



Fuzzy community detection on the basis of similarities in structural/attribute in large-scale social networks

Mansoureh Naderipour¹ · Mohammad Hossein Fazel Zarandi¹ · Susan Bastani²

Published online: 12 April 2021

© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Community detection aims to partition a set of nodes with more similarities in the set than out of it based on different criteria like neighborhood similarity or vertex connectivity. Most present day community detection methods principally concentrate on the topological structure, largely ignoring the heterogeneous properties of the vertex. This paper proposes a new community detection model, based on the possibilistic c-means model, by using structural as well as attribute similarities in a large scale in social networks. In the majority of real social networks, different clusters share nodes, resulting in the formation of overlapping communities. The proposed model, on the basis of structural and attribute similarity (PCMSA), serves as a fuzzy community detection model addressing the overlapping community detection problem, and detecting communities in a way that each community has a densely connected sub-graph with homogeneous attribute values. The function of the proposed model is assessed by a trade-off between intra-cluster and inter-cluster density and homogeneity. Therefore, to validate the proposed community detection algorithm (PCMSA) and its results, an index, compatible with the proposed model, is defined; and to assess the efficiency of the proposed fuzzy community detection, several experimental results in variety sizes from very small to very large sizes of real social networks are given, and the results are contrasted with other community detection models like FCAN, CODICIL, SA-cluster, K-SNAP and PCM. The experimental findings reveal the superiority of this novel model and its promising scalability and computational complexity over others.

Keywords Community detection · Large-scale social networks · Overlapping communities · Possibilistic c-means · Structural and Attribute similarity · Validation index

✉ Mohammad Hossein Fazel Zarandi
zarandi@aut.ac.ir

Mansoureh Naderipour
m.naderipour@aut.ac.ir

Susan Bastani
sbastani@alzahra.ac.ir

¹ Department of Industrial Engineering, Amirkabir University of Technology (Polytechnic of Tehran), P.O. Box 15875-4413, Tehran, Iran

² Department of Sociology, Alzahra University, 19938-93973, Tehran, Iran

1 Introduction

Networks have been studied in many fields such as biology, mathematics, quantitative geography, sociology, and information science (Fortunato 2010). A graph is made by a number of nodes (vertices) and some links (edges) that join them to each other (Schaeffer 2007). A community (cluster) is made of a series of vertices with common or similar properties based on various criteria (Fortunato 2010). Graph clustering or community detection refers to grouping nodes connected with edges, but not to outside the group (Fortunato 2010; Schaeffer 2007). For example, strategies related to target marketing can be designed well if community detection is possible in such networks. If users are regarded as vertices, and friendship relationships are regarded as edges, graph clustering can formulate the issue of detecting communities for these users for target marketing.

1.1 Research challenge

A significant characteristic of real-world networks is community structure (Fu et al. 2013) in which people have social relations sharing similar personal or professional interests, records or real-life relationships. Different communities can share nodes in graphs; therefore, overlapping can be formed among communities (Zarandi and Razaee 2010), and detecting such communities in social networks, which exists in most real social networks, is of great importance. In recent research, some fuzzy methods such as fuzzy c -means (FCM) clustering (Zhang et al. 2007; Zarinbal et al. 2014) and the possibilistic c -means clustering model (PCM) (Krishnapuram and Keller 1993) have been put forward for discovering these overlapping communities.

The problem studied in this research is to detect communities on the basis of attribute and structural similarities. The goal is to partition the graph into c communities, each of which has cohesive structures and homogeneous attributes. This is somewhat challenging because these two similarities are independent or even conflicting goals. For example, authors who cooperate with each other may have different attributes, such as research topics, whereas those who search the same topics may come from different groups and never cooperate. It is unknown how to balance these two sources of data. Most researches design a distance function between two vertices combining the structural distance and attribute distance with two different weighting factors. Although this procedure is simple, it is hard to set the factors and interpret the function so that it is not clear whether the weight of coauthor relationship should be larger or smaller than that of research topic. Moreover, making quantitative decisions on the weights is even harder.

1.2 Main contribution

This study proposes a fuzzy model and algorithm for detecting communities that overlap on the basis of analyzing the semantic of social networks data. Nodes share common attributes in groups or communities and they have many connections among themselves. Therefore, there are two sources of data for performing the clustering task. The first is information about the nodes and their attributes such as known properties, users' profiles in social networks or authors' publications; and the second comes from the set of connections between nodes such as interactions and collaborations that form among users.

Fuzzy clustering is very useful for cluster analysis (Yang 1993; Valente de Oliveira and Pedrycz 2007). Considering fuzzy sets in detecting communities can make it possible to

identify clusters due to their various impressions in link and attribute information. Nodes with fuzzy clustering are assigned to one or more clusters with different membership functions, making it possible to have overlapping and interesting clusters of various and flexible structures. Because of these advantages, a fuzzy clustering is proposed in order to identify clusters existing in complex networks using both link information and node attribute. Determining the membership functions for assigning each node to clusters based on node attribute and link information is challenging in fuzzy clustering. Considering this problem, a new model called PCMSA is suggested to identify overlapping clusters on the basis of attribute and structural similarities. The findings indicate that PCMSA is a considerable model for detecting communities in a complex network. Here is a summary of major contributions of this research:

- (1) Community detection in social networks based on both link information and node attribute due to the importance of these sources of data in some real graphs such as social networks
- (2) Fuzzy clustering that makes it possible for overlapping clusters in which nodes are assigned to one or more clusters that have various degrees of memberships
- (3) Strict structural and attribute similarity: in the last algorithms of graph clustering, most algorithms consider weighting factors to balance between attribute and structural similarities; however, the algorithm in this paper (PCMSA) strictly considers the two similarities.

The organization of this paper is as follows: The 2nd section presents the related works. The 3rd section introduces fuzzy clustering based on center, and the 4th section deals with the explanation of the proposed algorithm and importance of weak ties. Sections 5, 6 present the clustering validation index and experimental results, respectively, and conclusions and suggestions for further research work will be presented in the 7th section.

2 Related works

In recent years, fuzzy clustering has been created and widely used in general clustering, but little research has applied it in graph clustering (Schaeffer 2007) such that using fuzzy clustering in graph clustering has been observed less during the past decade (Wang et al. 2013). It is still possible to improve the performance of some methods that are meant for discovering fuzzy overlapping communities (Schaeffer 2007). FCM is among the most common fuzzy clustering models used along with other techniques to detect communities (Zhang et al. 2007). In these studies, the structure of the models is not adapted well enough for graph clustering. Golsefid et al. proposed a fuzzy duo-centric model for community detection in social networks for which the nodes' properties are not considered in this paper (Golsefid et al. 2015).

Most graph clustering considers only one aspect of the graph and ignores the other (Andersen et al. 2006; Flake et al. 2000; Girvan and Newman 2002; Tian et al. 2008). Consequently, the clusters either have a random distribution of vertex attributes in them, or have a non-cohesive intra-cluster structure. A good graph clustering ought to balance similarities that are both structural and attribute in order to have an intra-cluster structure that is cohesive and has homogeneous vertex properties. However, considering node attributes

and network topology together is also challenging so that one must combine two very different pieces of information.

Recently, some attempts take both sources of data into consideration. Such clustering algorithms are based either on distance or on model. Distance-based methods (Zhou et al. 2009, 2010; Ruan et al. 2013) initially form an augmented network by adding the virtual links in order to connect the attributes with nodes. In that case, the clusters can be identified with the similarity between two nodes that have standard clustering algorithms (Markov or K-Medoids clustering), calculated by the distance between nodes in the augmented network. There are two challenges with this algorithm: adding new nodes and new edges leads to a big graph that cannot be solved in some cases. Moreover, it is not clear how to cluster this heterogeneous graph with two types of nodes and edges (Zhou et al. 2010). Model-based approaches, both generative and discriminative models, have been developed to simulate the complex network generated by various Bayesian networks with topic modeling. In the literature, for the first one, there exist papers such as CART (Pathak et al. 2008), iTopic Model (Sun et al. 2009), and for the second one, there exist papers such as PCL-DC (Yang et al. 2009). Moreover, Cao et al. detect prosumer-community groups considering nodes' attributes and network structure, but they do not consider the overlapping communities. Their algorithm also cannot detect communities in large-scale networks (Cao et al. 2019). Bu et al. propose GK-mean algorithm which is formulated as a multi-objective optimization problem (MOOP). Although Graph K-means (GK-mean) algorithm considers two topological structure and attribute information, but do not work well on large-scale networks (Bu, et al. 2019).

The above clustering algorithms have a drawback in clustering in large-scale networks. Some of them did not consider both sources of data or some have taken the assumptions to the problem more easily. Moreover, they did not consider weak ties in their model in addition to the strong ties (weak ties are discussed in Sect. 4).

This is while, the proposed clustering approach based on structural and attribute similarity (PCMSA), is designed based on the semantic of social networks data in addition to the fuzzy sets by considering the important theorems in overlapping social networks such as, weak ties and homophily. These two theorems are explained in Sect. 4. Moreover, the proposed model is evaluated by an extensive evaluation using different network sizes and even real large graphs. Results show clusters with high quality, homogeneous attributes and cohesive structures.

Until now, methods that consider both topology structure and node attribute have not considered the fuzzy nature in extent of membership in graph clustering. Although, Hu and Chan propose a fuzzy clustering based on two sources of data (Hu and Chan (2016)), but in their problem formulation, the structure of the model is not adapted well enough for analyzing the semantic of fuzzy social networks and overlapping complex networks. As discussed above, in their problem formulation, the structural distance and attribute distance are combined, while they are two seemingly independent, or even conflicting goals and it does not make sense semantically.

3 Background

Fuzzy Center-Based Clustering (FCM) and the possibilistic c-means clustering (PCM) are presented in this section.

3.1 Fuzzy clustering based on center

In fuzzy clustering, each node forms a part of a cluster that has a membership function between 0 (not belonging), and 1(belonging), and each node can form a part of several clusters that have different membership degrees which are crisp values over the interval [0,1] (Höppners 1999). The most famous fuzzy clustering suggested by Dunn (Dunn (1974)) and continued by Bezdek (Bezdek (1981)) is FCM clustering algorithm.

If $X = \{x_1, x_2, \dots, x_n\} \in R^\alpha$ is a series of feature vectors (α and n are the dimension and the number of nodes, respectively), FCM assigns nodes to clusters by making the subsequence function minimum, and partitions nodes into c clusters.

$$J = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|x_k - v_i\|_R^2 \tag{1}$$

where, $u_{ik} \in [0, 1]$ is the membership degree of each node $k = 1, \dots, n$ in cluster $i = 1, \dots, c$, and $v = \{v_1, \dots, v_c\} \in R^{c\alpha}$ is indicative of a series of clusters' centers. R shows the distance norm (Fortunato 2010) and $1 \leq m < \infty$ indicates the fuzzifier parameter. Clustering based on the objective function can be considered an optimization problem, which is solved by the gradient descent technique (Tan et al. 2007).

Krishnapuram and Keller proposed the PCM clustering algorithm in order to decrease the impact of outliers on FCM with relaxation of the condition of membership values to all clusters for each node, which equals 1, and replaced it with $\max_i(u_{ik}) > 0, 1 \leq k \leq n$. The PCM objective function is defined as:

$$J = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|x_k - v_i\|_R^2 + \sum_{i=1}^c \beta_i \sum_{k=1}^n (1 - u_{ik})^m \tag{2}$$

where β_i is the average fuzzy intra-cluster distance of cluster i .

In a complex network, a number of nodes are joined to one another in a topological structure. By considering nodes as vertices, and links as edges, complex networks could be considered graphs. Many real-world problems such as social networks (Myspace, Facebook,...) have millions of users who are connected to one another as friends(Fortunato 2010). By considering users as vertices, and friendship relationships as edges, we can formulate the issue of assigning users to communities as a problem of community detection. An ideal cluster should possess an intra-cluster structure that is cohesive and has homogeneous vertex attributes. PCM considers only one aspect of the graph related to the nodes' attributes, and ignores the other aspect related to the structure of the nodes. In this paper, both aspects are considered to detect communities considering attribute/structural similarities based on the PCM algorithm.

