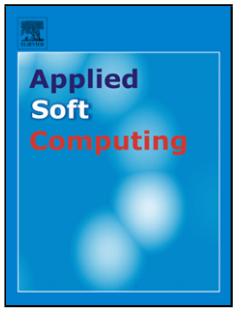


Title: Fuzzy Clustering in Community Detection Based on Nonnegative Matrix Factorization with Two Novel Evaluation Criteria

Author: Neda Binesh Mansoor Rezghi

PII: S1568-4946(16)30637-8
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2016.12.019
Reference: ASOC 3958

To appear in: *Applied Soft Computing*

Received date: 6-4-2016
Revised date: 8-12-2016
Accepted date: 9-12-2016

Please cite this article as: Neda Binesh, Mansoor Rezghi, Fuzzy Clustering in Community Detection Based on Nonnegative Matrix Factorization with Two Novel Evaluation Criteria, *<![CDATA[Applied Soft Computing Journal]]>* (2017), http://dx.doi.org/10.1016/j.asoc.2016.12.019

Fuzzy Clustering in Community Detection Based on Nonnegative Matrix Factorization with Two Novel Evaluation Criteria

Neda Binesh , Mansoor Rezghi

Department of Computer Science, Tarbiat Modares University, Tehran, Iran
Corresponding author: Rezghi@modares.ac.ir

Abstract

Clustering or community detection is one of the most important problems in social network analysis, and because of the existence of overlapping clusters, fuzzy clustering is a suitable way to cluster these networks. In fuzzy clustering, in addition to the correctness of the clusters assigned to each node, the produced membership of one node to each cluster is also important. In this paper, we introduce a new fuzzy clustering algorithm based on the nonnegative matrix factorization (NMF) method. Despite the well-known fuzzy clustering techniques like FCM, the proposed method does not depend on any parameter. Also, it can produce appropriate memberships based on the network structure and so identify the overlap nodes from non-overlap nodes, well. Also, to evaluate the validity of such fuzzy clustering algorithms, we propose two new evaluation criteria (*SFEC* and *UFEC*), which are constructed based on the neighborhood structure of nodes and can evaluate the memberships. Experimental results on some real-world networks and also many artificial networks show the effectiveness and reliability of our proposed criteria.

Keywords: Community detection, Fuzzy clustering, Nonnegative Matrix Factorization, Fuzzy C-means, Fuzzy membership matrix.

1. Introduction

Recently, analysis of large networks in biology, science, technology and social systems has become very popular [1, 2]. Community detection in networks is an important area of current research with many applications [3,

4, 5]. In a complex network, a community is a set of entities that share some closely correlated features with each other and have less connection to other entities. In addition to the variety of definitions of community, they have a number of common interesting features such as overlapping configuration. This means that the communities could have some common actors in the overlaps between them which are named as *multi-cluster* nodes and the other as *single-cluster* nodes.

Due to the existence of multi-cluster nodes in real networks, the overlapping community detection methods should be used. These methods can be classified into fuzzy and non-fuzzy methods.

Methods in non-fuzzy class work like hard clustering algorithms except that allow nodes belong to different clusters [4, 6, 7]. But fuzzy methods are more general and give more information. In this approach, the memberships of every node to each cluster can be computed. The result of a fuzzy clustering algorithm is a stochastic membership matrix U, where its elements are named memberships and show the probability of belonging every node to different clusters. So for a multi-cluster node, in addition to the labels of clusters, its tendency to them is also known[8]. By using any threshold, fuzzy methods could be used as non-fuzzy methods.

In an appropriate fuzzy clustering algorithm, the produced memberships of a multi-cluster node to different clusters should be soft and non-crisp, but for a single-cluster node, the memberships should be crisp like hard clustering.

Computing an appropriate fuzzy membership matrix depends on the choice of a suitable fuzzy clustering algorithm. Fuzzy C-Means method (FCM) is the most known fuzzy clustering algorithm that was initially proposed by Dunn [9] and generalized by Bezdek [10, 11]. Also, this method has been used as a part of many other fuzzy clustering algorithms implicitly[12]. To perform FCM algorithm, in addition to the number of clusters, we should determine a fuzziness parameter. Although this parameter has a significant role in computing fuzzy memberships, there is no way to find the optimal value of this parameter for each network.

Also in the last decades, some new fuzzy overlapping community detection methods have been proposed. Zhang [12] proposed a method based on the combination of the spectral mapping and fuzzy c-means. In [13] the task is modeled as a nonlinear constrained optimization problem. In [8] a fuzzy clustering algorithm is proposed based on the local random walk (LRW) and a new distance metric. Although in [8] it has been shown that this method is

better than the proposed method by Nepusz [13], this algorithm still depends on the number of random walk step and a fuzzy parameter for applying fuzzy C-means in the last step. SLPA is another overlapping community detection method which does not compute memberships[7, 14]. This non-fuzzy algorithm is a general speaker-listener based information propagation process.

Recently nonnegative matrix factorization (NMF) widely is used for community detection in social networks [15, 16, 17, 18]. The factors of NMF decomposition are nonnegative which is suitable for analysis of nonnegative data. We show that this property enables us to use NMF as a fuzzy clustering method that unlike FCM does not depend on any extra fuzzy parameter. We also show that our proposed algorithm based on NMF able to assign proper membership to each node and detect overlapping nodes well. Experiments on some test problems show that the membership matrix produced by NMF based method is consistent with the structure of the networks. Therefore in the membership matrix produced by our NMF based method, the corresponding rows to the single-cluster nodes are crisp, and the rows corresponding to multi-cluster nodes are fuzzy. We show that this does not occur for the other fuzzy methods like FCM and LRW. For comparing the quality of different fuzzy clustering methods, some appropriate evaluation criteria should be used.

Fuzzy clustering evaluation criteria can be divided into two main categories. In the first one, the membership matrix doesn't appear in the evaluation criteria directly, and only the labels of clusters are used.

In this approach, Normalized Mutual Information (NMI) [19] and Omega-Index [20] for ground truth networks (Networks with known clusters), and modularity measure (Q) [21] and its extension named (Q_{ov}) [22] for networks without ground truth are the most widely used criteria. Although in [23] the authors proposed three methods to adapt the existing crisp quality measures to handle graph overlaps even they don't use the memberships.

In the second approach, the memberships from fuzzy clustering algorithms are used to compute evaluation criterion. This approach is more consistent than the first approach with the nature of fuzzy clustering. Some validity indices in this approach such as the partition coefficient (PC) [24] and the partition entropy (PE) [25] don't consider the neighborhood structure of the network and the connections between the nodes. Some other validity index methods such Xie-Beni's index [26] and the Fukayama-Sugno cluster-validity criterion [27] or WGLI index [28] involves both the membership values and

the input feature vectors. But these criteria are constructed based on the FCM object function and therefore are not appropriate as a general method for evaluating any fuzzy partition matrix such the one produced by NMF.

Although, in fuzzy clustering, the produced memberships are important, there aren't straightforward measures to evaluate them. In this paper, two new evaluation criteria based on the neighborhood structure of nodes are proposed. The first is the Supervised Fuzzy Evaluation Criterion *SFEC*, which can evaluate the fuzzy membership matrix for the networks with the known clusters. The other proposed criterion is Unsupervised Fuzzy Evaluation Criterion(*UFEC*), that is proposed for the networks without ground truth. This criterion unlike the most validity indices that are constructed based on FCM objective function doesn't depend on the method that computes the fuzzy memberships.

The rest of the paper is organized as follows: in section 2 fuzzy community detection and FCM method are described and section 3 introduce NMF as a fuzzy clustering algorithm. Section 4 define our measure to evaluate fuzzy clustering methods and section 5 contains some experimental results on some social networks to see the performance of proposed criteria and compare the NMF based fuzzy algorithm with some other fuzzy clustering method. Experiments show the superiority of the proposed fuzzy clustering and evaluation criteria.

2. Background

2.1. Fuzzy community detection

The clustering algorithms classified into hard(crisp) and soft(fuzzy) methods. In the hard clustering, the data divided into distinct clusters, where each data belongs exactly to one cluster, but In soft clustering, the data can belong to more than one cluster. In this approach, real values in [0,1] assigned to each node that indicates the amounts of its dependency on each cluster. After computing the fuzzy membership matrix, each node belongs to the one or more clusters that its membership to them is more than the other clusters.

Consider a network with n node and the weight matrix $W = (w_{ij})_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, where $w_{i,j}$ for connected nodes i and j is nonzero and denotes the similarity between these nodes.

A fuzzy clustering method uses the weight matrix W and gives the fuzzy membership matrix $U = (u_{i,k}) \in \mathbb{R}^{n\times c}$, where c is the number of clusters and

$\forall i, k,\ 0 \leq u_{i,k} \leq 1$ denotes the probability of belonging node i to the cluster k and so

$$\sum_{k=1}^{c} u_{ik} = 1, \quad 1 \leq i \leq n. \tag{1}$$

2.2. Fuzzy C-means

FCM algorithm is an iterative clustering method that finds the membership matrix $U \in \mathbb{R}^{n \times c}$, by minimizing the following object function:

$$J_{FCM} = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^{f} d^2(\mathbf{x}_i, \mathbf{m}_k). \tag{2}$$

Here, the fuzziness parameter f is a real-valued number greater than 1 which controls the fuzziness of the resulting clusters and $\mathbf{m}_k$ is the prototype of the center of cluster k. Also $d^2(\mathbf{x}_i, \mathbf{m}_k)$ is a distance measure between object $\mathbf{x}_i$ and cluster center $\mathbf{m}_k$ that $\|\mathbf{x}_i - \mathbf{m}_k\|_F^2$ almost is used. The solution of the object function J_{FCM} can be approximated by the following algorithm.

1. Set values for c , f, $t = 0$ and initial membership matrix $U^{(0)}$.
2. Calculate the cluster centers $\{\mathbf{m}_k^{(t)}\}$ at step t as follows:

$$\mathbf{m}_k^{(t)} = \frac{\sum_{i=1}^{n} (u_{ik}^{(t)})^f \mathbf{x}_i}{\sum_{i=1}^{n} (u_{ik}^{(t)})^f} \tag{3}$$

3. Calculate the updated membership matrix $U^{(t+1)}$:

$$u_{ik}^{(t+1)} = \left[\sum_{l=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{m}_k^{(t)}\|}{\|\mathbf{x}_i - \mathbf{m}_l^{(t)}\|} \right)^{\frac{2}{f-1}} \right]^{-1} \tag{4}$$

4. If converged , stop ; otherwise, set $t = t + 1$ and go to step 2.

Recently a new fuzzy clustering that uses FCM has been proposed in [8]. This is based on the local random walk (LRW) and a new distance metric. Based on the new distance measurement, the dissimilarity index between each node of a network is calculated, and then the network is mapped into low-dimensional space by the multidimensional scaling (MDS). Finally, FCM clustering is employed to find fuzzy communities in the network. Although this method is better than the FCM, still depends on the number of random walk step and a fuzzy parameter for applying fuzzy C-means in the last step.

