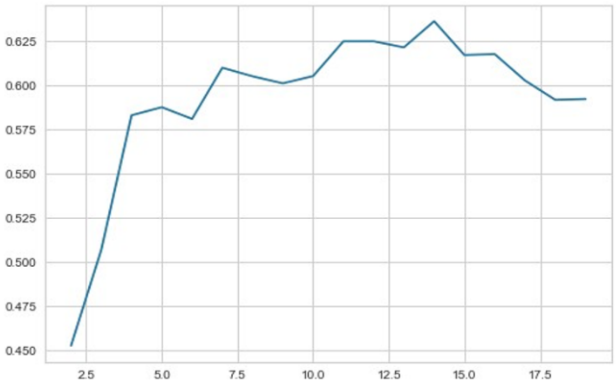


Fig. 13. The optimal-k of Zachary dataset given by silhouette method

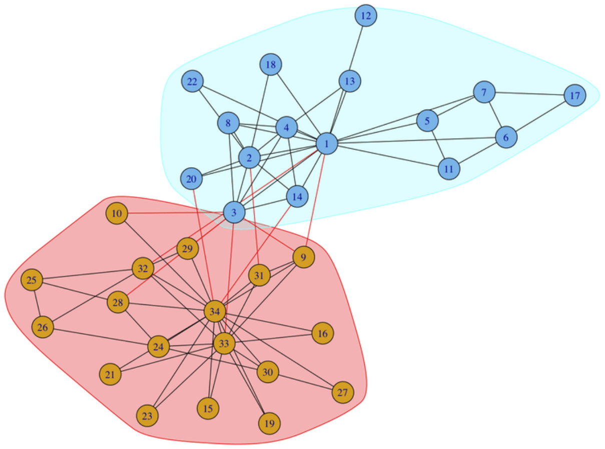


Fig. 14. The ground-truth community structure of Zachary karate club

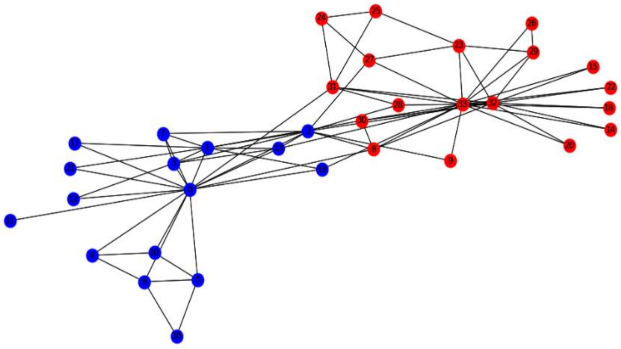


Fig. 15. The community structure detected by the proposed approach

Table VI presents the obtained results on Zachary compared to other algorithms of community detection like Louvain [47], Newman [48], Walktrap [5], Licod [18] and Yasca [27]. Compared to the ground-truth community structure, the proposed approach also detects 2 communities in Zachary as displayed in Fig. 15. Although some nodes deviate from the ground-truth community structure, the quality of the discovered communities is higher than other algorithms as mentioned in Table VI. NMI and ARI values are higher compared to other community detection algorithms.

TABLE VI. COMPARISON BETWEEN ALGORITHMS OF COMMUNITY DETECTION AND THE PROPOSED APPROACH ON ZACHARY KARATE CLUB.

	Algorithm	**ARI**	**NMI**	**Detected communities**
Zachary (2 groups)	Newman	0,46	0,57	5
	Louvain	0,46	0,58	4
	Walktrap	0,33	0,50	5
	Licod	0,62	0,60	3
	YASCA	0,69	0,77	2
	Proposed ML approach	**0,88**	**0,83**	2

VI. CONCLUSION

In this paper, a new seed-expanding algorithm to detect communities in social networks is introduced. The selection of seeds is based on centrality measures, gathered as multi-attribute in TOPSIS. The Top-K influential nodes obtained by TOPSIS methodology are used as starting centroids to run the k-means algorithm. In the experimental study, k-means is implemented using centrality measures and adjacency matrix as training data to compare the results. We notice that on large real networks like Facebook, the centrality measures behave well than adjacency matrix in terms of nodes distribution except that we found difficulties to determine the good values of k for the clustering. That is why we proceed to dynamic k-means using Elbow and silhouette methods to get the optimal values of k. Thereafter we observe that the results seem more reasonable with the dynamic k-means. Eventually, for assessing the quality of the discovered communities, we use Zachary karate club to validate the proposed model instead of Facebook because of the unavailability of a ground-truth for Facebook. ARI and NMI are two measures used for the test, and the results show that the proposed approach is better than other available algorithms for community detection. Many possible future directions have been opened up by this work. For example, implementing other clustering methods for the purpose of detecting communities such as K-medoids, C-means, fuzzy k-means and compare the results with the K-mean algorithms. As we can extend this work on other networks apart from social networks by using other centralities in the initial phase which are adequate for these types of networks.