4 Proposed fuzzy clustering considering attribute and structural similarities

When a graph is given, different criteria can be defined to identify different graph clusters. By considering both structure and properties of the nodes, the proposed model PCMSA detects ideal communities with the following criteria:

- (1) It identifies communities that have more densely connected nodes.

- (2) It identifies communities that have nodes more strongly related to each other.
- (3) The probability of adjacent nodes to belong to the same community is higher.

For PCMSA, the above criteria are considered by employing fuzzy clustering through formulation of the community detection problem as an optimization problem based on the PCM algorithm. Therefore, the main role of PCMSA is finding the best degree of membership to assign nodes to clusters so that clusters that are the most consistent with the discussed criteria can be achieved.

In order to detect communities that satisfy the above criteria, a minimum optimization problem based on PCM is formulated for the community detection problem of complex networks by considering both datasets related to node attributes and link structure. Then, the solution to this optimization problem is presented (Fig. 1).

4.1 Proposed model

This section deals with the proposed “fuzzy community detection model (PCMSA)” in an attempt to detect overlapping communities with regard to structural and attribute similarities in the complex networks. In the PCMSA, we want to cluster nodes on the basis of analyzing the semantic of social networks data considering the two important homophily and weak ties theorems (Kadushin 2004).

Assume that $G(N, L)$ is a graph in which N indicates a number of nodes ($N = \{1, 2, \dots, n\}$) and L is indicative of some links ($L = \{1, 2, \dots, l\}$). The relevant terminologies are described as follows:

n	Number of nodes
l	Number of edges
e	Number of attributes
t	Number of repetitions (periods)
$m \in [1, \infty)$	Fuzzy parameter (weighting exponent called the fuzzifier)
x_i	i th object
c	Number of clusters
g_i	Set of nodes in cluster i
v_i	i th cluster centroid
u_{ik}	Membership degree of k th node to i th cluster
Δ_i	Density of cluster i
v_i^0	The initial center of cluster i
δ_i	Entropy of cluster i
χ_j	The set of values of attribute \mathcal{Y}_j

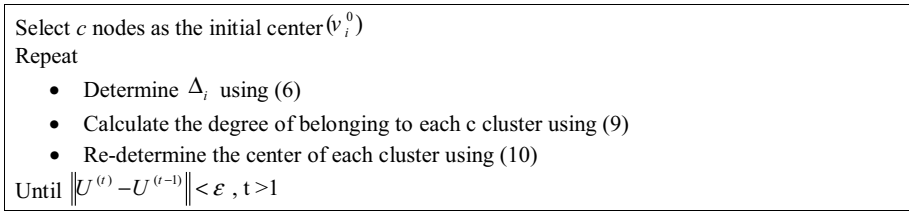


Fig. 1 Algorithm 1: Center-based fuzzy community detection algorithm based on link structure

4.1.1 Structural similarity

The center-based community detection objective function considering link structure is formulated as:

$$J = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m D_{ik} \tag{3}$$

In this formulation, D_{ik} indicates the distance of node k from the center of cluster i (v_i) that can be calculated with the following formulation:

$$D_{ik} = \begin{cases} 0 & \text{if there is a link between nodes } v_i \text{ and } k \\ \sum_{j=1}^n |a_{vij} - a_{jk}| & \text{otherwise} \end{cases} \tag{4}$$

The transitivity theorem (Kadushin 2004) (if node A is connected to node B and node B is also connected to node C, most probably node A will be connected to node C) of social networks is used to define the structural distance between nodes. Therefore, for two connected nodes (first case) the special case is considered and the second case is based on transitivity theorem. In this equation, each a_{ij} ($i = 1, \dots, n, j = 1, \dots, n$) is indicative of the entry in the i th row and j th column of the adjacency matrix denoted by A . The entries in the adjacency matrix (a_{ij}) indicate adjacent nodes. In this matrix, if nodes i and j are adjacent, then $a_{ij} = 1$, and if nodes i and j are not adjacent, $a_{ij} = 0$ (Wasserman and Faust 1994).

Our article focuses on undirected graphs, and the links are not signed or valued. Therefore, if $a_{ij} = 1$ then $a_{ji} = 1$, thus, the matrix is symmetric (Wasserman and Faust 1994).

In this article, first a community detection model based on PCM is proposed to cluster nodes considering their link information where nodes represent objects, and links indicate the relationship among objects. Therefore, each cluster is presented as a number of interconnected objects which are not connected to objects out of the group (Wasserman and Faust 1994). The objective function that has been proposed is formulated as:

$$\min J_m(u, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m D_{ik} + \sum_{i=1}^c \Delta_i \sum_{k=1}^n (1 - u_{ik})^m$$

The first part of this objective function lessens the distance from cluster centers to the extent that is possible considering the data link structure (D_{ik}). The second term causes u_{ik} to become as large as possible, and in this way avoid the trivial solution (Malek et al.

2015). Δ_i equals the proportion of existing links in a cluster to all the links that can be presented in this cluster ($|L_i|$), which has become maximized.

$$\Delta_i = \frac{|\{(p, q) | p, q \in g_i, (p, q) \in L_i\}|}{|L_i|}$$

Now, the first proposed model identifies a new fuzzy clustering model that is center-based in order to identify communities that overlap in complex networks. This model is defined on the basis of the PCM clustering model and detects overlapping communities on the basis of the link structure. The defined model is formulated as:

$$\min J_m(u, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m D_{ik} + \sum_{i=1}^c \Delta_i \sum_{k=1}^n (1 - u_{ik})^m \tag{7}$$

subject to

$$\left[\begin{array}{l} u_{ik} \in [0, 1], 1 \leq i \leq c, 1 \leq k \leq n \\ \max_i (u_{ik}) > 0, 1 \leq k \leq n \\ 0 < \sum_{k=1}^n u_{ik} < n, 1 \leq i \leq c \end{array} \right. \tag{8}$$

Theorem 1. Assume that $G(N, L)$ is a graph in which N indicates the set of nodes ($N = \{1, 2, \dots, n\}$) and L is indicative of some links ($L = \{1, 2, \dots, l\}$). In our model, U_k indicates the k th column of U ; that is, $U_k = \{u_{1k}, \dots, u_{ck}\}, 1 \leq k \leq n$. Then, U could be a global minimum for $J_m(U, V)$ only if the updating fuzzy membership value is:

$$u_{ik} = \left(1 + \left(\frac{D_{ik}}{\Delta_i} \right)^{1/(m-1)} \right)^{-1} \tag{9}$$

and the center of cluster is as follows:

$$v_i^* = \arg \min_{v_i \in [1, n]} \left(\sum_{k=1}^n \sum_{j=1}^n u_{ik}^m D_{ik} \right) \tag{10}$$

Theorem 1 will be proved in ‘‘Appendix 1’’.

Now, consider the algorithm of fuzzy community detection to identify overlapping communities based on link structure.

The Δ_i moves to 0 if there are no links in a cluster, and it moves to 1 if all the links exist. The higher the value of Δ_i , the more connections exist between nodes, leading to a denser cluster.

The most important step in a clustering that is center-based is choosing the proper initial central node. An approach that is more common is random choice of initial centers, however, the outcomes are often weak (Malek et al. 2015). The Nodal degree, the number of lines incident with the node in the graph, can be a good criterion to choose the initial centers. This can be achieved by summing with regard to elements in the adjacency matrix as follows (Wasserman and Faust 1994):

$$v_i^0 = \max_{j' \in \{1, \dots, n\}} \sum_{j=1}^n a_{ij'} \tag{11}$$

In the proposed clustering, clusters have an intra-cluster structure that is cohesive. A favorable cluster should have an intra-cluster structure that is cohesive and has homogeneous vertex attributes. Therefore, in this paper, re-clustering is proposed to re-cluster communities considering a threshold on the basis of the homophily theorem in social networks in which, if two people have characteristics that match in a proportion greater than expected in the population from which they are drawn or the network of which they are apart, then they are more likely to be connected. The converse is also true: if two people are connected, then they are more likely to have common characteristics or attributes (Kadushin 2004).

Moreover, “The strength of weak ties”, which has attracted a lot of research attention, is an article that has been presented by Mark Granovetter (Granovetter 1977). Weak ties concentrate on holes in the network (Kadushin 2004). Our acquaintances (weak ties) may have less relationship with us than our close friends (strong ties). Thus, if we have a set of people with their acquaintances in whom many of the possible ties are absent, their network will constitute a low-density network (Kadushin 2004).

Weak ties cause the information to easily flow from remote parts of a network. Objects that have few weak ties are deprived of information from remote parts of a network and only get provincial news and information from their close friends. Compared to the strong ties, weak ties may serve as bridges between network segments. Thus, social systems that do not have weak ties are incoherent and will be fragmented as weak ties helping to integrate social systems. Without considering weak ties, new ideas will spread slowly, and scientific efforts will not achieve their success (Kadushin 2004). Due to the importance of weak ties they are considered in this paper with the proposed model that detect communities based on two sources of data, structural and attribute similarities.

Therefore, the re-clustering should be employed by the following measurement according to the above theorems.

This operation works by using a pair wise similarity measure to find groups of clusters that could benefit from re-clustering their component nodes and edges. In order to find groups of clusters needing to be re-clustered, the similarity of each pair of clusters based on their nodes’ attributes is found. The similarity measure for two especial communities (clusters) i and j is defined as:

$$p_{i,j} = 1 - \frac{\sum_{q \in g_i} \sum_{q' \in g_j} \|x_q - x_{q'}\|}{\sum_{w=1}^{c-1} \sum_{v=w+1}^c \sum_{h \in g_w} \sum_{h' \in g_v} \|x_h - x_{h'}\|}$$

Equation (12) indicates the proposed similarity measure that calculates the percentage of similarity levels of cluster i and cluster j considering their nodes’ attributes. For especial case $c = 2$, the re-clustering is done and it is then decided by the proposed validation index whether the re-clustering is good or not.

By using this similarity measure, groups of clusters that could benefit from re-clustering their component nodes and edges can be defined. If $p_{i,j} > B$, (B identifies by try and error) then the re-clustering algorithm is run to re-cluster all the nodes and edges in cluster i and cluster j . This is similar to the “The rich get richer and the poor get poorer” phrase as the

nodes in the cluster with the structural algorithm that possess similar attributes become denser with the re-clustering algorithm.

It is worth mentioning that B is not limited and depends on the considering graph. For each clustering, the value of B is defined by try and error and the scenario with the highest validity index. According to the experiments, the value of B is obtained smaller in graphs with more weak ties. As a result our method is flexible in which it is considered for each graph according to its similarities. If the attribute similarity is greater than B that is identified by the validity index, the re-clustering is applied. This is while some graphs don't need to be re-clustered according to the weak ties and homophily theorems.

As it is mentioned before the value of B depends on the graph and its attribute similarities. Therefore the minimum and maximum values of $p_{i,j \in [1,c]}$ are obtained for the graph. B is searched in this bound ($[min, max]$) and the value with the highest validity index in clustering is selected.

4.1.2 Attribute similarity (Re-clustering)

The proposed re-clustering objective function is defined as follows:

$$\min J_m(u, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik} - \sum_{i=1}^c \delta_i \sum_{k=1}^n (1 - u_{ik})^m \tag{13}$$

In this equation, δ_i is measured as:

$$\delta_{i \in [1,c]} = entropy(C_i) = \sum_{j=1}^e \frac{|g_i|}{n} entropy(\gamma_j, C_i) \tag{14}$$

$$entropy(\gamma_j, C_i) = - \sum_{g \in \chi_j} p_{ijg} * \ln p_{ijg} + (1 - p_{ijg}) * \ln(1 - p_{ijg})$$

In this equation, p_{ijg} refers to the percentage of vertices existing in cluster i with value γ_{jg} on attribute γ_j . δ_i measures the weighted entropy from all attributes over c clusters. Moreover, for continues values of an attribute, the fuzzy membership function of that attribute is used and then ($\alpha - cut$) in fuzzy sets (Mendel and Mendel 2017) is applied to create a finite set of values.