2.3. The fuzziness parameter f

The results of FCM algorithm completely depend on the fuzzy parameter f.Whatever f be near to 1 the fuzzy partition matrix U is more crisp and its components are closer to 0 or 1. So the clustering will be a *Hard Clustering* such as k-means, but when f is larger, the components of each row of U are fuzzier and far away from 0 and 1.

Estimation of the fuzzy parameter for each network is not straightforward. In the literature about FCM, f usually is 2, which is not an appropriate fuzziness parameter for all networks [29]. In (2) when f tends to infinity, from equation (4), it is clear that for every i, k $u_{ik} = \frac{1}{c}$. In practice this occurs for $f \gg 1$. In this case, based on equation (3) it is clear that for every i, $\mathbf{m}_i$ become equal to the mass center of the data set. Thus in these cases, FCM failed to extract any clusters. So by increasing f all of the rows of matrix U become fuzzy simultaneously and for small values of f all of them become crisp. Therefore, this method is not able to find an appropriate membership matrix that distinguishes single and multi-cluster nodes.

In the following section, we propose a new fuzzy clustering algorithm based on NMF method that unlike FCM doesn't need any fuzzy parameter and able to assign reasonable memberships.

3. A new fuzzy clustering algorithm based on Nonnegative Matrix Factorization

The Nonnegative Matrix Factorization technique (NMF) is a machine-learning algorithm, which has been used in different applications as a dimension reduction, classification or clustering method[16, 30, 31].

For a given index c, this method estimates the data matrix $X \in \mathbb{R}^{m\times n}$ by product of two nonnegative matrices $V \in \mathbb{R}^{m\times c}$ and $H \in \mathbb{R}^{c\times n}$ from the following minimization problem [32]

$$\min_{V,H\geq 0} \|X - VH\|, \tag{5}$$

which $\|.\|$ denotes the Frobenious norm, I-divergence [33] or other distances. In clustering problems the parameter $c << min(m, n)$ is considered as the number of clusters and can be estimated by various ways [34]. For a symmetric data matrix W, the factors W and H can be considered as $W = H^{\mathsf{T}}$[35] and 5 becomes

$$\min_{H\leq 0} \|X - H^{\mathsf{T}}H\| \tag{6}$$

The NMF factorization is a nonconvex problem, and different type of algorithms such as methods based on alternating nonnegative least square (ANLS) approach have been proposed to solve this problem[36, 37]. At each step of ANLS type methods, one factor is considered to be known, and so the another one is estimated, and this process will continue intermittently until the convergence. The main idea of this approach is that by fixing one factor, finding the another one becomes a convex problem, that can be solved by well known nonnegative least square methods such as an active set based ANLS method proposed by Kim [38].

For a small value of c, the dimensions of produced matrices will be reduced and also if the factor matrices are computed well, they able to identify hidden information of the data. The ability of NMF in clustering and community detection discussed in many papers [16, 39, 40]. In the following, we show that how NMF can be used as a fuzzy clustering technique and produce a fuzzy membership matrix.

From equation (5) data matrix X can be approximated as

$$X \simeq VH = [V\mathbf{h}_1, V\mathbf{h}_2, \cdots, V\mathbf{h}_n], \tag{7}$$

which $\mathbf{h}_i$ is the i-th column of H. Now if $\mathbf{x}_i$ and $\mathbf{v}_j$ be the i-th and j-th columns of X and V, respectively, we have

$$\mathbf{x}_i \simeq V\mathbf{h}_i = \sum_{j=1}^{c} h_{ji}\mathbf{v}_j. \tag{8}$$

In linear algebra viewpoint, $\mathbf{v}_j$, $1 \leq j \leq c$, are corresponding to the new bases and according to (8), $\mathbf{x}_i$ is approximated by the linear combination of these new bases. Here the components of $\mathbf{h}_i$ are its new coordinates.

In using NMF as a clustering method, $\mathbf{v}_j$, $1 \leq j \leq c$ can be considered as the prototype of clusters and for every i, $\mathbf{h}_i$ is a vector in $\mathbb{R}^c$ that h_{ji} show the amounts of resemblance of $\mathbf{x}_i$ in the j-th cluster. The advantage of NMF is producing nonnegative matrices which cause all the new coordinates be positive and show the amounts of belonging each node to each cluster. Therefore if

$$l = \underset{1\leq j\leq c}{\operatorname{argmax}} h_{ji},$$

$\mathbf{x}_i$ will be in the cluster l. This property does not exist in another matrix decomposition methods like singular value decomposition[32].

This property of NMF enables us to consider it as a fuzzy clustering method. We can normalize $\mathbf{h}_i$ by dividing its components on $\|\mathbf{h}_i\|_1$, so that we obtain a vector $\mathbf{h}_i/\|\mathbf{h}_i\|_1$ where denotes the probability of belonging $\mathbf{x}_i$ to different clusters. So

$$U = \begin{bmatrix} \mathbf{h}_1^T/\|\mathbf{h}_1\|_1 \\ \vdots \\ \mathbf{h}_n^T/\|\mathbf{h}_n\|_1 \end{bmatrix}, \tag{9}$$

can be interpreted as a fuzzy membership matrix and NMF is a matrix method that can do fuzzy clustering. Our method is summarized in Algorithm 1.

Algorithm 1: **NMF based Fuzzy clustering Algorithm**

Input: Network's weight matrix W , The number of clusters c
Output: The Membership matrix U
Steps:
1. Decomposition matrix W by an appropriate NMF method
2. Calculate U from (9)

The complexity of this fuzzy clustering method is dominated by computing the NMF factorization in step 1, which is of order cn^2. Also, this for sparse matrices like the matrices of the social networks will be reduced[16].

Unlike fuzzy C-means, in using of NMF as a fuzzy clustering method, there is no need to specify any parameter. Also in our experiments, we observed that the memberships from NMF have more consistency with the reality and NMF can distinguish nodes in the center and the border of clusters well.

For example, consider the sample network in Figure 1 from [16], that have 4 clusters with some nodes on the overlap.

In this network nodes from 1 to 30 are belong to clusters 1 to 4 which are corresponding to the colors red, blue, green and yellow, respectively. Table 1 shows the memberships computed from the NMF based method for some nodes of this network. Also, since we do not know the optimal value of fuzzy parameter for FCM we report the obtained membership values from different fuzzy parameter for the same nodes in tables 2, 3, 4 and 5. Also tables 6 and 7 show the memberships from LRW method. In this method, t is a parameter for the number of random walk step which must be smaller than

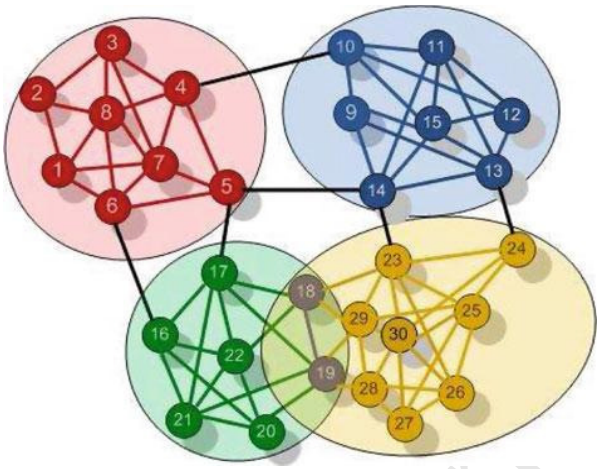


Figure 1: A simple social network from [16] with some multi-cluster nodes

Table 1: The memberships of some nodes in the network 1 by the NMF baed method

Node number	c_1	c_2	c_3	c_4
5	0.6052	0.1771	0.2177	0
8	1.0	0	0	0
14	0.0648	0.7891	0.0222	0.1238
15	0	1	0	0
18	0	0.0014	0.6672	0.3314
24	0	0.2429	0	0.7571

Table 2: The memberships of some nodes in the network 1 by the FCM with $f = 1.05$

Node number	c_1	c_2	c_3	c_4
5	1	0	0	0
8	1	0	0	0
14	0	1	0	0
15	0	1	0	0
18	0	0	1	0
24	0	0	0	0

the diameter of the network. The diameter of this network is 7, so we show two membership matrix in cases $t = 2$ and $t = 7$.

The results show that membership matrix produced by the NMF based method is better than FCM and LRW methods. Also, this method able to distinguish between nodes in the center of a cluster and nodes in the

Table 3: The memberships of some nodes in the network 1 by the FCM with $f = 1.1$

Node number	c_1	c_2	c_3	c_4
5	0.9950	0.0014	0.0033	0
8	1.0	0	0	0
14	0.00004	0 .9990	0.00002	0.00005
15	0	1	0	0
18	0	0.9970	0.2	0.0028
24	0.00005	0.0002	0.00004	0.9996

Table 4: The memberships of some nodes in the network 1 by the FCM with $f = 1.3$

Node number	c_1	c_2	c_3	c_4
5	0.6888	0.1075	0.1439	0.0598
8	0.9676	0.0108	0.0122	0.0094
14	0.0377	0.8936	0.0298	0.0389
15	0.0058	0.9853	0.0044	0.0045
18	0.0492	0.0396	0.7706	0.1406
24	0.0347	0.0649	0.0335	0.8669

Table 5: The memberships of some nodes in the network 1 by the FCM with $f = 1.5$

Node number	c_1	c_2	c_3	c_4
5	0.4866	0.1794	0.2113	0.1226
8	0.8070	0.0637	0.0708	0.0583
14	0.1249	0.6393	0.1098	0.1257
15	0.0461	0.8738	0.0404	0.0395
18	0.1212	0.1061	0.5496	0.2229
24	0.1051	0.1544	0.1050	0.6352

Table 6: The memberships of some nodes in the network 1 by the LRW with $t = 2$

node number	c_1	c_2	c_3	c_4
5	0.8372	0.0364	0.0988	0.0276
8	0.9023	0.0228	0.0576	0.0173
14	0.0106	0.9707	0.0081	0.0105
15	0.0211	0.9444	0.0153	0.0193
18	0.0777	0.0391	0.6522	0.2310
24	0.0615	0.1092	0.1132	0.7162

Table 7: The memberships of some nodes in the network 1 by the LRW with $t = 7$

node number	c_1	c_2	c_3	c_4
5	0.8055	0.0496	0.1038	0.0410
8	0.9191	0.0221	0.0414	0.0173
14	0.0079	0.9765	0.0063	0.0093
15	0.0151	0.9578	0.0112	0.0159
18	0.0504	0.0313	0.6822	0.2361
24	0.0569	0.1199	0.1203	0.7028

border or even in overlap well. In this method, the rows of a membership matrix for the nodes in the center of a cluster are crisp and for nodes in the border are fuzzy. Unlike NMF, FCM could not distinguish between nodes in the center of a cluster and nodes in the border. The FCM method for small fuzzy parameter gives crisp memberships even for nodes in the border of clusters, and by increasing this fuzzy parameter all memberships become fuzzy even for nodes in the center of clusters. To show the performance of the NMF based method we should have some criteria to evaluate the obtained membership matrix. Unfortunately, based on our knowledge, there are no general criteria in this case. These observations prompted us to research for methods to evaluate fuzzy clustering membership matrix.