REFERENCES

[1] D. J. Watts et S. H. Strogatz, « Collective dynamics of "small-world" networks », Nature, vol. 393, no 6684, p. 440-442, juin 1998, doi: 10.1038/30918.

[2] A.-L. Barabási et R. Albert, « Emergence of Scaling in Random Networks », Science, vol. 286, no 5439, p. 509-512, oct. 1999, doi: 10.1126/science.286.5439.509.

[3] S. Bilke et C. Peterson, « Topological properties of citation and metabolic networks », Phys. Rev. E, vol. 64, no 3, p. 036106, août 2001, doi: 10.1103/PhysRevE.64.036106.

[4] J. Eriksen, M. K. Jensen, P. Sjøgren, O. Ekholm, et N. K. Rasmussen, « Epidemiology of chronic non-malignant pain in Denmark », Pain, vol. 106, no 3, p. 221-228, déc. 2003, doi: 10.1016/S0304-3959(03)00225-2.

[5] P. Pons et M. Latapy, « Computing communities in large networks using random walks », p. 20.

[6] G. Cordasco et L. Gargano, « Label propagation algorithm: a semi-synchronous approach », Int. J. Soc. Netw. Min., vol. 1, no 1, p. 3, 2012, doi: 10.1504/IJSNM.2012.045103.

[7] M. E. J. Newman et M. Girvan, « Finding and evaluating community structure in networks », Phys. Rev. E, vol. 69, no 2, p. 026113, févr. 2004, doi: 10.1103/PhysRevE.69.026113.

[8] S.-J. Chen et C.-L. Hwang, « Fuzzy Multiple Attribute Decision Making Methods », in Fuzzy Multiple Attribute Decision Making, vol. 375, Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, p. 289-486. doi: 10.1007/978-3-642-46768-4_5.

[9] L. C. Freeman, « Centrality in social networks conceptual clarification », Soc. Netw., vol. 1, no 3, p. 215-239, janv. 1978, doi: 10.1016/0378-8733(78)90021-7.

[10] Ulrik Brandes, « A faster algorithm for betweenness centrality », The Journal of Mathematical Sociology, p. 163-177, 2001.

[11] P. Bonacich et P. Lloyd, « Eigenvector-like measures of centrality for asymmetric relations », Soc. Netw., vol. 23, no 3, p. 191-201, juill. 2001, doi: 10.1016/S0378-8733(01)00038-7.

[12] R. Nainggolan, R. Perangin-angin, E. Simarmata, et A. F. Tarigan, « Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method », J. Phys. Conf. Ser., vol. 1361, no 1, p. 012015, nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[13] P. J. Rousseeuw, « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », J. Comput. Appl. Math., vol. 20, p. 53-65, nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[14] H. Gupta et M. K. Barua, « Supplier selection among SMEs on the basis of their green innovation ability using BWM and fuzzy TOPSIS », J. Clean. Prod., vol. 152, p. 242-258, mai 2017, doi: 10.1016/j.jclepro.2017.03.125.

[15] « (26) Analizing Topsis Method for Selecting the Best Wood Type | Ria Sari - Academia.edu ». https://www.academia.edu/38541712/Analizing_Topsis_Method_for_Selecting_the_Best_Wood_Type (consulté le 21 avril 2022).

[16] F. El Allaki, J. Christensen, et A. Vallières, « A modified TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) applied to choosing appropriate selection methods in ongoing surveillance for Avian Influenza in Canada », Prev. Vet. Med., vol. 165, p. 36-43, avr. 2019, doi: 10.1016/j.prevetmed.2019.02.006.

[17] A. Kelemenis et D. Askounis, « A new TOPSIS-based multi-criteria approach to personnel selection », Expert Syst. Appl., vol. 37, no 7, p. 4999-5008, juill. 2010, doi: 10.1016/j.eswa.2009.12.013.

[18] R. Kanawati, LICOD: Leaders Identification for Community Detection in Complex Networks. 2011, p. 582. doi: 10.1109/PASSAT/SocialCom.2011.206.

[19] S. Papadopoulos, Y. Kompatsiaris, et A. Vakali, « A Graph-Based Clustering Scheme for Identifying Related Tags in Folksonomies », in Data Warehousing and Knowledge Discovery, Berlin, Heidelberg, 2010, p. 65-76. doi: 10.1007/978-3-642-15105-7_6.

[20] B. Bollobas et O. Riordan, « Clique percolation », Random Struct. Algorithms, vol. 35, no 3, p. 294-322, oct. 2009, doi: 10.1002/rsa.20270.

[21] D. Shah et T. Zaman, « Community Detection in Networks: The Leader-Follower Algorithm », ArXiv, 2010.

[22] D. Parthasarathy, D. Shah, et T. Zaman, « Leaders, Followers, and Community Detection », p. 9, 2018.

[23] M. Danisch, J.-L. Guillaume, et B. Le Grand, « Unfolding Ego-Centered Community Structures with "A Similarity Approach" », in Complex Networks IV, vol. 476, G. Ghoshal, J. Poncela-Casasnovas, et R. Tolksdorf, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 145-153. doi: 10.1007/978-3-642-36844-8_14.