The parameterization of d_{ik} should be specified. Referring to the Gustafson and Kessel's definition, d_{ik} can be obtained as follows (Gustafson and Kessel 1978):

$$d_{ik}(\Omega_i) = (x_k - v_i)^T H_i (x_k - v_i), 1 \leq i \leq c \tag{15}$$

This form of d_{ik} indicates the norm metric of an inner product with H_i symmetric and positive-definite matrix. Note that we take $\Omega_i = \{v_i, H_i\}$ and J is linear in H_i inducing a singular problem (Gustafson and Kessel 1978). Gustafson and Kessel restricted the determinant $|H_i|$ of matrix H_i in order not to allow the metric to grow without bound (Gustafson and Kessel 1978).

Now the proposed re-clustering model is as follows:

$$\min J_m(u, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik} - \sum_{i=1}^c \delta_i \sum_{k=1}^n (1 - u_{ik})^m \tag{16}$$

subject to

$$\begin{cases} |H_i| = v_i, v_i > 0 \\ u_{ik} \in [0, 1], 1 \leq i \leq c, 1 \leq k \leq n \\ \max_i(u_{ik}) > 0, 1 \leq k \leq n \\ 0 < \sum_{k=1}^n u_{ik} < n, 1 \leq i \leq c \end{cases} \tag{17}$$

Constraint $|H_i| = v_i$ guarantees that H_i is positive-definite (Bezdek et al. 1999). Now, the augmented objective function is defined as:

$$\begin{aligned} \min J_m(u, \Omega, \lambda) = & \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik}(\Omega_i) - \sum_{i=1}^c \delta_i \sum_{k=1}^n (1 - u_{ik})^m \\ & + \sum_{i=1}^c \lambda_i [|H_i| - v_i] \end{aligned} \tag{18}$$

where $\{\lambda_i\}$ is a set of Lagrange multipliers.

Theorem 2. U could be a global minimum for $J_m(U, V)$ only if the updating fuzzy membership value is:

$$u_{ik}^* = \left(1 + \left(\frac{d_{ik}(\Omega_i)}{-\delta_i} \right)^{1/(m-1)} \right)^{-1} \tag{19}$$

and the cluster center is:

$$v_i^* = \frac{\sum_{k=1}^n u_{ik}^m x_k}{\sum_{k=1}^n u_{ik}^m} \tag{20}$$

and finally,

$$H_i^{*-1} = \frac{1}{\lambda_i |H_i^*|} \sum_{i=1}^c u_{ik}^m (x_k - v_i^*)(x_k - v_i^*)^T \tag{21}$$

Theorem 2 will be proved in Appendix 1.

Now, FC_i is the fuzzy covariance matrix that can be defined as follows (Gustafson and Kessel 1978):

$$FC_i = \frac{\sum_{k=1}^n u_{ik}^m (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^n u_{ik}^m}; m > 1 \tag{22}$$

Then, using (22) and $|H_i| = v_i$ in (21), H_i^{*-1} gives:

$$H_i^{*-1} = \left[\frac{1}{v_i |FC_i|} \right]^{1/\alpha} FC_i \tag{23}$$

In which α is the feature space dimension.

The previous discussion and then re-clustering algorithm induce the following proposed algorithm for community detection considering both sources of data related to topological structure and vertex properties. Figure 2 illustrates the proposed algorithm.

The essential condition for converging the algorithm suggested in Fig. 2 is met when:

$$\lim_{t \rightarrow \infty} \|U^{(t)} - U^{(t-1)}\| = 0$$

The reason for this condition as well as the proposed algorithm convergence is to be offered in Appendix 2.

In the PCMSA algorithm, steps 1–4 detect communities on the basis of the structural similarities and steps 5–11 detect the last communities based on the attribute similarities.

For the proposed algorithm the complexity is $O(c * n * t + c * (c - 1) * t * e * n_g)$ in which c , n , t , e , and n_g indicate the number of communities, the number of nodes, the number of iterations, the number of attributes, and the number of nodes in two re-clustering communities, respectively.

1. Select c nodes as initial center (v_i^0) , $i = 1, \dots, c$

Repeat

2. Determine Δ_i using (6)
3. Calculate degree of belonging to each c cluster using (9)
4. Re-determine the center of each cluster using (10)

Until centers do not change.

5. Calculate $P_{i,j \in \{1,c\}} (i < j)$ using (12) for each pair of clusters $(i < j)$. If $P_{i,j} > B$, then re-cluster all the nodes in clusters i and j by the following procedure:

Repeat

6. Consider $\omega = i, j$
7. Compute $\Omega_\omega^{(0)} = \{v_\omega^0, FC_\omega^0\}$. The center for each cluster from the last step is considered as the initial center. Calculate FC_ω^0 using (22)
8. Determine δ_ω using (14)
9. Compute $\{d_{\omega k}(\Omega_\omega)\}$ using (15)
10. Calculate $u_{\omega k}$ using (19)
11. Re-determine Ω_ω using (20), (22) and (23)

Until $\|U^{(t)} - U^{(t-1)}\| < \epsilon, t > 1$

Fig. 2 PCMSA Algorithm: Center-based fuzzy community detection algorithm to identify overlapping communities based on link structure and nodes' attributes

5 Clustering validation index

The appropriate criteria for evaluating the performance of clustering process can be the number of links in the community and those outside the community, which are the base of most community definitions (Zarandi et al. 2010). Suppose a sub-graph G_i of a graph G in which $|G_i| = g_i$ and $|G| = g$. The internal and external degree of sub-graph G_i can be defined as the number of links in the sub-graph connecting nodes to each other and that of links connecting nodes inside the sub-graph to the remainder of the graph, respectively. The quality of clusters with two measures of density, intra-cluster density and inter-cluster density is evaluated. The ratio between the number of internal links of cluster C_i and that of possible internal edges (L_i) is called the intra-cluster density.

$$\Delta_i = \frac{|\{(p, q) | p, q \in C_i, (p, q) \in L_i\}|}{|L_i|} \tag{25}$$

In addition, the inter-cluster density can be defined as the ratio between the number of edges from the nodes of C_i to the remainder of the graph and that of possible inter-cluster edges.

$$\Delta_i^{ext} = \frac{|\{(p, q) | p \in C_i, q \notin C_i, (p, q) \in L'_i\}|}{|L'_i|} \tag{26}$$

In this equation, L'_i indicates edges between nodes inside cluster C_i to the remainder of the graph.

For C_i to be a community with homogeneous attributes, the homogeneity is expected to be appreciably maximum inside a community and minimum between the communities. This is defined as the separation measure. The separation measure is defined as follows:

$$hom_i = \frac{\sum_{j=1}^c \sum_{z \in g_i} \sum_{z' \in g_j} \|x_z - x_{z'}\|}{\sum_{w=1}^{c-1} \sum_{h=w+1}^c \sum_{p \in g_w} \sum_{p' \in g_h} \|x_p - x_{p'}\|}$$

The suggested index considers two criteria: compactness and separation. The compactness measure is determined based on inter- and intra-density of communities $\Delta dense_i = \Delta_i - \Delta_i^{ext}$.

As a result, a desired community is one with a maximum level of compactness and a larger level of separation. Therefore searching for the best trade-off between density and homogeneity is the goal of our algorithm, and one method to do that is maximization:

$$\Lambda_i = \Delta dense_i * hom_i \tag{28}$$

It is expected that by maximizing Λ_i , C_i can be a community. By considering this criterion for each community, the validity index Λ is determined by Eq. (29):

$$\Lambda = \Delta dense * hom \tag{29}$$

In this equation $\Delta dense$ is the average of $\Delta dense_i (i = 1, \dots, c)$ and hom is the average of $hom_i (i = 1, \dots, c)$. Equation (29) is considered the criterion to assess the performance of the proposed community detection and determines the most favorable number of clusters.

6 Experimental results

In this section, the performance of the proposed model is tested in several artificial and large scale real networks.

Example 1 In the first step, a simple dataset with 10 nodes is considered as shown in Fig. 3a, which indicates a co-worker graph where nodes show workers, and edges indicate relationships between them. Each number shows a worker ID. Moreover, two attributes describe features of a node. The first letter indicates gender (Male “M” or Female “F”) and the second letter indicates where they live (Montreal “M” or Toronto “T”). As shown in Fig. 3a, workers 4, 6 and 7 have the same properties, worker 3 is male and lives in Montreal and the others have the same properties. Suppose the cluster number is $c = 2$. As it can be seen, depending on the clustering criteria, several clustering ways are obtained:

Figure 3b indicates clusters based on structure similarity that only considers relationships between co-workers and ignores their attributes where each cluster has varieties of genders and places. In this clustering method, co-workers in clusters are closely connected and coherent.

Figure 3c indicates clusters considering attribute similarity. This means that clusters have co-workers with the same properties as much as possible; therefore, the resulted clusters are homogenous, however, the vertex connectivity may not be considered.

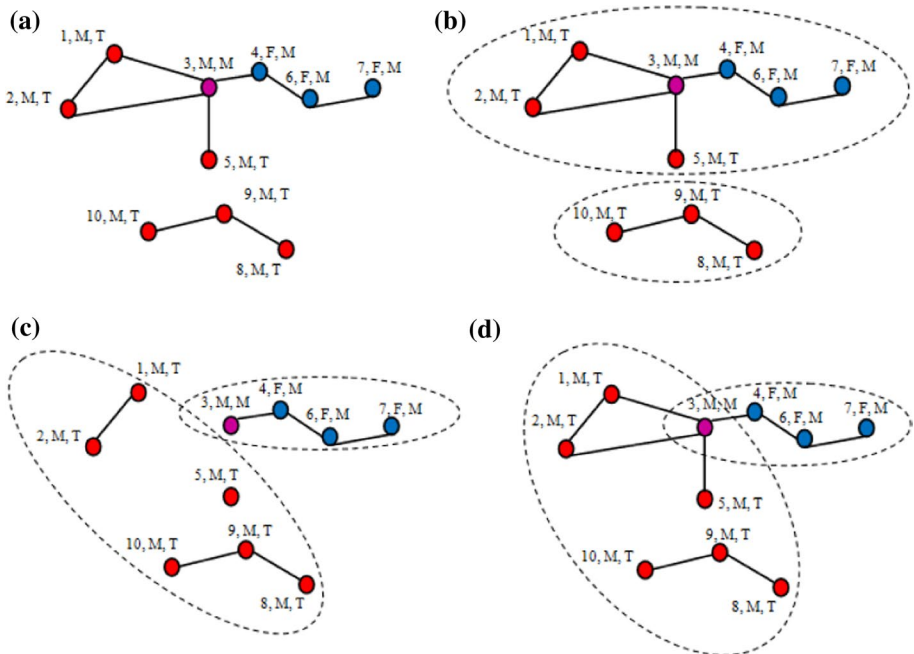


Fig. 3 a co-worker Graph. b Structure-based cluster. c Attribute-based Cluster. s Proposed clustering (PCMSA)

Figure 3d illustrates the result of the proposed clustering algorithm considering both sources of data related to topology structure and the node’s attribute. The workers in one cluster are closely connected and have the same property so that the coherent and homogenous clusters are resulted. In addition, as discussed earlier in Sect. 4, the proposed algorithm detects overlap communities where node 3 is assigned to both communities.

In this section, the function of the proposed model is tested in “ m ” (fuzzy parameter) and v ($|H|$ in (17)). By considering the proposed validity index, to find the best value for fuzzy parameter (m), the values of $1 - \Lambda$ for different values of m (fuzzy parameter) and c (number of communities) are obtained as follows. Thus by minimizing $1 - \Lambda$, the best value for “ m ” can be 2.5.