4. Neighborhood-based evaluation criteria for fuzzy clustering

In this section, we describe two proposed criteria to evaluate fuzzy clustering membership matrix. These measures are made based on the neighborhood structure of nodes in the network.

Let us consider the simple network presented in Figure 1. In this network some nodes are in the center, and some others are on the border of clusters. For example, node 8 that all of its neighbors are in the cluster 1 is located at the center of this cluster and so its membership to this cluster should be near to 1. On the other hand node 14 has some connections with clusters 1, 2 and 4, but we expect that its membership to cluster 2 be more than other clusters, because most of its neighbors are in cluster 2. Also, the multi-cluster nodes 18 and 19 belong to two clusters 3 and 4, because their neighbors are in those clusters with the nearly same proportion.

The main point which is extracted from this example is that the probability of belonging one node to one cluster is related to probability of belonging

its neighbors to that cluster. So we expect that in a favorite fuzzy clustering, the memberships of a node to every cluster should be proportional to memberships of its neighbors to that cluster. We call this fact *Neighborhood influence*. In the next section, we introduce the two proposed criteria in supervised and unsupervised modes based on this fact.

4.1. Supervised fuzzy evaluation criterion based on neighborhood influence

In supervised mode, we have a network with ground truth, and we know the correct clusters. Now, we assign a membership matrix to this network, which its rows denote the favorite probabilities of belonging one node to all clusters. In this matrix based on Neighborhood influence, the probability of belonging one node to one cluster estimated by the average of the probability of belonging its neighbors to that cluster. This membership matrix distinguishes the nodes in the center and the nodes in the border of clusters. In a weighted network, the edges between the nodes show their similarities, and whatever the weight w_{ik} be larger, the effect of neighbors k in computing the memberships of node i is more.

Therefore, we define the probability of belonging one node to each cluster as the weighted average of the probabilities of belonging its neighbors to that cluster as bellow:

$$p_{ij}^{(s)} = \frac{\sum_{k \in n_i \cup \{i\}} w_{ik} p_{kj}}{\sum_{l \in n_i \cup \{i\}} w_{il}}, \tag{10}$$

where n_i is the set of neighbors of the node i and p_{kj} is defined as

$$p_{kj} = \begin{cases} \frac{1}{m_k} & k \in c_j, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Here m_k is the number of clusters that node k belongs them. In fact, we say that the probabilities of belonging one node to its different clusters(ground truth) are equal and become $1/m_k$. It is clear that for every i, $\sum_{j=1}^{c} p_{ij}^{(s)} = 1$ and so the matrix $P^{(s)} = \left(p_{ij}^{(s)}\right) \in \mathbb{R}^{n \times c}$ denote the ideal membership matrix consistent with the structure of the network. As an example, in Figure 1 by considering node 5 in two cluster 1 and 3 and nodes 18 and 19 in two clusters 3 and 4 the corresponding $P^{(s)}$ matrix is shown in Table 8. From Table 8, it is evident that matrix $P^{(s)}$ assign acceptable memberships to each node. Now, we expect that the fuzzy membership matrix becomes the same as the

Table 8: Matrix $P^{(s)}$ for the network in Figure 1

Node number	c_1	c_2	c_3	c_4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0.8520	0.0987	0.0493	0
5	0.5328	0.0655	0.4017	0
6	0.8520	0	0.1480	0
7	0.9551	0	0.0449	0
8	1	0	0	0
9	0	1	0	0
10	0.0987	0.9013	0	0
11	0	1	0	0
12	0	1	0	0
13	0	0.9102	0	0.0898
14	0.0449	0.8204	0.0449	0.0898
15	0	1	0	0
16	0.0987	0	0.9013	0
17	0.0449	0	0.8653	0.0898
18	0	0	0.5000	0.5000
19	0	0	0.5307	0.4693
20	0	0	0.9453	0.0547
21	0	0	0.9507	0.0493
22	0	0	0.9507	0.0493
23	0	0.0824	0.0412	0.8764
24	0	0.1095	0	0.8905
25	0	0	0	1
26	0	0	0	1
27	0	0	0	1
28	0	0	0.0493	0.9507
29	0	0	0.0898	0.9102
30	0	0	0	1

matrix $P^{(s)}$ and the quality of a fuzzy clustering method can evaluated by the closeness of its membership matrix to the favorite matrix $P^{(s)}$.

Comparing the fuzzy matrix U with matrix $P^{(s)}$ can be done in different ways. When the columns of U and $P^{(s)}$ denote the same clusters, as a simple method similarity between the corresponding rows of U and $P^{(s)}$ denote the amount of closeness of these two matrices.

For example, we perform our proposed NMF based fuzzy clustering, FCM, and LRW on the small network presented in Figure 1. The similarity between corresponding rows of $P^{(s)}$ and each fuzzy membership matrix U based on cosine similarity is shown in Figure 2.

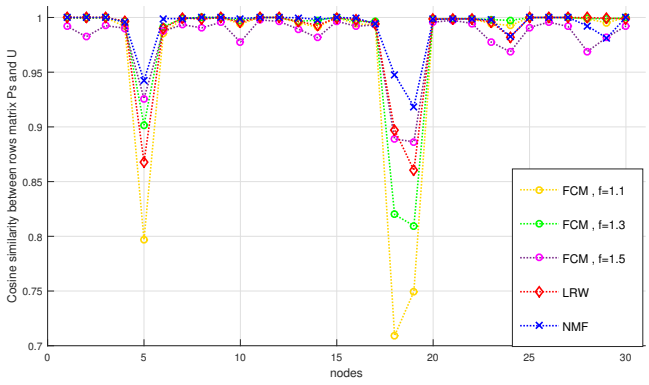


Figure 2: The cosine similarity between the expected membership matrix $P^{(s)}$ and the membership matrices provided by NMF based method and FCM and LRW for the network in Figure 1.

For all nodes in this diagram, the similarity between the membership matrix U from NMF and $P^{(s)}$ matrix is more than the others, especially, for the nodes in the overlap such as nodes 5, 18, 19 and the nodes in the border of clusters such as 10,14, 24. This way is not appropriate in general. Because comparing the various diagrams is not straightforward and the most significant problem is that the cluster labels exchanged and the columns of $P^{(s)}$ and U do not correspond to each other, and their reordering is not easy for large-scale social networks. To overcome this problem we introduce the following approach. For matrix $P^{(s)} \in \mathbb{R}^{n\times c}$, if $\mathsf{Simrows}(P^{(s)}) \in \mathbb{R}^{n\times n}$, denotes the similarity between rows of matrix $P^{(s)}$, for an appropriate fuzzy clustering with membership matrix $U \in \mathbb{R}^{n\times c}$, we expect that its corresponding $\mathsf{Simrows}(U) \in \mathbb{R}^{n\times n}$ be close to $\mathsf{Simrows}(P^{(s)})$. So, our supervised fuzzy

evaluation criterion (*SFEC*) can be defined as follows:

$$SFEC = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|(\mathsf{Simrows}(U))_{ij} - (\mathsf{Simrows}(P^{(s)}))_{ij}\right|. \qquad (12)$$

It is clear that *SFEC* doesn't depend on the ordering of the assigned labels to the clusters. Whatever *SFEC* be smaller, the fuzzy membership matrix U is more coincide with $P^{(s)}$ and is a better membership matrix. For example Table 9 shows the values of *SFEC* for the network introduced in Figure 1. Here, we used cosine similarity in construction of $\mathsf{Simrows}(P^{(s)})$ and $\mathsf{Simrows}(U)$. In this table, according to *SFEC* measure the memberships from NMF is better than the others, and LRW is better than FCM. Also U matrix from FCM with $f = 1.3$ is better than other FCM clustering results. These results confirm the Figure 2 and show the accuracy of *SFEC*. More tests will be presented in the experimental section.

Table 9: *SFEC* values for NMF and FCM with various fuzzy parameters

Algorithm	*NMF*	*FCM* $f = 1.1$	*FCM* $f = 1.2$	*FCM* $f = 1.3$	*FCM* $f = 1.4$	*FCM* $f = 1.5$	*LRW*
SFEC	**0.0395**	0.0712	0.0666	0.0600	0.0794	0.1460	0.0576

4.2. Unsupervised evaluation criterion based on neighborhood influence

In many real networks, we don't have any information about the community structure and correct clusters, so evaluation of membership matrices produced by different fuzzy clustering algorithms is very significant and difficult. Two most important aspects that should be considered in every clustering algorithm are the intra-cluster *compactness* and the inter-cluster *separation*. The first means the nodes of one cluster should have more connections with each other and the second means that different clusters should be separated and have few connections with each other.

In this section, based on the Neighborhood Influence and mentioned compactness and separation, we propose an unsupervised evaluation method that only uses an adjacency matrix of a network.

Consider the membership matrix $U \in \mathbb{R}^{n\times c}$, produced by an appropriate fuzzy clustering method. Based on the Neighborhood influence fact, we expect that in a good clustering algorithm, the behavior of one node should

be consistent with the behavior of its neighbors. For example, the algorithm should not assign one node to one cluster when based on this clustering algorithm none of its neighbors are not in that cluster. So, for an appropriate fuzzy clustering method, the membership of node i in cluster j , i.e., u_{ij} should be proportional to weighted average membership

$$p_{ij}^{(u)} = \frac{\sum_{k \in n_i} w_{ik} \times u_{kj}}{\sum_{l \in n_i} w_{il}}. \tag{13}$$

of its neighbors in the cluster j. Therefore, in a good fuzzy clustering method,

$$\sum_{i=1}^{n} \sum_{j=1}^{c} |u_{ij} - p_{ij}^{(u)}|, \tag{14}$$

should be small. From another view, this equation denotes the compactness of the clustering. By replacing (13) in (14) we have

$$\sum_{i=1}^{n} \sum_{k \in n_i} \sum_{j=1}^{c} \frac{w_{ik}}{\sum_{l=1}^{c} w_{il}} |u_{ij} - u_{kj}| = \sum_{i=1}^{n} \sum_{k \in n_i} \frac{w_{ik}}{\sum_{l=1}^{c} w_{il}} \|\mathbf{u}'_i - \mathbf{u}'_k\|_1 \tag{15}$$

where $\mathbf{u}'_i$ is the i-th row of U and denotes the membership values of node i.