[24] J. J. Whang, D. F. Gleich, et I. S. Dhillon, « Overlapping community detection using seed set expansion », in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13, San Francisco, California, USA, 2013, p. 2099-2108. doi: 10.1145/2505515.2505535.

[25] M. Weskida et R. Michalski, « Evolutionary algorithm for seed selection in social influence process », in Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Davis, California, août 2016, p. 1189-1196.

[26] Y. Wang, B. Zhang, A. V. Vasilakos, et J. Ma, « PRDiscount: A Heuristic Scheme of Initial Seeds Selection for Diffusion Maximization in Social Networks », in Intelligent Computing Theory, Cham, 2014, p. 149-161. doi: 10.1007/978-3-319-09333-8_17.

[27] R. Kanawati, YASCA: An Ensemble-Based Approach for Community Detection in Complex Networks. 2014. doi: 10.1007/978-3-319-08783-2_57.

[28] A. Zakrzewska et D. A. Bader, « A Dynamic Algorithm for Local Community Detection in Graphs », in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, New York, NY, USA, août 2015, p. 559-564. doi: 10.1145/2808797.2809375.

[29] B. R V, E. Kanaga, et P. Bródka, « Overlapping community detection using superior seed set selection in social networks », 10 août 2018.

[30] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, et P. Spyridonos, « Community detection in Social Media », Data Min. Knowl. Discov., vol. 24, no 3, p. 515-554, mai 2012, doi: 10.1007/s10618-011-0224-z.

[31] B. Bollobas et O. Riordan, « Clique percolation », Random Struct. Algorithms, vol. 35, no 3, p. 294-322, oct. 2009, doi: 10.1002/rsa.20270.

[32] J. J. Whang, D. F. Gleich, et I. S. Dhillon, « Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion », IEEE Trans. Knowl. Data Eng., vol. 28, no 5, p. 1272-1284, mai 2016, doi: 10.1109/TKDE.2016.2518687.

[33] C.-T. Chen, C.-T. Lin, et S.-F. Huang, « A fuzzy approach for supplier evaluation and selection in supply chain management », Int. J. Prod. Econ., vol. 102, no 2, p. 289-301, 2006.

[34] « Plant location selection based on fuzzy TOPSIS | SpringerLink ». https://link.springer.com/article/10.1007/s00170-004-2436-5 (consulté le 16 mai 2022).

[35] Y.-M. Wang et T. M. S. Elhag, « Fuzzy TOPSIS method based on alpha level sets with an application to bridge risk assessment », Expert Syst. Appl., vol. 31, no 2, p. 309-319, août 2006, doi: 10.1016/j.eswa.2005.09.040.

[36] T. Kaya et C. Kahraman, « Multicriteria decision making in energy planning using a modified fuzzy TOPSIS methodology », Expert Syst. Appl. Int. J., vol. 38, no 6, p. 6577-6585, juin 2011, doi: 10.1016/j.eswa.2010.11.081.

[37] S. Wasserman et K. Faust, Social network analysis: methods and applications. 1994.

[38] J. Scott, Social Network Analysis: A Handbook. SAGE, 2000.

[39] J. M. Kleinberg, « Authoritative sources in a hyperlinked environment », J. ACM, vol. 46, no 5, p. 604-632, sept. 1999, doi: 10.1145/324133.324140.

[40] S. P. Borgatti et M. G. Everett, « Models of core/periphery structures », Soc. Netw., vol. 21, no 4, p. 375-395, oct. 2000, doi: 10.1016/S0378-8733(99)00019-2.

[41] N. Krislock et H. Wolkowicz, « Euclidean distance matrices and applications », in Handbook on semidefinite, conic and polynomial optimization, Springer, 2012, p. 879-914.

[42] « SNAP: Network datasets: Social circles ». https://snap.stanford.edu/data/ego-Facebook.html (consulté le 17 mai 2022).

[43] W. W. Zachary, « An Information Flow Model for Conflict and Fission in Small Groups », J. Anthropol. Res., vol. 33, no 4, p. 452-473, 1977.

[44] « Objective Criteria for the Evaluation of Clustering Methods on JSTOR ». https://www.jstor.org/stable/2284239 (consulté le 13 juillet 2022).

[45] L. Hubert et P. Arabie, « Comparing partitions », J. Classif., vol. 2, no 1, p. 193-218, déc. 1985, doi: 10.1007/BF01908075.

[46] « [PDF] Robust data clustering | Semantic Scholar ». https://www.semanticscholar.org/paper/Robust-data-clustering-Fred-Jain/9e7a86fd9e15bf37d937a79ccb7efb78bb070f74 (consulté le 13 juillet 2022).

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, et E. Lefebvre, « Fast unfolding of communities in large networks », J. Stat. Mech. Theory Exp., vol. 2008, no 10, p. P10008, oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[48] M. E. J. Newman, « Fast algorithm for detecting community structure in networks », Phys. Rev. E, vol. 69, no 6, p. 066133, juin 2004, doi: 10.1103/PhysRevE.69.066133.