The membership values generated by the proposed model with different values for v are shown in Table 1. In Fig. 3d the left cluster is cluster 1 and the other is cluster 2. By increasing the value of “ v ” only, the variations in the shape of clusters can become greater without any limitation that leads to the generation of clusters without homogenous attributes or coherent structures depending on the data. The historical membership functions for two critical points are shown in Table 1. Note that point 3 is strongly related to cluster 1 for $v = 0.5$ and $v = 1$, but it starts to form correctly from $v = 1.5$. Moreover, node 4 starts to form incorrectly from $v = 2$. Therefore, v can be set to 1.5, according to Table 1, causing reasonable and desired membership functions. Moreover, there is a weak tie between node 3 and node 8 in which they are not coworker. But they are assigned to the same cluster considering PCMSA algorithm based on their similar attributes. The same results can be seen for node 3 compared to nodes 9 and 10 (Fig. 4).

Note that in this paper, v_i is considered the same for all clusters.

Example 2 The second sample having 11 nodes is depicted in Fig. 5 (Zhou et al. 2009). There are two communities in this example. This graph contains vertices that indicate authors, and the edges between them indicate the co-author relationship between them. In addition to an ID number for every author, the associated topic related to an author is given by describing its attribute. Authors 1–7 work on XML, authors 9–11 work on Skyline and author 8 works on both. Obtained clusters with different clustering ways are shown in Fig. 5b–e. Figure 5d is a clustering algorithm, named SA-clustering, proposed by Zhou et al. (2009). Although it considers both relationships between nodes and their attributes, it cannot detect overlap communities as well as the PCMSA algorithm. Note that the proposed community detection PCMSA detects overlap communities in addition to considering both sources of data.

Table 1 Membership functions history for different values of v

v	u_{14}	u_{24}	u_{13}	u_{23}
0.5	0	0.92	0.71	0
1	0	0.91	0.71	0
1.5	0	0.91	0.69	0.3
2	0.12	0.90	0.69	0.3
2.5	0.12	0.88	0.69	0.32
3	0.13	0.88	0.68	0.32

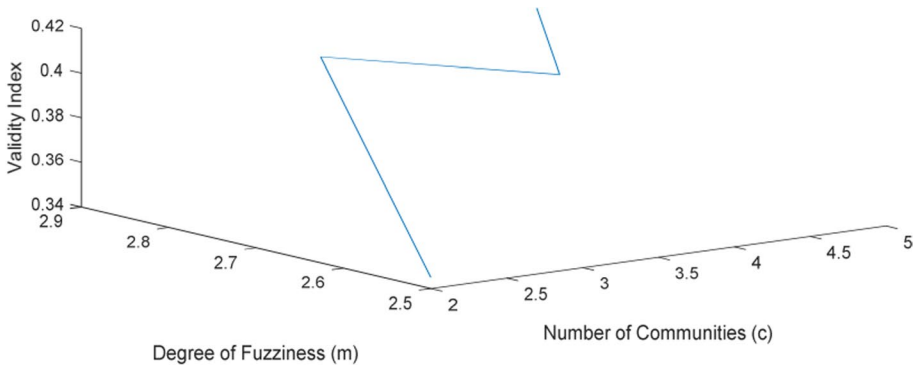


Fig. 4 Proposed validity index for different degrees of fuzziness (Example 1)

Figure 6 indicates membership values generated by PCMSA for different values of m . Again By considering the proposed validity index, to find the best value for fuzzy parameter (m), the values of $1 - \Lambda$ for different values of m (fuzzy parameter) and c (number of communities) are obtained as follows. Thus by minimizing $1 - \Lambda$, the best value for “ m ” can be 2.

Table 2 shows the membership values generated by the proposed model with different values for v . Note that the membership function of nodes 3, 4 and 8 assign correctly until $v = 1$. Therefore, v can be set to 1 according to Table 2. In Fig. 5e, the left cluster is named cluster 1 and the other is cluster 2.

Moreover, there is a weak tie between node 8 and node 5 in which they are not coauthor. But they are assigned to the same cluster considering PCMSA algorithm based on their similar attributes. The same results can be seen for node 8 compared to nodes 6 and 7.

Example 3 The third example to evaluate the proposed model is a partial Facebook network. Facebook is an American online social medium and social networking service company. Sample three is part of this social network from Northwestern University (Traud et al. 2011) (The nodes are selected randomly from the original dataset). This network has 256 nodes with 401 links as shown in Fig. 7a. A total of six communities have been identified in this dataset (Traud et al. 2011). Moreover, the other sizes of the network have also evaluated in Fig. 9 that the largest of which has 3452788 nodes, and 76849635 edges.

In addition, each node has 7 attributes which indicate its properties. These attributes include student/faculty status flag, gender, major, second major/minor (if applicable), dorm/house, year, and high school.

As it is mentioned before, parameter B is defined by a procedure which is explained in Sect. 4. In this graph p_{ij}^{\min} and p_{ij}^{\max} are obtained 0.01 and 0.78, respectively. B is searched in this bound $[0.01, 0.78]$ and $B = 0.2$ resulting in a best validity index.

The distinction between inter-cluster and intra-cluster density is explained in Table 3. The smaller the cluster number is, the lower the difference between intra-cluster and inter-cluster distance is. Additionally, the value of homogeneity is obtained according to (27). As the results show, the smaller the cluster number is, the lower the homogeneity is. The trade-off between $\Delta dense$ and hom by using (29) indicates that the first local maximum occurs in $c = 6$.

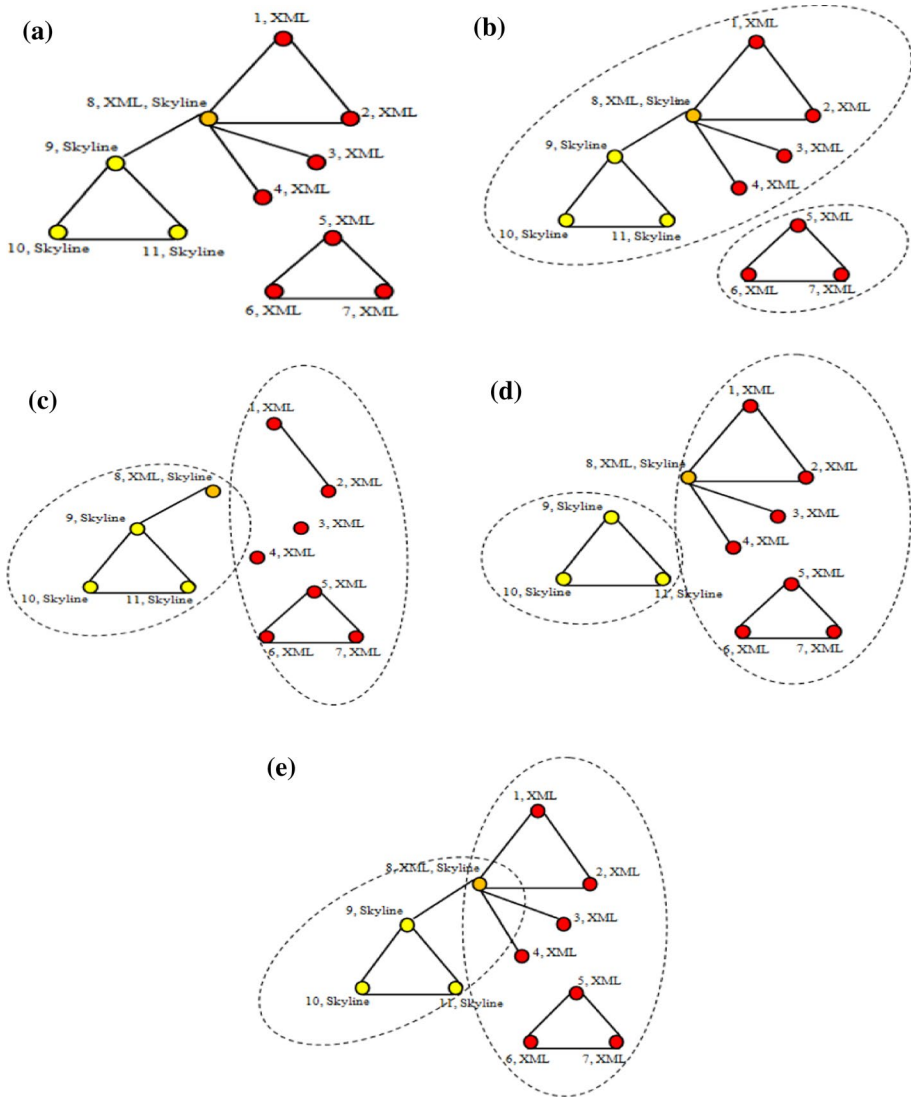


Fig. 5 a Co-author graph. b Structure-based cluster. c Attribute-based Cluster. d SA-cluster Approach. e Proposed clustering (PCMSA)

In this part, two validity indices (*NMI* and *Accuracy* (Hu and Chan 2016)) in addition to the proposed validity index (Λ) are tested on the dataset in which the optimum number of communities c for each index is shown in Table 7. The reason why we choose these two metrics is that they are widely used to evaluate the community detection algorithms. *NMI* and *Accuracy* metrics (Hu and Chan 2016) are defined in Appendix 3.

The parameters of the PCMSA are set to $\epsilon = 0.001$ and $m = 2.5$. As discussed in Appendix 3, *NMI* and *Accuracy* obtain optimal value with maximization. Table 4 shows the optimal values of the validity indices for $c = 2, 3, \dots, c_{\max} = \sqrt{n} = 16$. Among the indices, only

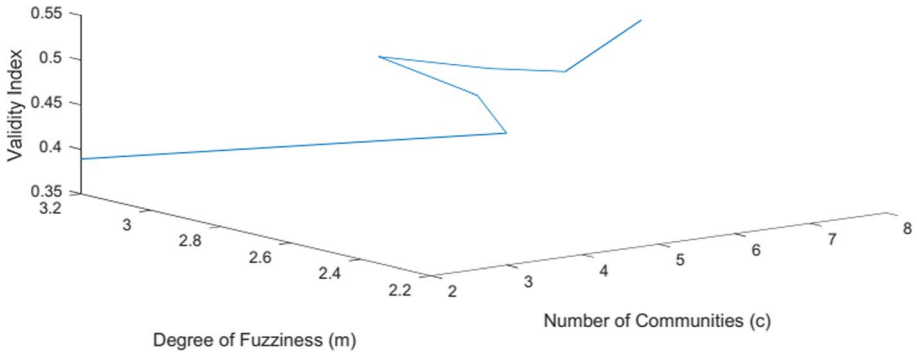


Fig. 6 SC index for different degrees of fuzziness (Example 2)

Table 2 Membership functions history for different values of v

v	u_{13}	u_{23}	u_{14}	u_{24}	u_{18}	u_{28}
0.5	0	1	0	0.94	0.28	0
0.75	0	1	0	0.94	0.30	0
1	0	1	0	0.92	0.30	0.25
1.5	0.12	1	0.23	0.92	0.30	0.29
2	0.12	1	0.25	0.93	0.30	0.30

Fig. 7 Facebook network

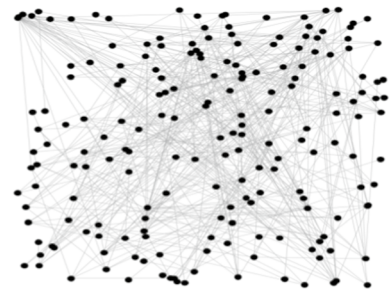


Table 3 Validity Index

c	Δ_{dense}	hom	Λ
2	0.12	0.23	0.02
3	0.19	0.24	0.04
4	0.23	0.32	0.07
5	0.32	0.48	0.15
6	0.46	0.59	0.28
7	0.34	0.58	0.20
8	0.35	0.63	0.23
9	0.36	0.69	0.25
10	0.36	0.70	0.25

Table 4 Cluster validity values for $c = 2, \dots, c_{\max} = \sqrt{n} = 16$

c	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Λ	0.02	0.04	0.07	0.15	0.28	0.20	0.23	0.25	0.25	0.27	0.24	0.20	0.19	0.21	0.15
Accuracy	0.55	0.57	0.67	0.69	0.74	0.72	0.71	0.62	0.59	0.60	0.64	0.58	0.52	0.51	0.51
NMI	0.11	0.18	0.23	0.20	0.19	0.12	0.11	0.11	0.15	0.13	0.10	0.11	0.12	0.14	0.12

Accuracy and Λ correctly result in the six clusters and NMI does not recognize the optimal c value and results in four clusters as the optimal number.