Note that in $\|\mathbf{u}'_i - \mathbf{u}'_k\|_1$ nodes i and k are adjacent so small value of (15) emphasis that the memberships of adjacent nodes to different clusters should be similar. Therefore, (15), can be used as a measure for compactness of clusters.

The relation (14) is a summation of $n \times c$ numbers and so for small values of c the result always is less. So we should correct our criterion to be able to compare between different membership matrices with various numbers of clusters. The following normalization can do this correction

$$comp = \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} |u_{ij} - p_{ij}^{(u)}|}{nc} \tag{16}$$

Although *comp* can be used to compare the compactness of membership matrices well, but it does not consider the separation of clusters. So, the closeness of U and $P^{(u)} = (P_{ij}^{(u)}) \in \mathbb{R}^{n \times c}$ is not sufficient for verification of U matrix. As mentioned in 2.2, in the fuzzy membership matrix U from FCM with $f > \delta$ all components u_{ik} are similar and equal to $\frac{1}{c}$. In this case from

definition of $P^{(u)}$, for every i , j we have $P_{ij}^{(u)} = \frac{1}{c}$ and so *comp* becomes zero which is the minimum value. But it is clear that the obtained membership matrix cannot detect cluster of nodes and all the clusters overlap with each other entirely.

To complete our criterion we add a separation measure which controls the overlap between clusters. In this case, the distance between two clusters with the most overlap can be an appropriate measure for separation. So we define the separation of a fuzzy c-partition as follows,

$$sep = \min_{i,j} d(\mathbf{u}_i, \mathbf{u}_j). \tag{17}$$

where $d(.)$ could be any appropriate distance function. In fact *sep* is the distance between two columns of U with minimum distance, which is equal to the distance of clusters with the maximum overlap. So, large value of *sep* denotes that the minimum distance of obtained clusters is large and so they are more separated.

For an ideal fuzzy clustering, our defined quantities *comp* and *sep* should be small and large, respectively. Therefore the combination of them can be used to set a sophisticated validity criterion as follows:

$$UFEC = \frac{comp}{sep} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{c}|u_{ij} - p_{ij}^{(u)}|}{n \times c \times \min_{i,j} d(\mathbf{u}_i, \mathbf{u}_j)}, \tag{18}$$

that will be named Unsupervised Fuzzy Evaluation Criterion (UFEC). A smaller value for *UFEC* indicates a good cohesion within clusters and a little overlap between pairs of clusters. Therefore, it is an appropriate criterion for evaluation of membership matrices.

We performed FCM with various fuzzy parameters on the network in Figure 1 to see the performance of our proposed evaluation criteria. Table 10 shows the results. In this table the best fuzzy parameter corresponding to the minimum value of *UFEC* is $f = 1.4$ which is near to the values of *SFEC* in the Table 9. As it is explained in section 2.2 by increasing the fuzzy parameter f, U will be very fuzzy and so is useless. In this case, we expect that UFEC is large which is consistent with the results in Table 10. Therefore our proposed criterion can detect an appropriate fuzzy partition matrix that is not very crisp or very fuzzy. Also, we used our NMF based fuzzy clustering on this network. Here the obtained values for *comp*, *sep*, *UFEC* are 0.0375, 0.8777 and 0.0427 respectively. The comparison of these results with the values in Table 10 shows the superiority of the proposed NMF based fuzzy clustering method.

Table 10: The values of *comp*, *sep* and *UFEC* measure for network 1 by FCM with different fuzzy parameters

$fuzzy$ $parameter$	*comp*	*sep*	*UFEC*
1.05	0.0612	1	0.0612
1.08	0.0611	0.9999	0.0611
1.1	0.0611	0.9995	0.0611
1.2	0.0588	0.9823	0.0598
1.3	0.0533	0.9310	0.0572
1.4	0.0466	0.8514	**0.0548**
1.5	0.0413	0.7458	0.0554
1.6	0.0371	0.6048	0.0613
1.63	0.0295	0.4170	4.7957
1.65	0.0088	0.1306	24.6670
1.7	8.6058e-05	1.9855e-06	163.0747

5. Experimental results

In this section, we present our experimental results on some artificial network from LFR benchmark [41] and some well known real-world networks with overlapping clusters. All Experiments have been done with MATLAB in a system with 2.4GHz cpu and 16 Gb RAM. To evaluate different fuzzy clustering methods on the networks with ground truth, we use the fuzzy NMI [19], and Omega-Index (OI) [20]. Also for the networks without ground truth we use modularity measure (Q_{ov}) [22, 6]. The parameters of Q_{ov} are assigned as[6]. All of these measures are between 0 and 1, and the maximum values of them are ideal. Also, to verify SFEC and UFEC measures we applied them on networks with and without ground truth, respectively. For these measures, we used the cosine distance as a distance function. In experiments, the NMF based method, FCM and local random walk (LRW) [8] algorithms have been used for fuzzy clustering. The LRW algorithm depends on the number of random walk step ($\mathbf{t}$ parameter) which, according to the paper [8] it must be less than the diameter of the network. All, NMF and FCM and so LRW are non-convex problems and maybe give local optimal. Therefore, to overcome this issue in each case we executed them many times (about 50 iterations) and reported the mean value of them. According to these measures the overlapped communities or memberships produced by the NMF are better than the other algorithms.

5.1. Tests on networks with ground truth

In this section we present the experimental results on some networks with ground truth.

A simple test network

Consider the simple network presented in the Figure 3, with two main clusters and one node in the overlap of these clusters. For this network,

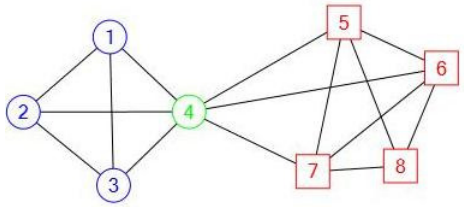


Figure 3: A sample network with node 4 in the overlap.

the fuzzy membership matrix from NMF, FCM with some different fuzzy parameters and LRW has been presented in Table 11. In this table, the

Table 11: The fuzzy partition matrix from NMF and FCM with various fuzzy parameters for sample network in Figure 3.

$FCM, f = 1.1$	$FCM, f = 1.5$	$FCM, f = 1.9$	NMF	LRW
$\begin{pmatrix} 1.0 & 0 \\ 1.0 & 0 \\ 1.0 & 0 \\ 0.9966 & 0.0034 \\ 0 & 1.0 \\ 0 & 1.0 \\ 0 & 1.0 \\ 0 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 0.9669 & 0.03306 \\ 0.9669 & 0.0330 \\ 0.9669 & 0.0330 \\ 0.5705 & 0.4295 \\ 0.04789 & 0.9521 \\ 0.04789 & 0.9521 \\ 0.0478 & 0.9521 \\ 0.0647 & 0.9352 \end{pmatrix}$	$\begin{pmatrix} 0.861 & 0.139 \\ 0.861 & 0.139 \\ 0.861 & 0.139 \\ 0.5302 & 0.4698 \\ 0.1705 & 0.8295 \\ 0.1705 & 0.8295 \\ 0.1705 & 0.8295 \\ 0.1942 & 0.8058 \end{pmatrix}$	$\begin{pmatrix} 1.0 & 0 \\ 1.0 & 0 \\ 1.0 & 0 \\ 0.5805 & 0.4195 \\ 0.1095 & 0.8905 \\ 0.1095 & 0.8905 \\ 0.1095 & 0.8905 \\ 0 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 0.9923 & 0.0077 \\ 0.9923 & 0.0077 \\ 0.9923 & 0.0077 \\ 0.4098 & 0.5902 \\ 0.0056 & 0.9944 \\ 0.0056 & 0.9944 \\ 0.0056 & 0.9944 \\ 0.4815 & 0.5185 \end{pmatrix}$
$UFEC = 0.1561$ $Q_{ov} = 0.6750$ $SFEC = 0.1641$ $NMI = 0.78$ $OI = 0.7742$	$UFEC = 0.1305$ $Q_{ov} = 0.6750$ $SFEC = 0.0374$ $NMI = 1$ $OI = 1$	$UFEC = 0.1424$ $Q_{ov} = 0.6735$ $SFEC = 0.1027$ $NMI = 1$ $OI = 1$	$UFEC = 0.1221$ $Q_{ov} = 0.6750$ $SFEC = 0.0306$ $NMI = 1$ $OI = 1$	$UFEC = 0.2668$ $Q_{ov} = 0.5728$ $SFEC = 0.1520$ $NMI = 0.78$ $OI = 0.6522$

memberships from FCM for small values of f are more crisp, even for the nodes in overlap such as node 4, and with increasing f, all the rows of U become fuzzy even for other single-cluster nodes such as node 8. But the rows of produced membership based on NMF for overlapping nodes such as 4 are fuzzy and for other single-cluster nodes are crisp. In memberships from LRW, the single-cluster node 8 has fuzzy memberships which are not correct. We evaluate each membership matrix with the mentioned measures. If we

don't use the known clusters (unsupervised mode) and use UFEC and Q_{ov}, one can see that Q_{ov} can not distinguish between memberships from NMF and FCM with $f = 1.1$ and $f = 1.5$ but UFEC in the best membership matrix is minimum. On the other hand, if we use the known cluster and

Table 12: The membership matrix $P^{(s)}$ for network 3

node number	c_1	c_2
1	0.9259	0.0741
2	0.9259	0.0741
3	0.9259	0.0741
4	0.5000	0.5000
5	0.0645	0.9355
6	0.0645	0.9355
7	0.0645	0.9355
8	0	1

use SFEC, NMI, and Omega-Index, according to the results, NMI and OI for NMF and FCM with $f = 1.5$ and $f = 1.9$ are equal to 1 (the maximum value) and don't distinguish between these memberships. It is because that these measures are not based on all produced memberships and depend on the cluster labels assigned to each node. But SFEC can detect the best membership matrix well. Table 12 shows the $P^{(s)}$ matrix of this network.