The objective function of the PCMSA depends on the weighting exponent $m \in [1, \infty)$. When a validity index is insensitive to change in m , it can be said that it is reliable. Therefore, here, all validity indices are considered for various values of both c and m . The results of reliability of each index by changing m are reported, and the results are listed in Tables 5, 6, 7.

Tables 5, 6 show the results of three validity indices for $c = 2, \dots, 8$ for weighting exponent $m = 1.5$ and 2, respectively. In Table 5, for $m = 1.5$, only Λ results in the optimal value for c ($c = 6$). In Table 6, for $m = 2$, Λ and NMI are the only indices that result in the optimal $c = 6$.

Table 7 lists the optimal number of clusters for all validity indices in varieties values of m . As shown in Table 7, Λ demonstrates a better result, and is the least sensitive to change in m . The bold numbers in the Table 7 show the correct number of communities and the bold numbers in the Tables 4, 5, and 6 show the best validity values for each validity index and different values of m .

From the tests, the suggested validity index Λ is the only index which obtains the optimal c in all of experiments; moreover, when the different values of m are considered again, Λ has a better performance. Therefore, the proposed index Λ is used to evaluate PCMSA.

Table 5 Cluster validity values for $c = 2, \dots, 8; m = 1.5$

C	2	3	4	5	6	7	8
Λ	0.17	0.04	0.09	0.14	0.24	0.21	0.19
Accuracy	0.42	0.52	0.32	0.56	0.42	0.47	0.50
NMI	0.24	0.23	0.12	0.27	0.14	0.22	0.25

Table 6 Cluster validity values for $c = 2, \dots, 8; m = 2$

C	2	3	4	5	6	7	8
Λ	0.03	0.13	0.04	0.17	0.19	0.11	0.16
Accuracy	0.43	0.33	0.23	0.59	0.42	0.45	0.52
NMI	0.32	0.28	0.15	0.18	0.34	0.23	0.33

Table 7 Values of c preferred by validity indices with $c = 2, \dots, 8$ and different values of m

m	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
Λ	6	5	6	6	6	6	6	6	6	6	6
Accuracy	5	4	4	4	7	5	7	6	6	5	6
NMI	5	5	6	6	5,4	6	5	6	4	4	4

The result of community detection by algorithm 1 (structure-based community detection) is illustrated in Fig. 8a. It clusters nodes based on their relationship in the university and does not pay attention to their properties compared to Fig. 8c which clusters nodes based on the PCMSA algorithm considering both the nodes' structure and properties. As shown in Fig. 8c, nodes in a community do not necessarily only have a relationship with each other, but they also have the same properties according to the dataset. Moreover, the result of community detection by SA-cluster algorithm (Zhou et al. 2009) is shown in Fig. 8b. In Fig. 8, the exclusive members of 6 clusters are shown using colors: yellow, dark blue, pink, green, light blue, and red. However, the black nodes are shared between each pair of clusters. The function of the suggested model (PCMSA) is compared with

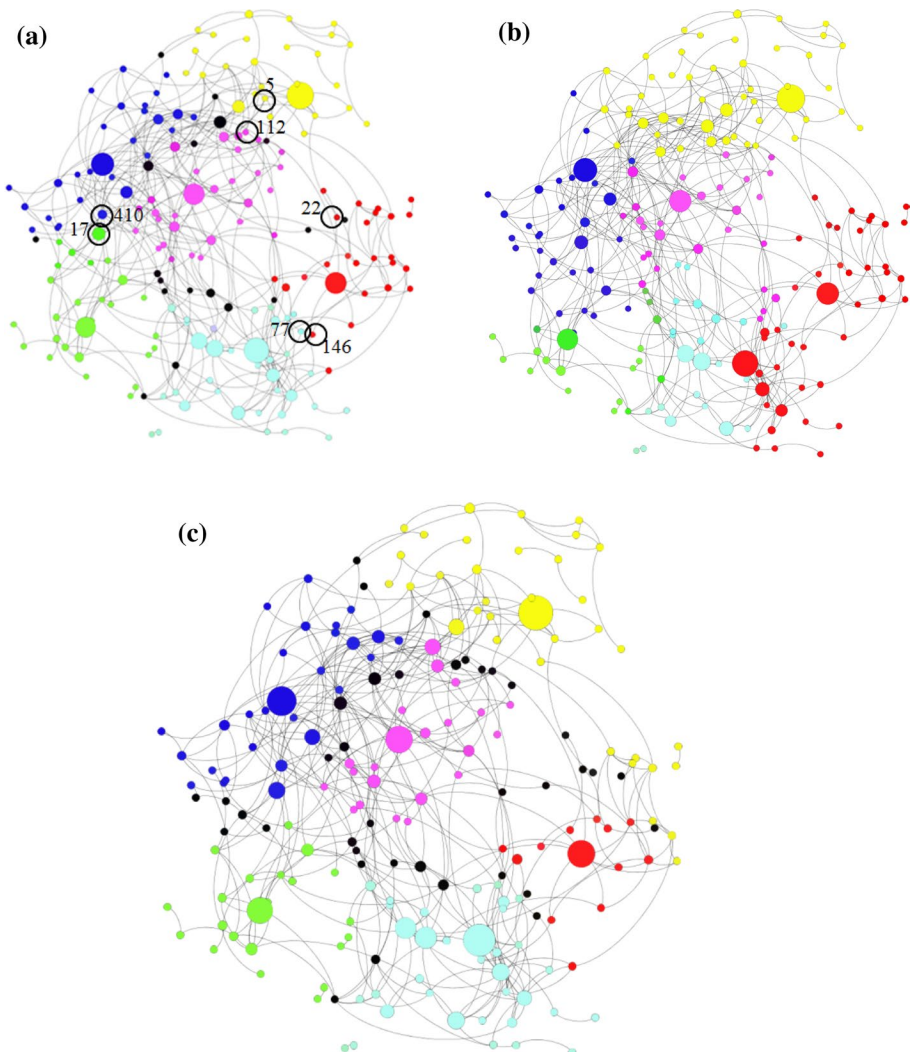


Fig. 8 **a** Community detection based on node's structure. **b** SA-clustering. **c** PCMSA Algorithm

algorithm 1 and SA-cluster using the validation index which is equal to $\Lambda_{algorithm1} = 0.14$ and $\Lambda_{SA-cluster} = 0.11$ for algorithm 1 and SA-cluster, respectively, using (29). However, the validity index for PCMSA is obtained $\Lambda_{PCMSA} = 0.28$. Hence, the function of the suggested model, PCMSA, is superior to that of other models.

In this experiment, the efficiency of different clustering algorithms is compared. Figure 9 indicates the clustering time on different number of nodes. The following observations on the runtime costs of different methods are made. First, PCMSA is usually 1.9–2.4 times slower than Algorithm 1, as it iteratively computes the re-clustering procedure. Although PCMSA is more expensive, but the iterative re-clustering improves the clustering quality a lot, as it is demonstrated in Fig. 8 and Table 15. According to our analysis CODICIL (Ruan et al. 2013), SA-cluster (Zhou et al. 2009), and FCAN (Hu and Chan 2016) algorithms are slower than the proposed algorithm (PCMSA). K-SNAP (Tian et al. 2008) is a little faster than PCMSA in three samples, but the quality of PCMSA is more than K-SNAP according to Table 15. Figure 9 suggests our method has promising scalability and computational complexity for analysis on even large social networks.

To understand more about the unique capabilities of the PCMSA, we choose this network and looked in to its communities in details. To investigate how content information may affect cluster determination, we have also considered the attributes of nodes. Nodes 17 and 410 in Fig. 8a are assigned to different communities because there is not any edge between them, while with PCMSA algorithm (Fig. 8c) they are assigned to the same community according to their similar attributes(nodes 17 and 410 attributes: student, male, chemical engineering, -, dorm, 1980, chio). This confirms the existence of a weak tie between nodes 17 and 410 which they do not have any link (strong link) to each other but they are assigned to the same community considering their attribute similarities. This result is also clear for node 22, so that, it is shared between two communities with PCMSA according to its attribute similarities with two communities, while in Fig. 8a it is assigned to one community. Moreover, nodes 5 and 112 are assigned to different communities in Fig. 8a, but in Fig. 8c, node 112 is shared between two communities according to PCMSA algorithm (node 5 attributes: faculty, female, management, art, house, 1976, berth and node 112 attributes: faculty, male, management,-, house, 1975, berth). These results can also be seen for nodes 77 and 146.This, again, indicates the importance of considering content information in addition to the link structure when determining communities.

Example 4s The fourth sample is a social network that has received particular attention in the context of community detection known as a Political Blogs Dataset, which could be downloaded from <http://www-personal.umich.edu/mejn/netdata/> (Adamic and Glance

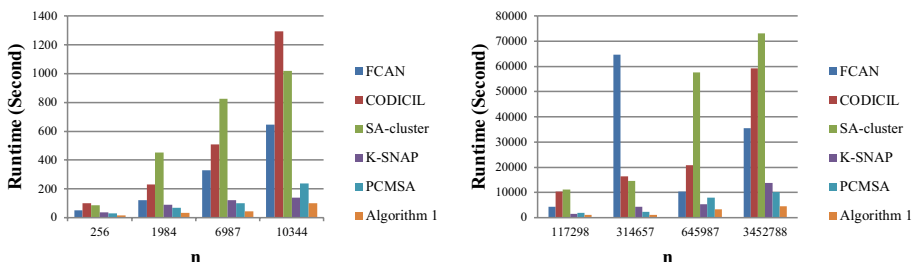


Fig. 9 Clustering efficiency

2004). 1490 weblogs politics are considered as vertices with 19025 hyperlinks between these weblogs as shown in Fig. 10a. These nodes are distributed in two clusters (Adamic and Glance 2004). Moreover, the other size of the network has also considered in Fig. 12 which has 6577 nodes, and 87345 edges. Political leaning as either liberal or conservative describes an attribute for each blog in the dataset. Several papers use this network for evaluating their cluster analysis. The average degree of this network is 12.768 and the average density is 0.009.

Fig.10 Political blogs dataset

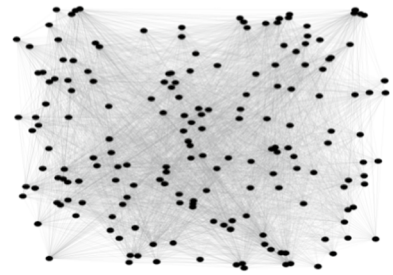


Fig. 11 difference in density ($\Delta dense$)

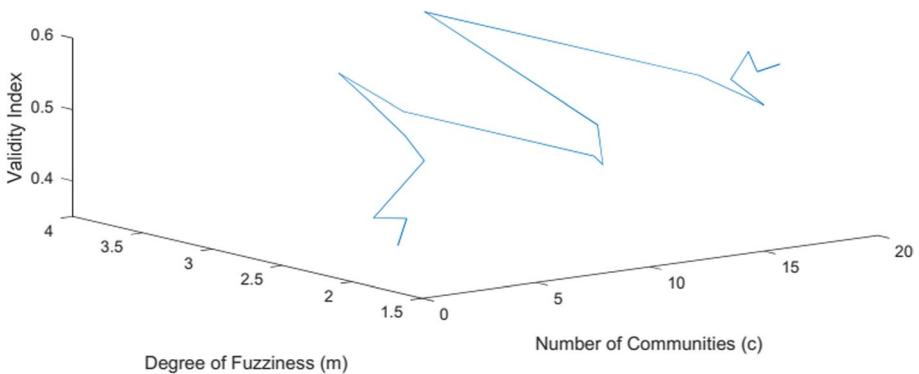
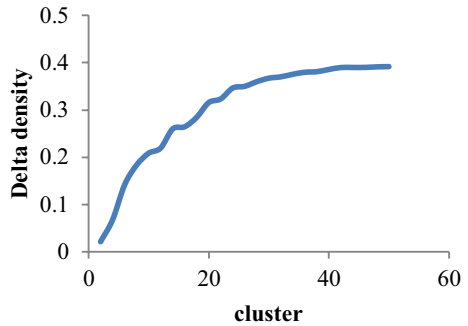


Fig. 12 Validity index for different degrees of fuzziness (Example 4)

The intra and inter-cluster density are calculated by using (36) and (37). Then, $\Delta dense$ is calculated and illustrated in Fig. 11. According to Fig. 11, with an increase in the number of clusters, $\Delta dense$ has an increasing trend with a lower rate for $c \geq 10$.

By considering the proposed validity index, to find the best value for fuzzy parameter (m), the values of $1 - \Lambda$ for different values of m (fuzzy parameter) and c (number of communities) are obtained as follows in Fig. 12. Thus by minimizing $1 - \Lambda$, the best value for “ m ” can be 2. Figure 12 indicates an average trending of Λ . The first local maximum (minimum of $1 - \Lambda$) occurs in $c = 2$. However, there are other favorable numbers of clusters like $c = 14, 28, 48, \dots$

In this part, two validity indices (*NMI* and *Accuracy* (Hu and Chan 2016)) in addition to the proposed validity index (Λ) are tested on the dataset in which the optimum number of communities c for each index is shown in Table 10.

The parameters of the PCMSA are set to $\epsilon = 0.001$ and $m = 2$. Table 8 shows the optimal values of the validity indices. It is worth noticing that $c_{max} = \sqrt{n} = 38$, and numbers after $c = 16$ are not presented in Table 8 due to the space consideration. Among the indices, only Λ correctly results in the two clusters and *Accuracy* and *NMI* do not recognize the optimal c value and results in 3 clusters as the optimal number.

The objective function of the PCMSA depends on the weighting exponent $m \in [1, \infty)$. When a validity index is insensitive to change in m , it can be said that it is reliable. Therefore, here, all validity indices are considered for various values of both c and m . The results of reliability of each index by changing m are reported, and the results are listed in Tables 9, 10, 11.

Tables 9, 10 show the results of three validity indices for $c = 2, \dots, 8$ for weighting exponent $m = 1.5$ and 2.5 , respectively. In Table 9, for $m = 1.5$, only Λ and *NMI* result

Table 8 cluster validity values for $c = 2, \dots, 16$

c	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Λ	0.59	0.38	0.35	0.29	0.28	0.25	0.25	0.22	0.23	0.25	0.20	0.19	0.16	0.18	0.16
<i>Accuracy</i>	0.64	0.67	0.55	0.54	0.52	0.53	0.50	0.50	0.48	0.44	0.45	0.43	0.37	0.32	0.32
<i>NMI</i>	0.25	0.23	0.29	0.26	0.19	0.23	0.23	0.15	0.11	0.10	0.17	0.17	0.15	0.14	0.12

Table 9 Cluster validity values for $c = 2, \dots, 8; m = 1.5$

c	2	3	4	5	6	7	8
Λ	0.33	0.31	0.29	0.29	0.19	0.17	0.18
<i>Accuracy</i>	0.54	0.57	0.55	0.48	0.49	0.39	0.44
<i>NMI</i>	0.31	0.25	0.29	0.28	0.24	0.18	0.17

Table 10 Cluster validity values for $C = 2, \dots, 8; m = 2.5$

c	2	3	4	5	6	7	8
Λ	0.48	0.41	0.39	0.35	0.31	0.29	0.25
<i>Accuracy</i>	0.71	0.70	0.65	0.67	0.64	0.62	0.60
<i>NMI</i>	0.39	0.33	0.31	0.32	0.29	0.23	0.26

Table 11 Values of c preferred by validity indices with $c = 2, \dots, 8$ and different values of m

m	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
Λ	2	2	2	2	2	2	2	2	2	2	2
Accuracy	3	2	3	3	2	3	4	4	3	4	2
NMI	2	3	2	3	2	4	3	4	3	3	2

in the optimal value for c ($c = 2$). In Table 10, for $m = 2.5$, all indices result in the optimal $c = 2$.

Table 11 lists the optimal number of clusters for all validity indices in varieties values of m . As shown in Table 11, Λ demonstrates a better result, and is the least sensitive to change in m . The bold numbers in the Table 11 show the correct number of communities and the bold numbers in the Tables 8, 9, and 10 show the best validity values for each validity index and different values of m .

From the tests, the suggested validity index Λ is the only index which obtains the optimal c in all of experiments; moreover, when the different values of m are considered again, Λ has a better performance. Therefore, the proposed index Λ is used to evaluate PCMSA.

The result of clustering by center-based clustering algorithms such as PCM is shown in Fig. 13a which reveals no difference in assigning nodes to clusters in comparison to previous algorithms. Moreover, Fig. 13b indicates the result of the SA-cluster algorithm proposed by Zhou et al. (2009). It is considered that the left cluster is cluster 1 (blue color) and the other is cluster 2 (green color). According to Fig. 13b node 1037 is assigned to cluster 2 instead of cluster 1. In addition, nodes 167, 170, 564 and 979 and some others are assigned to one cluster; however, they have similar attributes with the other cluster according to dataset. This means that the SA-cluster might not consider overlap clusters. Figure 13c indicates the results of the proposed community detection model (PCMSA). In this figure, shared nodes between two clusters are shown with red color. The function of the proposed model, PCM, and SA-cluster models are evaluated considering validation index using (29). The validation index for PCM and SA-cluster are $\Lambda_{PCM} = 0.29$ and $\Lambda_{SA-cluster} = 0.32$, respectively, however, it is $\Lambda_{PCMSA} = 0.59$ for proposed model which indicates the better performance of PCMSA compared with PCM and SA-cluster models.

The efficiency of different clustering algorithms is compared in Fig. 14. This Figure indicates the clustering time on different number of nodes. The following observations on the runtime costs of different methods are made. First, PCMSA is usually 1.5–2.2 times slower than Algorithm 1, as it iteratively computes the re-clustering procedure. Although PCMSA is more expensive, but the iterative re-clustering improves the clustering quality a lot, as it is demonstrated in Fig. 13 and Table 15. The following observations show PCMSA is faster than the CODICIL (Ruan et al. 2013), SA-cluster (Zhou et al. 2009), K-SNAP (Tian et al. 2008), and FCAN (Hu and Chan 2016). K-SNAP is a little slower than PCMSA but the quality of PCMSA is much better than K-SNAP according to Table 15. Figure 14 suggests our method has promising scalability and computational complexity for analysis on even large social networks.

Example 5 This example is used to identify communities from social networks extracted from Facebook. This dataset is available in Stanford Large Network Dataset Collection <http://snap.stanford.edu/data/index.html>. This network consists of 4089 nodes and 170714 edges as it is shown in Fig. 15. A total of 193 communities have been distinguished and gender, job titles, institutions, etc., are considered as the attributes and they are 175 in total.

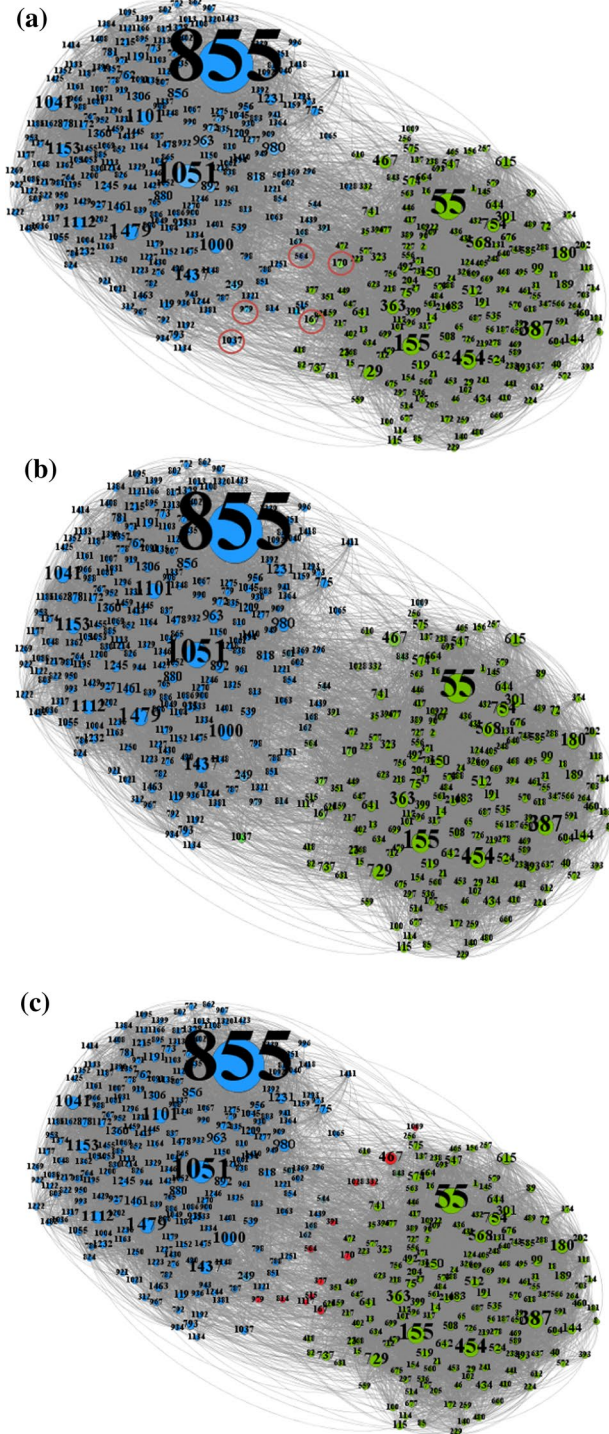


Fig. 13 a PCM algorithm. b SA-cluster algorithm. c PCMSA algorithm

Fig. 14 Clustering efficiency

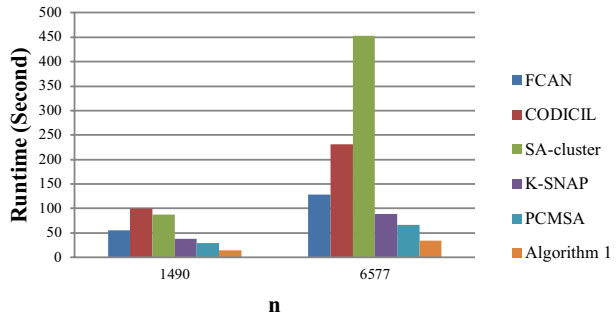


Fig.15 Facebook dataset

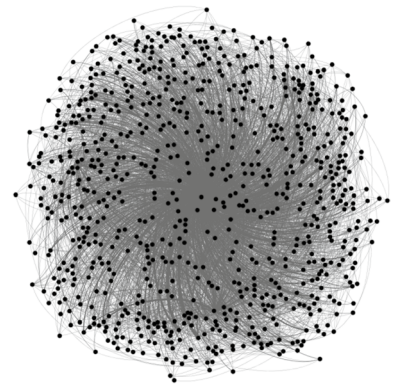


Table 12 Values of c preferred by validity indices with different values of c and m

m	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
Λ	193	193	190	193	193	193	193	193	190	193	193
Accuracy	186	193	186	189	193	189	186	186	181	181	189
NMI	191	193	187	187	189	189	189	191	193	191	193

As it is mentioned before, the objective function of the PCMSA depends on the weighting exponent $m \in [1, \infty)$. When a validity index is insensitive to change in m , it can be said that it is reliable. Therefore, here, all validity indices are considered for various values of both c and m . After running PCMSA, the results of reliability of each index by changing m are reported in Table 12.