The LFR benchmark

In this section, we use artificial networks obtained from the public LFR benchmark network algorithm [1] [42]. The networks produced by this algorithm have the main features of all the social networks such as power-law distribution and community structure[43]. We used networks produced by this benchmark with 500 and 1000-node with the various percent of overlap. For example, we consider a network with parameters $N = 500$, $k = 15$, $\mu_t = 0.3$, and $\mu_w = 0.2$ according to [42] which 10 percent of its nodes are in overlap. Here N , k denote the number of nodes and the average degree of nodes in the network. Also the parameters μ_t, and μ_w are mixing parameters for the topology and weights, respectively [42]. Table 13 shows the results for this network. For another network with these parameters, with 30 percent of nodes in the overlap, the results reported in Table 14.

[1] https://sites.google.com/site/santofortunato/inthepress2

Table 13: The comparison of various measures for a 500-node LFR type network which 10 percent of nodes in the overlap.

Algorithm	NMI	OI	$SFEC$	$UFEC$	Q_{ov}
NMF	**0.9775**	**0.9858**	**0.0194**	**0.0288**	0.5658
FCM ,$f = 1.043$	0.8050	0.5479	0.0573	0.0380	**0.6022**
LRW	0.7166	0.5462	0.0736	0.0452	0.5714

Table 14: The comparison of various measures for a 500-node LFR type network which 30 percent of nodes in the overlap.

Algorithm	NMI	OI	$SFEC$	$UFEC$	Q_{ov}
NMF	**0.8305**	**0.8471**	**0.0512**	**0.0393**	0.3499
FCM ,$f = 1.035$	0.5781	0.5325	0.1354	0.0637	**0.4319**
LRW	0.5241	0.4732	0.4412	0.0759	0.3891

We performed FCM with various f and shows the best result. One can see that the values of measure NMI, Omega-Index, SFEC, and UFEC are proportional and these measures verify each other. It is clear that according to these measures, memberships from the NMF based method are better than others. But Q_{ov} measure has little different behavior. To compare the fuzzy clustering methods and also show the reliability of the proposed validity measures, we run these methods on the LFR benchmark networks with the same parameters as mentioned above. Figure 4 shows the results.

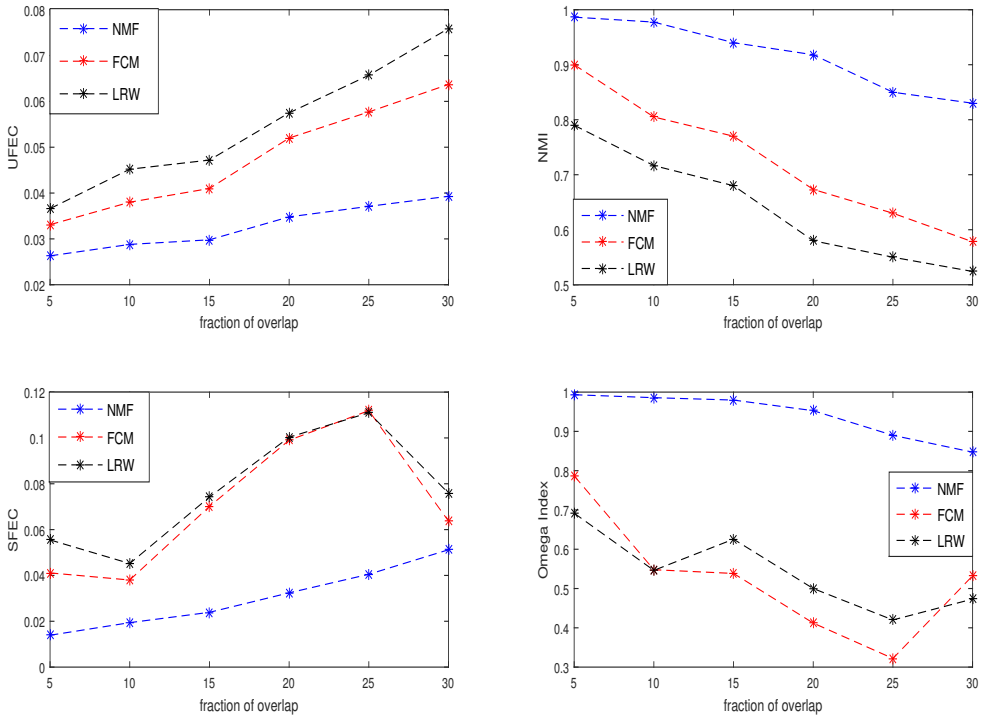


Figure 4: Average results (NMI, Omega Index, SFEC , UFEC) of NMF, FCM and LRW algorithms on the LFR networks with fuzzy overlapping.

In all these figures the results from the NMF based algorithm have the best values according to all measures.

5.2. Tests on real world social networks

In this section, we consider some well known networks, which their information can be seen in Table 15.

Table 15: Real world networks for evaluation

network	*nodes*	*edges*
Zachary	34	78
Dolphin	62	159
Jaz	198	2742
Football	115	613
Facebook	4039	88234
Youtube	19017	119470

Zachary Karate Club network

The famous karate club network, which is analyzed by Zachary widely is used as a test example for community detection methods in complex networks [44]. This network consists of 34 members of a karate club as nodes and 78 edges representing the friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and the clubs instructor, the club is split into two smaller ones. This network is shown in Figure 5.

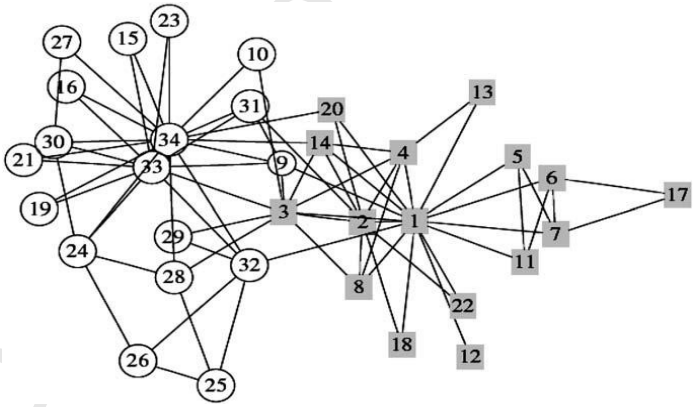


Figure 5: Zachary karate club network [45].

Figure 6 shows the membership matrices produced with FCM by different fuzzy parameters and also membership matrices of NMF and LRW methods.

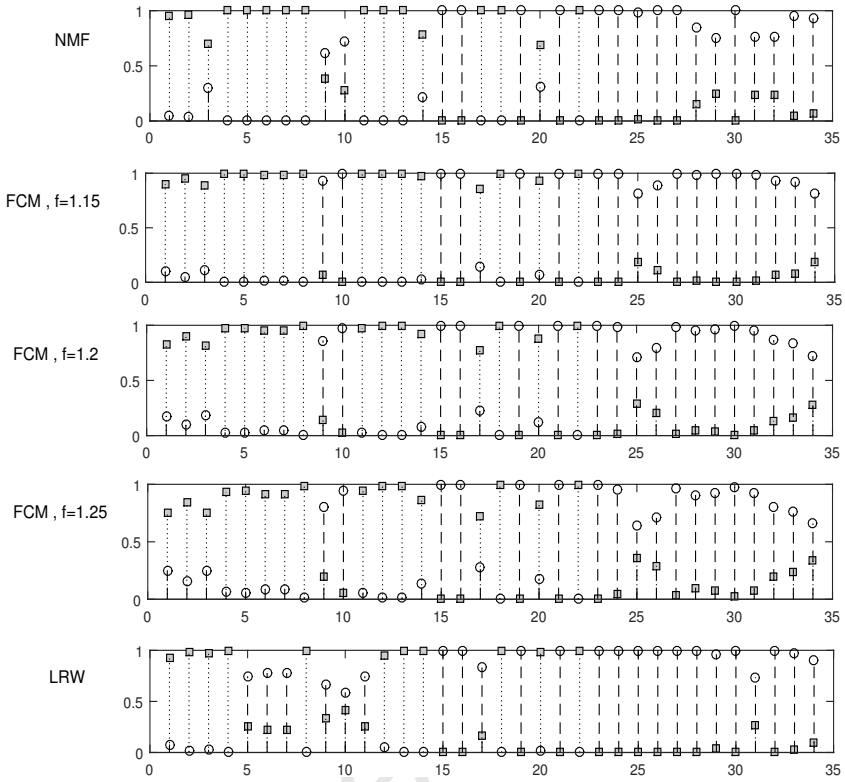


Figure 6: Memberships of Zachary network's nodes to clusters 1 and 2. In these diagrams, the horizontal axis is proportional with the nodes and for each node, its memberships to clusters 1 and 2, are shown by the circle and the square, respectively.

From this figure, it's clear that the produced membership matrix by NMF can distinguish overlapping and non-overlapping nodes. It means that for the single-cluster nodes the assigned memberships are crisp, but for the nodes in the overlaps, the assigned memberships are fuzzy. In the diagrams, we see that when the fuzzy parameter f is small ($f = 1.15$) the memberships by FCM are crisp and non-fuzzy, so the multi-cluster nodes can not be recognized. Also for larger fuzzy parameters, the memberships become fuzzy, which are not correct for single cluster nodes. For example in the third diagram ($FCM, f = 1.20$)the memberships of the single cluster nodes 17 and 34 are fuzzy, but from Figure 5, these nodes are not in the overlap of the clusters. Also when $f = 1.25$, many single-cluster nodes incorrectly clustered in the overlap of the clusters such as nodes 1, 2, 25, 26, 33 and 34 and their assigned memberships are not consistent with the network's structure.

Table 16: The values of *comp*, *sep*, *UFEC* and Q_{ov} measures for Zachary network.

Algorithm	*comp*	*sep*	*UFEC*	Q_{ov}
NMF	0.0590	0.8740	**0.0675**	0.7455
FCM , $f = 1.15$	0.1411	0.9269	0.1522	0.7455
FCM , $f = 1.2$	0.1697	0.8623	0.1968	0.7455
FCM , $f = 1.25$	0.1888	0.7937	0.2378	0.7455
LRW	0.1171	0.8722	0.1343	0.6318

In the diagram corresponding to LRW method, the memberships are better than FCM, but the nodes 5, 6 and 17 have fuzzy memberships incorrectly. Also, node 3 has the crisp membership to cluster 1, but it is a multi-cluster node and is in the overlap. Based on Figure 5 and the above description it is clear that the memberships produced by the NMF based method are more natural than FCM's and LRW's memberships. Also, LRW is better than FCM. Therefore, we expect that the $UFEC$ value for the membership matrix produced by the NMF be smaller than the others, which Table 16 verify this expectation.But Q_{ov} could not identify the favorite membership matrix. Here, the value of Q_{ov} for different memberships are equal and for LRW is less than the others which is not proportional with their memberships from Figure 6.

Dolphin social network

Another well known real world network is *dolphin social network* which is a social network of 62 bottlenose dolphins living in Doubtful Sound, New Zealand [46, 47]. The network naturally divided into two groups, represented by the squares and circles in Figure 7. But, this figure shows that the set of right-hand side nodes (circles) can be split into three subclasses, and the number of clusters can be 4. In the case of $c = 2$ the number of overlapping nodes is little but when $c = 4$, there are more overlapping nodes in the right-hand side clusters. So, according to the section 2.2 we expect that for $c = 2$ the best fuzzy parameter f be smaller than the case $c = 4$.

For testing *UFEC* we perform FCM with various fuzzy parameters, and for each parameter, we set the number of clusters to 2 and 4. The values of *UFEC* for all cases are shown in Table 17.

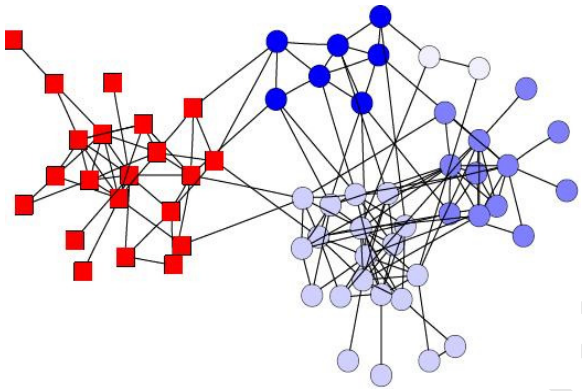


Figure 7: Dolphin social network [45].

Table 17: The values of UFEC for FCM with various fuzzy parameters, for $c = \{2, 4\}$ on Dolphin network

f	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.1	1.11	1.12
c=2	**0.0433**	0.603	0.0599	0.0588	0.0574	0.0558	0.0544	0.0529	0.0512	0.0516	0.0562	0.2372
c=4	0.1009	0.0995	**0.0994**	0.1024	0.1025	0.1026	0.1027	0.1402	0.1414	0.1669	0.2236	0.2224

In each row, the best case is bolded. It can be seen that for $c = 2$ the best fuzzy membership matrix of FCM, based on the *UFEC* measure occurs in $f = 1.01$ and for $c = 4$ it occurs in $f = 1.03$ which is coinciding with our expectation.

Jazz musician network

The network of collaborations between early jazz musicians of Gleiser and Danon [48] from the Red Hat Jazz Archive has 198 nodes and 2742 edges [2].

A link between two nodes means that they have at least one musician in common. Figure 8 that is drawn in Gephi software [3] show this network that is divided into two main clusters.

[2]https://www.infochimps.com/datasets/jazz-musicians-network

[3]https://gephi.org/

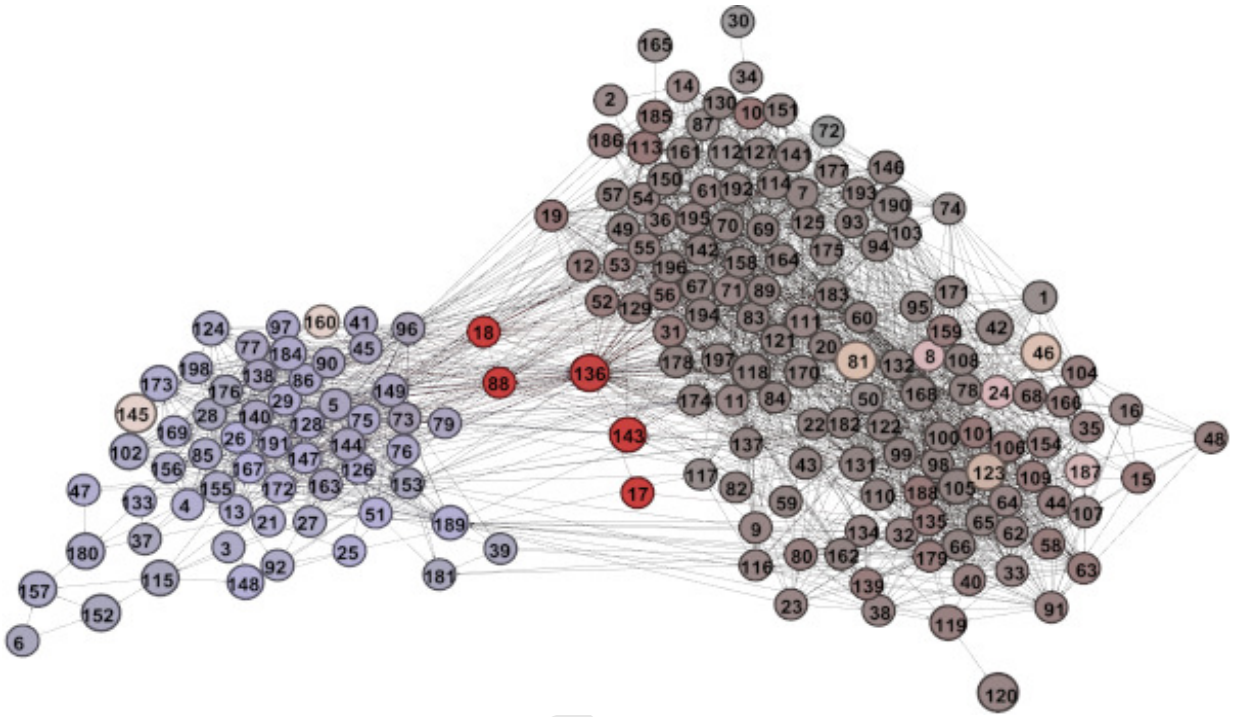
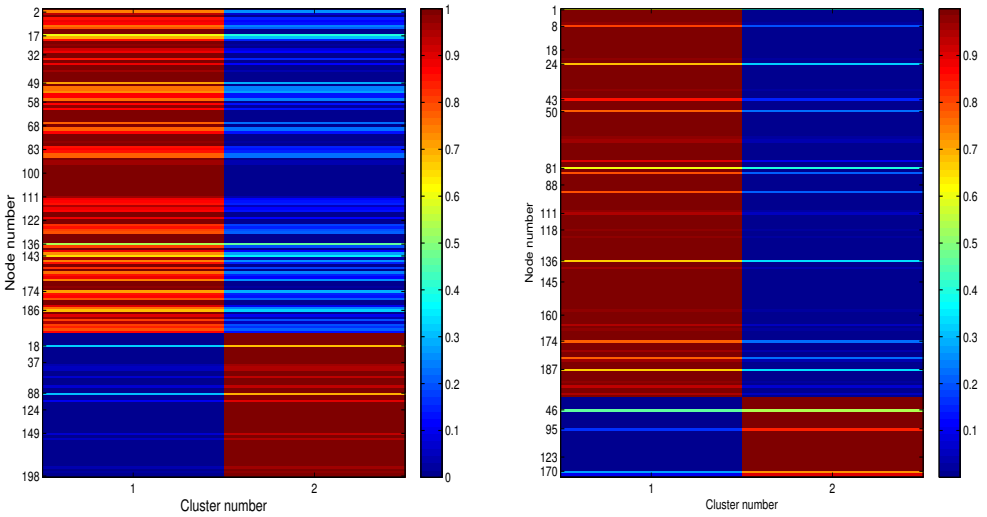


Figure 8: Jazz Musicians Network

For this network, we perform NMF and FCM with $f = 1.1$ and show their produced U fuzzy membership matrices in Figure 9. In each of these matrices, we sort the rows in a way that the nodes in the same cluster are adjacent. According to Figure 8, some node such as 17, 18, 88, 143 and 136 are in the overlap of the two clusters.

In Figure 9a the memberships assigned by the NMF based method to these mentioned nodes are fuzzy. But in Figure 9b by FCM, some single-cluster nodes such as 46 and 81 have fuzzy memberships, but the multi-cluster nodes 18 and 88 have quietly crisp memberships. Also in this figure some nodes such as 123 and 145 assigned to incorrect clusters. Table 18 shows The memberships of some nodes.

In table 18, Although the results of LRW are better than FCM, some node like 143 and 123 have crisp and fuzzy memberships incorrectly. These figures and Table 18 demonstrate that the memberships of the NMF based method are consistent with the structure of the network and are much better than FCM and LRW methods. So we expect that the values of *UFEC* for NMF be smaller than the others. Table 19 shows the values of *UFEC* and



(a) NMF's Fuzzy membership matrix (b) FCM's Fuzzy membership matrix

Figure 9: Fuzzy membership matrix from NMF and FCM with $f = 1.1$ for Jazz musician network.

Table 18: The membership produced by NMF, LRW and FCM with different fuzzy parameters for Jazz musician network.

	NMF		FCM		LRW	
node number	$cluster1$	$cluster2$	$cluster1$	$cluster2$	$cluster1$	$cluster2$
17	0.6025	0.3975	1	0	0.7291	0.2709
18	0.3260	0.6740	0.9992	0.0008	0.4611	0.5389
46	1	0	0.4639	0.5361	0.7072	0.2928
81	0.9972	0.0028	0.5980	0.4020	0.9295	0.0705
88	0.3018	0.6982	0.9981	0.0019	0.1708	0.8292
123	1	0	0	1	0.6020	0.3980
136	0.5399	0.4601	0.6659	0.3341	0.5696	0.4304
143	0.6505	0.3495	0.9997	0.0003	0.8657	0.1343
145	0	1	1	0	0.01	0.99
160	0	1	1	0	0.0363	0.9637

Q_{ov} measures. Here it's clear that the result of the NMF based method is better than FCM, which confirms our results from the figures 8, 9a,9b.

Table 19: The values of *comp*, *sep* in *UFEC* measure and Q_{ov} on Jazz network.

Algorithm	*comp*	*sep*	*UFEC*	Q_{ov}
NMF	0.0476	0.8643	0.0550	0.6521
FCM	0.1597	0.9478	0.1685	0.4424
LRW	0.0813	0.6397	0.1271	0.6518

The network of American college football teams

The network of American football games between division IA teams during the regular season Fall 2000 [4] [1]. There are 115 vertices in this graph that represent teams and edges represent regular season games between the two teams that they connect. The network incorporates a public community structure since the teams are divided into 12 conferences.