As shown in Table 12, Λ demonstrates a better result, and is the least sensitive to change in m . Therefore, the proposed index Λ is used to evaluate PCMSA. The bold numbers in the Table 12 show the correct number of communities.

After running algorithms (FCAN, CODICIL, SA-cluster, K-SNAP, Algorithm 1, and PCMSA) with this dataset, the results are obtained as follows: $\Lambda_{FCAN} = 0.45$, $\Lambda_{CODICIL} = 0.33$, $\Lambda_{SA-cluster} = 0.49$, $\Lambda_{K-SNAP} = 0.29$, $\Lambda_{Algorithm1} = 0.38$, $\Lambda_{PCMSA} = 0.56$. PCMSA, again, performs better than the other algorithms and SA-cluster is ranked second. As it is mentioned before SA-cluster does not consider the overlapping communities and cannot handle noise in data. After SA-cluster, FCAN is the better one. FCAN

cannot handle and balance attribute and structural similarities as well as PCMSA. In their problem formulation, the structural distance and attribute distance are combined, while they are two seemingly independent, or even conflicting goals and it does not make sense semantically.

Moreover, the efficiency of different clustering algorithms is compared in Fig. 16. According to our analysis PCMSA is faster than the all algorithms, except the Algorithm 1 as it iteratively computes the re-clustering procedure. Figure 16 suggests our method has promising scalability and computational complexity.

Example 6: The sixth sample is the twitter dataset that is available for download from Stanford Network Dataset Collection <http://snap.stanford.edu/data/index.html>. This dataset has 125,120 nodes and 2,248,406 edges and 3140 communities. This network is indicated in Fig. 17. The number of attributes for each node is 33569. The hashtags used by users in their tweets are considered as the attributes.

Fig. 16 Clustering efficiency

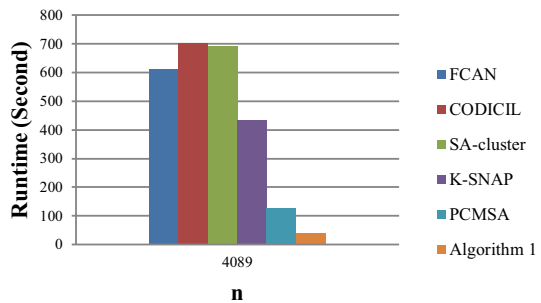


Fig. 17 Twitter dataset

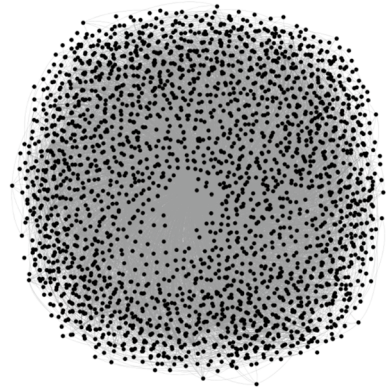


Table 13 Values of c preferred by validity indices with different values of c and m

m	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
Λ	3140	3138	3140	3140	3140	3140	3136	3140	3140	3140	3140
Accuracy	3140	3132	3132	3137	3140	3137	3131	3131	3131	3132	3132
NMI	3138	3138	3132	3140	3140	3140	3135	3135	3135	3129	3135

Fig. 18 Twitter dataset

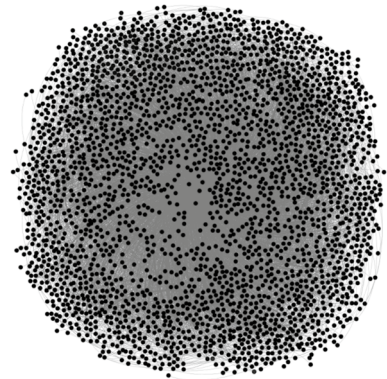
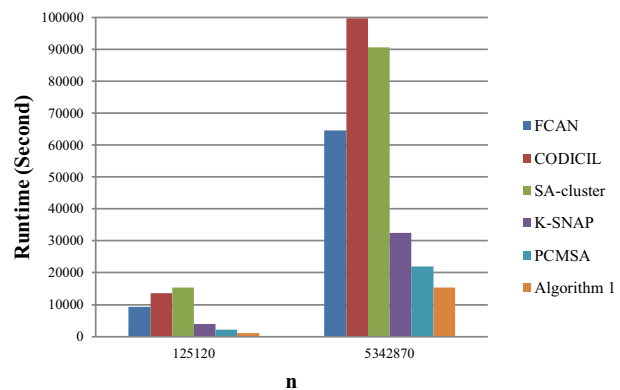


Table 14 Values of c preferred by validity indices with different values of c and m

m	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
Λ	5016	5020	5020	5020	5020	5018	5020	5020	5020	5020	5020
Accuracy	5018	5018	5020	5015	5015	5018	5020	5020	5018	5015	5015
NMI	5015	5017	5020	5017	5017	5017	5015	5015	5020	5015	5018

Fig. 19 Clustering efficiency



Like the previous examples, all validity indices are considered for various values of both c and m . After running PCMSA, the results of reliability of each index by changing m are reported in Table 13.

As shown in Table 13, Λ demonstrates a better result, and is the least sensitive to change in m . Therefore, the proposed index Λ is used to evaluate PCMSA. The bold numbers in the Table 13 show the correct number of communities.

After running algorithms (FCAN, CODICIL, SA-cluster, K-SNAP, Algorithm 1, and PCMSA) with this dataset, the results are obtained as follows: $\Lambda_{FCAN} = 0.34$, $\Lambda_{CODICIL} = 0.26$, $\Lambda_{SA-cluster} = 0.38$, $\Lambda_{K-SNAP} = 0.31$, $\Lambda_{Algorithm1} = 0.41$, $\Lambda_{PCMSA} = 0.62$.

Table 15 Comparison of different methods

Example	Sample	n	E	e	FCAN	CODICIL	SA-cluster	K-SNAP	Algorithm 1	PCMSA
1	1	10	9	1	0.39	0.30	0.38	0.26	0.36	0.45
2	2	11	12	2	0.45	0.36	0.48	0.40	0.52	0.59
3	3	256	401	7	0.21	0.10	0.11	0.19	0.14	0.28
	4	1984	3249	7	0.30	0.27	0.26	0.29	0.25	0.36
	5	6987	10645	7	0.35	0.29	0.32	0.31	0.41	0.48
	6	10344	15463	7	0.32	0.20	0.41	0.31	0.43	0.54
	7	117298	484673	7	0.48	0.43	0.42	0.49	0.51	0.57
	8	314657	981254	7	0.50	0.40	0.48	0.39	0.42	0.55
	9	645987	1029384	7	0.38	0.39	0.36	0.32	0.45	0.48
	10	3452788	76849635	7	0.45	0.41	0.43	0.44	0.50	0.52
4	11	1490	19025	2	0.42	0.26	0.32	0.30	0.38	0.59
	12	6577	87345	2	0.31	0.14	0.19	0.20	0.26	0.39
5	13	4089	170714	175	0.45	0.33	0.49	0.29	0.38	0.56
6	14	125120	2248406	33569	0.34	0.26	0.38	0.31	0.41	0.62
7	15	5342870	20037423	42342	0.41	0.21	0.50	0.16	0.52	0.59

PCMSA performs better than the other algorithms and Algorithm 1 is ranked second. After Algorithm 1, SA-cluster is the better one.

The efficiency of different clustering algorithms considering computational complexity is compared in Fig. 19.

Example 7 The final sample is the twitter dataset that is available for download from Stanford Network Dataset Collection <http://snap.stanford.edu/data/index.html>. This network consists of 5342870 nodes and 20037423 edges. A part of this network is shown in Fig. 18. A total of 5020 communities have been distinguished. The number of attributes for each node is 42342. The hashtags used by users in their tweets are considered as the attributes.

Here, all validity indices are considered for various values of both c and m . After running PCMSA, the results of reliability of each index by changing m are reported in Table 14.

As shown in Table 14, Λ demonstrates a better result, and is the least sensitive to change in m . Therefore, the proposed index Λ is used to evaluate PCMSA. The bold numbers in the Table 14 show the correct number of communities.

After running algorithms (FCAN, CODICIL, SA-cluster, K-SNAP, Algorithm 1, and PCMSA) with this dataset, the results are obtained as follows: $\Lambda_{FCAN} = 0.41$, $\Lambda_{CODICIL} = 0.21$, $\Lambda_{SA-cluster} = 0.50$, $\Lambda_{K-SNAP} = 0.16$, $\Lambda_{Algorithm1} = 0.52$, $\Lambda_{PCMSA} = 0.59$. PCMSA, again, performs better than the other algorithms.

The efficiency of different clustering algorithms for examples 6 and 7 is compared in Fig. 19. The clustering time is indicated in this figure. The following observations on the runtime costs of different methods are made. PCMSA is 1.4–2.1 times slower than Algorithm 1, as it iteratively computes the re-clustering procedure. Although PCMSA is more expensive, but the iterative re-clustering improves the clustering quality a lot, as it is demonstrated in Table 15. The following observations show PCMSA is faster than the CODICIL (Ruan et al. 2013), SA-cluster (Zhou et al. 2009), K-SNAP (Tian et al. 2008), and FCAN (Hu and Chan 2016). K-SNAP is a little slower than PCMSA but the quality of PCMSA is much better than K-SNAP according to Table 15. Figure 19 suggests our method has promising scalability and computational complexity for analysis on even large social networks.

Finally, the validity indices of different methods for all samples are illustrated in Table 15. As it is shown in Table 15, in all samples the performance of the PCMSA is better than the others. Samples 1 and 2 are described before in Examples 1 and 2, respectively. Samples 3–10 are described in Example 3 and discussed in Fig. 9. Samples 11 and 12 are discussed in Example 4 and described in Fig. 14. Sample 13 is discussed in example 5 and described in Fig. 16. Finally samples 14, and 15 are mentioned in Examples 6, and 7, respectively and described in Fig. 19.

Considering Table 15, PCMSA has performed better not only in low dimensional data sets, but also in high dimensional ones where complexity increases. Also as it is shown in Figs. 9, 14, 16, and 19, the time efficiency of PCMSA was better in all examples.

7 Conclusion

In this research, we proposed a fuzzy model for overlapping community detection based on the PCM algorithm in complex networks. This model (PCMSA) identifies communities in the social networks based on both resources of data related to the nodes' attributes and nodes' structure. Moreover, PCMSA strictly takes attribute and structural similarities into consideration instead of balancing them. Therefore, the proposed community detection algorithm is capable of clustering the graph into quality partitions that have high structural and attribute similarities. The performance of the proposed model was shown with several real-world and artificial networks from small to very large sizes. The trade-off between density and homogeneity was used to assess the model and to specify the most favorable number of clusters. In addition, the structure of the validity index was well adapted for graph clustering considering both link information and node attribute. Results indicated that the proposed model detects communities in a better manner than the other algorithms, especially when some nodes are shared between clusters. Our experiments showed that PCMSA detects meaningful and insightful patterns in both synthetic and large scale real social networks. Also the experimental findings reveal the superiority of this novel model and its promising scalability and computational complexity over others. In future works, the community detection based on other clustering algorithms will be considered.