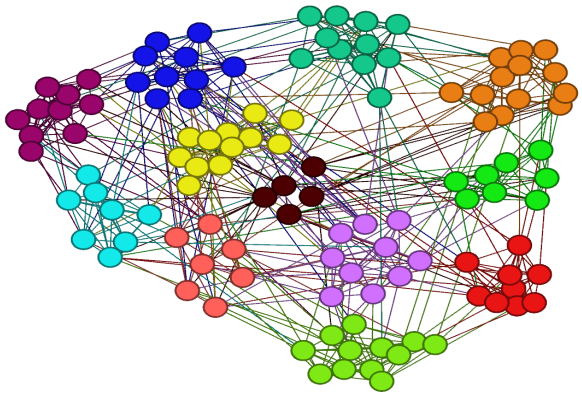


Figure 10: The network of American football team

[4]http://www-personal.umich.edu/ mejn/netdata/

Table 20: The values of *UFEC* and Q_{ov} for the NMF based method, LRW and FCM with various fuzzy parameters on Football network

Algorithm	NMF	FCM $f=1.01$	FCM $f=1.03$	FCM $f=1.05$	FCM $f=1.07$	FCM $f=1.09$	FCM $f=1.13$	FCM $f=1.17$	LRW
$UFEC$	**0.0428**	0.0539	0.0519	**0.0516**	0.0521	0.0523	0.0536	0.0540	0.0871
Q_{ov}	0.6715	0.6688	0.6892	**0.6892**	0.6892	0.6892	0.6844	0.6730	0.45

Table 20 shows the values of UFEC and Q_{ov} of NMF base fuzzy clustering, FCM and LRW methods on this data set.

According to this table the minimum value of UFEC occurs in NMF case and for FCM the best value of UFEC is obtained for $f = 1.05$. .

Facebook social network

Facebook social network is one of the popular social networks. We get this network's data with 4039 nodes and 88234 edges from SNAP cite [5] that its diagram is shown in 11. As it can be seen from Figure 11 the network is divided into 7 large clusters. Also, the number of clusters can be estimated by the number of eigenvalues of Laplacian matrix that are near to zero and have a significant difference with other eigenvalues.[49]

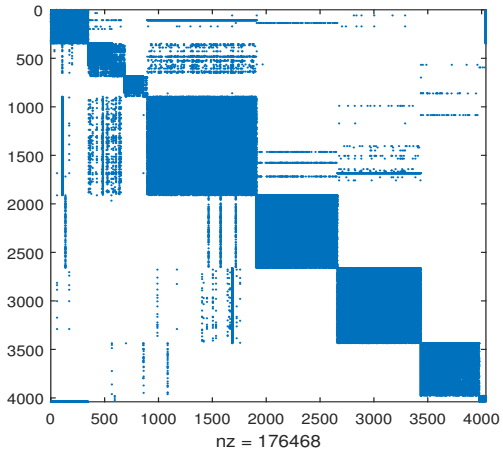


Figure 11: Facebook social network with 4039 nodes

[5] http://snap.stanford.edu/data/egonets-Facebook.html

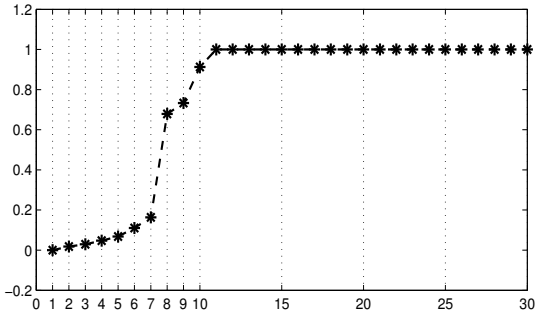


Figure 12: The 30 smallest eigenvalues of Laplacian matrix for Facebook network

In Figure 12 which show the first 30 minimum eigenvalues of Laplacian matrix of the Facebook network, the number of main clusters are 7 that coincides with the 7 large clusters in Figure 11. The results of UFEC, Q_{ov} and Time on this network are shown in 15, 16, 17. According to these diagrams, for the Facebook network also the best values of Time, UFEC and Q_{ov} occur in NMF based fuzzy clustering algorithm.

Youtube social network

Youtube is a video-sharing website that includes a social network [6]. In the Youtube social network, users form friendship each other. For simplicity and memory restrictions, we removed the ground-truth communities which have less than 70 nodes and so the network reduced to the network with 19017 nodes. Figure 13 shows this network.

[6]https://snap.stanford.edu/data/com-Youtube.html

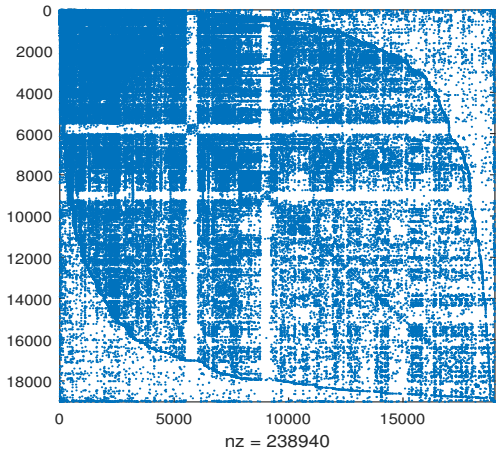


Figure 13: Youtube social network with 19019 nodes

To estimate the number of clusters we use the Laplacian matrix.

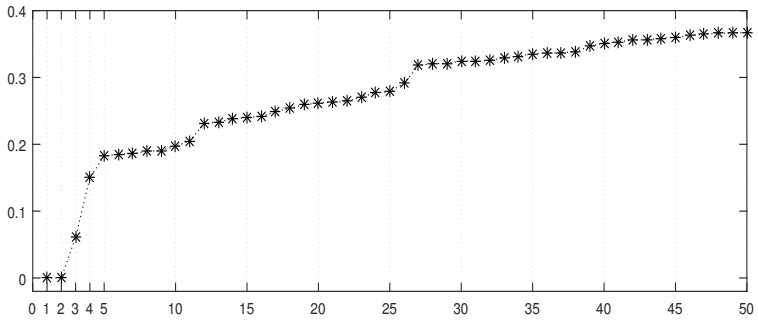


Figure 14: The 50 smallest eigenvalues of Laplacian matrix for Youtube network

In 14, which shows the first 50 minimum eigenvalues of Laplacian matrix of Youtube network, the number of eigenvalues near zero is 2 and a significant difference between the eigenvalues occurs in 3. So, the number of large clusters of this network can be estimated as 2 with less overlap and as 3 with more overlaps. So, we set the number of clusters to 3. The results of UFEC and Q_{ov} and time for mentioned network are shown in Figures 15, 16, 17, respectively. From these figures the quality of NMF for yotube network is more than the others.

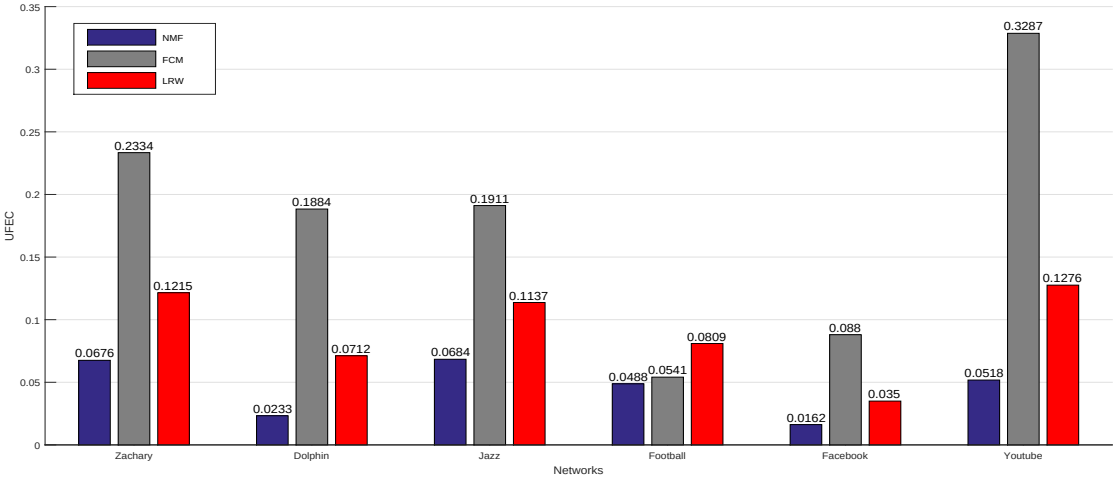


Figure 15: the mean value of UFEC for some real world networks which minimum values of UFEC are ideal.

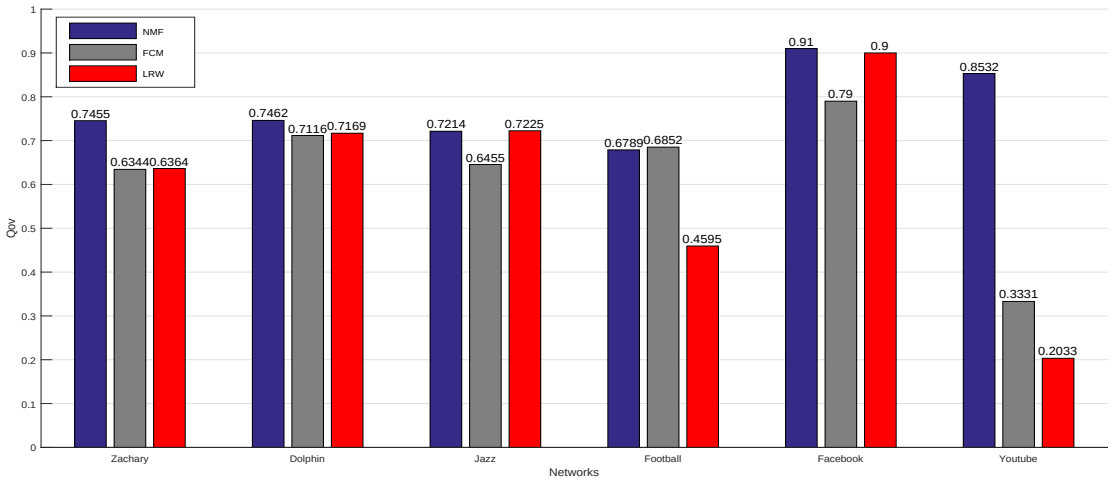


Figure 16: the mean value of Q_{ov} for some real world networks which maximum values of Q_{ov} are ideal.