Appendix 1

Proof of Theorem

Theorem 1 All u_{ik} in $U, \forall i, k$ are independent. Hence, minimization of $J(U, V)$ with regard to U is similar to that of $J(u_{ik}, v_i)$ regarding u_{ik} . The gradient of $J(u_{ik}, v_i)$ with respect to u_{ik} is set to zero in an attempt to find the first-order necessary conditions for optimality:

$$\frac{\partial J}{\partial u_{ik}} = m u_{ik}^{m-1} D_{ik} - m \Delta_i (1 - u_{ik})^{m-1} = 0 \Rightarrow u_{ik} = \left(1 + \left(\frac{D_{ik}}{\Delta_i} \right)^{1/(m-1)} \right)^{-1} \quad (30)$$

To find the most favorable node as the center of cluster, a node with the closest in structure to other members of the cluster considering their membership value u_{ik} should be selected. As a result, the center of cluster i is defined as follows:

$$v_i^* = \arg \min_{v_i \in [1, n]} \left(\sum_{k=1}^n \sum_{j=1}^n u_{ik}^m D_{ik} \right) \quad (31)$$

Theorem 2 From (18), the necessary conditions considering partial derivatives are as follows:

$$\begin{aligned} \frac{\partial J}{\partial u_{ik}} &= m(u_{ik})^{m-1}d_{ik}(\Omega_i) + m\delta_i(1 - u_{ik})^{m-1} = 0 \\ u_{ik}^* &= \left(1 + \left(\frac{d_{ik}(\Omega_i)}{-\delta_i}\right)^{1/(m-1)}\right)^{-1} \end{aligned} \tag{32}$$

$$\begin{aligned} \left.\frac{\partial J}{\partial v_i}\right|_* &= -2 \sum_{k=1}^n (u_{ik})^m H_i(x_k - v_i^*) = 0; i = 1, 2, \dots, c \\ v_i^* &= \frac{\sum_{k=1}^n u_{ik}^m x_k}{\sum_{k=1}^n u_{ik}^m} \end{aligned} \tag{33}$$

Note that $\frac{\partial}{\partial x}(x^T Hx) = 2Hx$ in which H is symmetric and is not a function of x .

and finally,

$$\begin{aligned} \left.\frac{\partial J}{\partial H_i}\right|_* &= \sum_{k=1}^n u_{ik}^m (x_k - v_i)(x_k - v_i)^T + \lambda_i |H_i^*| H_i^{*-1} = 0 \\ H_i^{*-1} &= \frac{1}{\lambda_i |H_i^*|} \sum_{k=1}^n u_{ik}^m (x_k - v_i^*)(x_k - v_i^*)^T \end{aligned} \tag{34}$$

The identities $\frac{\partial}{\partial H}(x^T Hx) = xx^T, \frac{\partial}{\partial H}|H| = |H|H^{-1}$ are used for a non-singular matrix H and any compatible vector x (Gustafson and Kessel 1978).

Appendix 2

The required condition for converging of the algorithm proposed in Fig. 2 is met when:

$$\lim_{t \rightarrow \infty} \|U^{(t)} - U^{(t-1)}\| = 0 \tag{35}$$

The iterative formula for u_{ik} originates from the classical gradient descent method (Newton–Raphson method) (Kelley 1999), with J_m as the error function which should be minimized, namely:

$$u_{ik}^{(t)} = u_{ik}^{(t-1)} - \tau^{(t)}(J_m(u_{ik}, \Omega_i, \delta_i)^{(t-1)}) \left(\frac{\partial J_m(u_{ik}, \Omega_i, \delta_i)^{(t-1)}}{\partial u_{ik}}\right)^{-1} \tag{36}$$

where $\tau^{(t)}$ denotes a positive parameter of learning rate and $2 \leq c \leq 20$ stands for the gradient of J_m with respect to u_{ik} at $(t-1)$ iteration. Re-writing (36) for U renders:

$$U^{(t)} - U^{(t-1)} = -\tau^{(t)}(J_m(U, \Omega, \delta)^{(t-1)}) \left(\frac{\partial J_m(U, \Omega, \delta)^{(t-1)}}{\partial U}\right)^{-1} \tag{37}$$

Now, putting (35) in (37):

$$\lim_{t \rightarrow \infty} \|U^{(t)} - U^{(t-1)}\| = \lim_{t \rightarrow \infty} \left(\left\| \tau^{(t)} \left\| J_m(U, \Omega, \delta)^{(t-1)} \right\| \left\| \frac{\partial J_m(U, \Omega, \delta)^{(t-1)}}{\partial U} \right\|^{-1} \right\| \right) \quad (38)$$

By considering $\tau^{(t)} = \frac{\psi^{(t)}}{\|J_m(U, \Omega, \delta)^{(t-1)}\| \left\| \frac{\partial J_m(U, \Omega, \delta)^{(t-1)} \right\|^{-1}}$, (38) becomes:

$$\lim_{t \rightarrow \infty} \|U^{(t)} - U^{(t-1)}\| = \lim_{t \rightarrow \infty} \|\psi^{(t)}\| = 0 \quad (39)$$

where, $\psi^{(t)} = \psi_0^{(t)}/t$ in which $\psi_0^{(t)}$ represents a constant value, and $\psi_0^{(t)} \rightarrow 0$ when $t \rightarrow \infty$. Hence, (B.5) is proved and, as a result, the suggested algorithm is convergent.

Appendix 3

NMI is a measurement index that measures the degree of matching between the communities identified by different algorithms and that of the expected. Consider $F = \{F_k\} (1 < k < c)$ is the expected communities. *NMI* is defined as follows (Hu and Chan 2016):

$$NMI = \frac{\sum_{k_1=1}^c \sum_{k_2=1}^c n_{C_{k_1}, F_{k_2}} \log \left(\frac{n_{*n_{C_{k_1}, F_{k_2}}}}{n_{C_{k_1}} * n_{F_{k_2}}} \right)}{\sqrt{\left(\sum_{k=1}^c n_{C_k} \log \frac{n_{C_k}}{n} \right) \left(\sum_{k=1}^c n_{F_k} \log \frac{n_{F_k}}{n} \right)}} \quad (40)$$

where n_{C_k} is the number of nodes in C_k , n_{F_k} is the number of nodes in F_k , and $n_{C_{k_1}, F_{k_2}}$ is the number of nodes discovered in both C_{k_1} and F_{k_2} .

For *Accuracy* measure (Hu and Chan 2016), the mapping function $Z : C_{k_1} \rightarrow F_{k_2}$ is needed. To find Z , the $n_{C_{k_1}, F_{k_2}}$ is determined for all combinations of C_{k_1} and F_{k_2} in C and F , respectively. Given that, this is an iterative process, for each iteration, starting from the largest $n_{C_{k_1}, F_{k_2}}$, the C_{k_1} that matches against F_{k_2} is determined; then the mapping $C_{k_1} \rightarrow F_{k_2}$ to Z is added and C_{k_1} and F_{k_2} are ignored in future iterations. The process ends when all C_{k_1} in C has found a match in F . This measurement is defined as follows:

$$Accuracy = \frac{1}{n} \sum_{k=1}^c n_{C_k, Z(C_k)} \quad (41)$$

Therefore, the values of *NMI* and *Accuracy* are larger when C_k matches better with the expected result F .

References

Adamic LA, Glance N (2004) The political blogosphere and the 2004 US election. In n Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem, 2005

Andersen R, Chung F, Lang K (2006) Local graph partitioning using pagerank vectors. Proc - Annu IEEE Symp Found Comput Sci FOCS, pp. 475–483

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms, vol 25, no 3. Utah state university, Logan, Utah, Plenum press, New York

- Bezdek JC, Keller J, Krisnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing, vol 18, no 2. Kluwer Academic Publisher, Boston, London, Dordrecht
- Bu Z et al (2019) Graph K-means based on leader Identification, dynamic game and opinion dynamics. *IEEE Trans Knowl Data Eng* 32(7):1348–1361
- Cao J, Bu Z, Wang Y, Yang H, Jiang J, Li H (2019) Detecting prosumer-community groups in smart grids from the multiagent perspective. *IEEE Trans Syst Man Cybern Syst* 49(8):1652–1664
- Dunn JC (1974) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 3:32–57
- Flake GW, Lawrence S, Giles CL (2000) Efficient identification of Web communities. *Proc sixth ACM SIG-KDD Int Conf Knowl Discov data Min - KDD '00*, 150–160
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174
- Fu X, Liu L, Wang C (2013) Detection of community overlap according to belief propagation and conflict. *Phys A Stat Mech its Appl* 392(4):941–952
- Höppner F (Ed.) (1999), *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*
- Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99(12):7821–7826
- Golsefid SMM, Fazel Zarandi MH, Bastani S (2015) Fuzzy duocentric community detection model in social networks. *Soc Networks* 43:177–189
- Granovetter MS (1977) The strength of weak ties, vol 1380. ACADEMIC PRESS, INC.
- Gustafson DE, Kessel WC (1978) Fuzzy-clustering-with-a-fuzzy-covariance-matrix. *IEEE*
- Hu L, Chan KCC (2016) Fuzzy clustering in a complex network based on content relevance and link structures. *IEEE Trans Fuzzy Syst* 24(2):456–470
- Kadushin C (2004) *Understanding social network*
- Kelley CT (1999) *Iterative methods for optimization*
- Krishnapuram R, Keller JM (1993) A Possibilistic approach to clustering. *IEEE Trans Fuzzy Syst* 1(2):98–110
- Mendel JM, Mendel JM (2017) *Uncertain rule-based fuzzy systems*
- Malek Mohamadi Golsefid S, Fazel Zarandi MH, and Susan B (2015) Fuzzy community detection model in social networks. *Int J Intell Syst* 30:1227–1244
- Pathak N, Delong C, Banerjee A (2008) Social topic models for community extraction, 2nd SNA-KDD Work., pp. 565–574
- Ruan Y, Fuhry D, Parthasarathy S (2013) Efficient community detection in large networks using content and links. In *WWW '13 proceedings of the 22nd international conference on World Wide Web*, pp. 1089–1098
- Schaeffer SE (2007) Graph clustering. *Comput Sci Rev* 1(1):27–64
- Sun Y, Han J, Gao J, Yu Y (2009) iTopicModel : information network-integrated topic modelling. In 2009 Ninth IEEE international conference on data mining
- Tan WW, Chua TW, Cliffs E (2007) Book review, no. February
- Tian Y, Hankins RA, Patel JM (2008) Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 567–580
- Traud AL, Mucha PJ, Porter MA (2011) Social structure of facebook networks. *PDF, Arxiv Prepr. arXiv1102.2166*, 2011., 391(16)pp. 4165–4180,
- Valente de Oliveira J, Pedrycz W (2007) *Advances in fuzzy clustering and its applications*. Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England
- Wang W, Liu D, Liu X, Pan L (2013) Fuzzy overlapping community detection based on local random walk and multidimensional scaling. *Phys A* 392(24):6578–6586
- Wasserman S, Faust K (1994) *Social network analysis: methods and applications, Methods and Applications*. p. 825
- Yang MS (1993) A survey of fuzzy clustering. *Mat Comput Model* 18(11):1–16
- Yang T, Jin R, Chi Y, Zhu S (2009) Combining link and content for community detection: a discriminative approach. *Proc 15th ACM SIGKDD Int Conf Knowl Discov data Min*, pp. 927–936
- Zarandi MHF, Razaee ZS (2010) A fuzzy clustering model for fuzzy data with outliers. *Int J Fuzzy Syst Appl* 1(2):1–18
- Zarandi MHF, Faraji MR, Karbasian M (2010) An exponential cluster validity index for fuzzy clustering with crisp and fuzzy data. *Sci Iran* 17(2):95–110
- Zarinbal M, Fazel Zarandi MH, Turksen IB (2014) Relative entropy fuzzy c-means clustering. *Inf Sci (Ny)* 260:74–97
- Zhang S, Wang R-S, Zhang X-S (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Phys A-Stat Mech Its Appl* 374:483–490

- Zhou Y, Cheng H, Yu JX (2009) Graph clustering based on structural / attribute similarities. *Vldb* 2(1):718–729
- Zhou Y, Cheng H, Yu JX (2010) Clustering large attributed graphs: an efficient incremental approach. Proc. - IEEE Int. Conf. Data Mining, ICDM, pp 689–698

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.