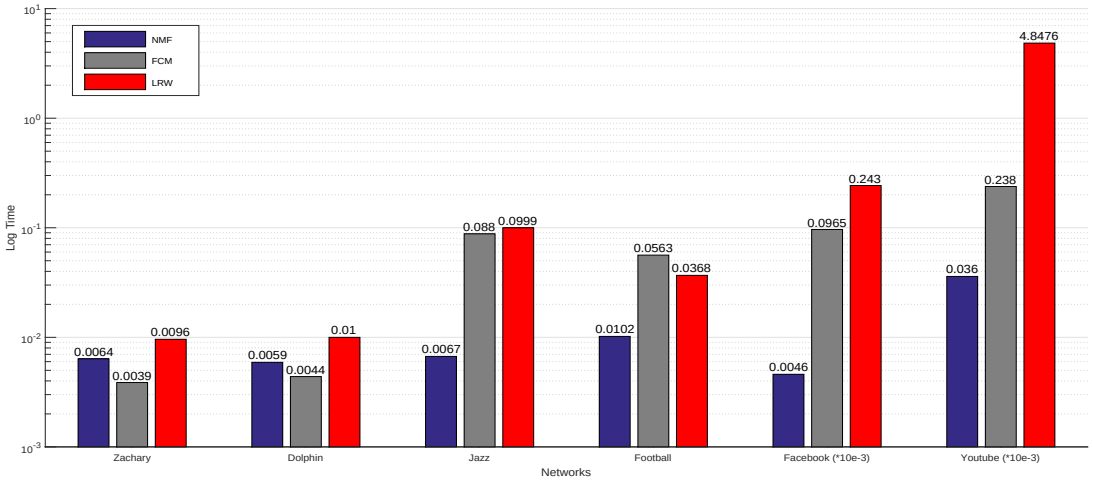


Figure 17: the mean value of running time for some real world networks which minimum values of Time are ideal.

In the end, to have an overall view about the quality of these methods on real networks, we summarize the results of all mentioned networks based on UFEC, Q_{ov} and consuming time in the figures15, 16 and 17, respectively. These diagrams show the superiority of NMF base method to do fuzzy clustering on overlapped networks. According to Figure 17 by increasing the size of the networks the distance between running times of NMF based method and others increase. Here NMF is faster than the others. Also, NMF based method does not depend on any parameter more than the number of the clusters. But for FCM and LRW we need some parameters that should be set manually for each class. In the experiments we found that the used memory by NMF is less than the others and LRW used the most memory.

5.3. Comparison between different overlapping community detection Algorithm

In this section, we compare the NMF based method with some other overlapping community detection algorithms based on Q_{ov}. All parameters of this measure are assigned as [6]. The results are shown in table 21. In addition to the mentioned networks the email network [7] with 1133 node is

[7]http://deim.urv.cat/ alexandre.arenas/data/welcome.htm

Table 21: Comparison between different fuzzy clustering algorithms with Q_{ov} on some real world network.

Network	*NMF*	*FCM*	*LRW*	*SLPA*	*Copra*
Zachary Karate club	**0.74**	0.6344	0.6364	0.65	0.44
Dolphin	**0.74**	0.7116	0.7169	0.76	0.70
Jazz	**0.7214**	0.64	0.7225	0.70	0.71
Football	0.67	0.68	0.45	**0.70**	0.69
Email	**0.67**	0.49	0.59	0.64	0.51

added. Also algorithms SLPA [7] and COPRA (Community Overlap Propagation Algorithm) [6] which are presented as methods with more better performance in [4], are compared.

6. Conclusion

In this paper, we proposed a novel NMF based fuzzy clustering method and demonstrates that the produced membership matrix is better than the other well-known fuzzy clustering methods. This means that its obtained memberships are consistent with the structure of the network. Also, the NMF based method is faster than the others and for large scale networks, this superiority becomes meaningful. Results on large scale networks like Facebook and YouTube confirm the quality of the proposed method speed and quality of the produced membership matrices. Although in fuzzy clustering method the quality of assigned memberships to different clusters are very important but unfortunately, there is not a fuzzy evaluation method based on the evaluation of membership matrix. So, we proposed two evaluation criteria for quantifying fuzzy membership matrix in networks with and without ground truth, that named *SFEC* and *UFEC*, respectively. Experimental results on well-known networks demonstrate the power of the proposed criteria to recognize the quality of the membership matrices in comparison with the other methods like Q_{ov}. Based on these proposed criteria and other evaluation methods we found that the quality of the fuzzy clustering method based on NMF, especially for large scale social networks is more than the others.

References

[1] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences

USA, 99, (2002),pp. 7821-7826.

[2] G. Palla, et al., Uncovering the overlapping community structure of complex networks in nature and society, Nature, 435,(2005) pp. 814-818.

[3] S. Fortunato, Community detection in graphs, Physics Reports, 486,(2010), pp. 75-174.

[4] J. Xie, et al, Overlapping community detection in networks: the state of the art and comparative study, ACM Computing Surveys (CSUR), 45,(2013), p. 1-35.

[5] X. Qi et al, Optimal local community detection in social networks based on density drop of subgraphs, Pattern Recognition Letters, 36, (2014), pp. 46-53.

[6] S. Gregory, Finding overlapping communities in networks by label propagation. New J.Phys. 12,(2010) 103018.

[7] J. Xie, B. K. Szymanski,and X. Liu, SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In Proc. ICDM Workshop. pp. 344349,2011.

[8] W . Wang , et al, Fuzzy overlapping community detection based on local random walk and multidimensional scaling, Physica A-statistical Mechanics and Its Applications,392,(2013),pp. 6578-6586.

[9] J. C .Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, (1973).

[10] J. C. Bezdek, Pattern Recognition with Fuzzy Object Function Algorithms, Plenum Press, NewYork, (1981).

[11] J. C. Bezdek, et al,FCM: The fuzzy C-means clustering algorithm, Computers and Geosciences, 10, (1984), pp. 191-203.

[12] S. Zhang, R. Wang and, X. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A, 374,(2007), pp. 483-490.

[13] T. Nepusz, A. Petroczi, L. Negyessy, and F. Bazso, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. 77, (2008) 016107.

[14] J. Xie, and B. K. Szymanski, Towards linear time overlapping community detection in social networks. In Proc. PAKDD Conf. pp. 25–36, 2012.

[15] I. Psorakis, et al., Overlapping community detection using bayesian nonnegative matrix factorization, Physical Review E, vol. 83, (2011) 066114.

[16] F. Wang, et al., Community discovery using nonnegative matrix factorization," Data Mining and Knowledge Discovery, vol. 22, (2011), pp. 493-521.

[17] R.-S. Wang, et al., Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures, Neurocomputing, vol. 72, (2008) , pp. 134-141.

[18] S. Zhang, et al., Uncovering fuzzy community structure in complex networks,Physical Review E, vol. 76, (2007), 046103.

[19] A. Lancichinetti, et al., Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics, vol. 11, (2009),pp. 033015.

[20] L. M. Collins, C. W. Dent, Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. Multivar. Behav. Res. 23, 2 (Feb.), (1988), pp. 231242.

[21] M. E. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences, vol. 103, 2006, pp. 8577-8582.

[22] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities, J. Stat. Mech., 03024,2009.

[23] T. Gossen, et al.,Graph clusterings with overlaps: Adapted quality indices and a generation model, Neurocomputing, vol. 123,(2014), pp. 13-22.

[24] J. C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 ,(1974), pp. 57-71.

[25] J. C. Bezdek, Cluster validity with fuzzy sets, J. Cybernit. 3 , (1974), pp. 58-72.

[26] X. L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8), (1991), pp. 841-847.

[27] Y. Fukayama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, Proc. 5th Fuzzy Syst. Symp, (1989), pp. 247-250.

[28] D. Zhang, et al., A novel cluster validity index for fuzzy clustering based on bipartite modularity, Fuzzy Sets and Systems,253, (2014), pp. 122-137.

[29] N. Anuar and Z. Zakaria, Determination of fuzziness parameter in load profiling via Fuzzy C-Means, Control and System Graduate Research Colloquium (ICSGRC) 2011 IEEE, (2011), pp. 139-142.

[30] J. Yu, R. Hong, M. Wang, J. You, Image clustering based on sparse patch alignment framework ,Pattern Recognition, 47 (2014), pp. 35123519.

[31] K. Zeng , J. Yu, C. Li , J. You , T. Jin., Image clustering by hyper-graph regularized non-negative matrix factorization. Neurocomputing, 138,(2014), pp. 209217.

[32] L. Elden, Matrix methods in data mining and pattern recognition, SIAM, (2007).

[33] N. P. Nguyen and M. T. Thai, "Finding Overlapped Communities in Online Social Networks with Nonnegative Matrix Factorization" in military communications conference, 2012-MILCOM, (2012), pp. 1-6.

[34] N. P. Nguyen, et al., "Overlapping communities in dynamic networks: their detection and mobile applications," in Proceedings of the 17th annual international conference on Mobile computing and networking ,(2011), pp. 85-96.

[35] , D. Kuang, S. Yun and H. Park H. SymNMF: nonnegative low rank approximation of a similarity matrix for graph clustering. Journal of Global Optimization. 62(2015), pp. 545-574 (2015).

[36] A. Cichocki, et al., Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation, Wiley, (2009).

[37] , M. Rezghi, M. Yousefi, A Projected Alternating Least square Approach for Computation of Nonnegative Matrix Factorization, Jsciences, 26, (2015),pp. 273 - 279.

[38] H. Kim H, H. Park, Non-negative matrix factorization based on alternating non-negativity constrained least squares and the active set method. SIAM journal on matrix analysis and applications, 30, (2008), pp. 713-730.

[39] F. Shahnaz, et al., "Document clustering using nonnegative matrix factorization," Information Processing and Management, 42, (2006), pp. 373-386.

[40] W. Xu, et al., "Document clustering based on non-negative matrix factorization," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, (2003), pp. 267-273.

[41] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E, 78 ,(2008), pp. 046110.

[42] A. Lancichinetti and S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," Physical Review E,80, (2009) , pp. 016118.

[43] L. Tang and H. Liu, Community detection and mining in social media," Synthesis Lectures on Data Mining and Knowledge Discovery, 2, (2010), pp. 1-137.

[44] W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33,(1977), pp. 452-473.

[45] M. E. Newman and M. Girvan, Finding and evaluating community structure in networks, Physical Review E, 69, (2004), 026113.

[46] D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. Lond. B 270 , (2003), pp. 186188.

[47] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, S. Dawson, The bottle nose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 , (2003), pp. 396-405.

[48] P. Gleiser, L. Danon., Community Structure in Jazz. J Advances in Complex Systems,6, (2003), pp. 565-573.

[49] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing, 17, (2007), pp. 395-416.

- In this paper , we proposed a novel fuzzy clustering based on NMF . Experimental results show the superiority of our proposed fuzzy clustering in Comparison with the other methods.
- Also, we propose two new fuzzy clustering evaluation method for networks with and without ground truth . The novelty of these two evaluation method is that unlike the other methods they use all information of membership matrix in evaluation process.