Journal Pre-proof

Overlapping community detection in networks based on Neutrosophic theory

Maryam Gholami, Amir Sheikhahmadi, Keyhan Khamforoosh, Mahdi Jalili

PII: S0378-4371(22)00281-3
DOI: https://doi.org/10.1016/j.physa.2022.127359
Reference: PHYSA 127359

To appear in: *Physica A*

Received date : 2 October 2021
Revised date : 23 March 2022

Please cite this article as: M. Gholami, A. Sheikhahmadi, K. Khamforoosh et al., Overlapping community detection in networks based on Neutrosophic theory, *Physica A* (2022), doi: https://doi.org/10.1016/j.physa.2022.127359.

Overlapping community detection in networks based on Neutrosophic theory

Maryam Gholami[1], Amir Sheikhahmadi[1,*], Keyhan Khamforoosh[1], Mahdi Jalili[2]

[1] Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
[2] School of Engineering, RMIT University, Melbourne, Australia

Abstract

Discovering community structure is one of the most intensively studied problems in network science. Many real networks are composed of nodes belonging to multiple communities. In this manuscript, a new overlapping community detection algorithm is proposed based on neutrosophic set (NS) theory. The proposed community detection method manages uncertainty arisen from imprecise definition of communities, by handling boundary and outlier nodes. In the first step, the proposed algorithm calculates the dissimilarity index between each pair of nodes in the network. Then, in order to keep the original distance between nodes as much as possible, the network structure is mapped into a low-dimensional space by multidimensional scaling. Finally, the neutrosophic c-means algorithm is employed to find communities in the network. The experimental results show that the proposed algorithm can detect communities on real and artificial datasets effectively and accurately.

Keywords: Overlapping community detection, Neutrosophic theory, multidimensional scaling, fuzzy c-means

1. Introduction

In nature and society, many real networks containing large number of entities and interaction among them, can be described as complex networks[1, 2]. Examples include the Internet, power grids, transportation systems, water distribution systems, social networks and biochemical networks. Real networks often share some common properties, such as small-worldness, scale-free degree distribution, rich club and community structure. A community is made of a group (or cluster) of nodes within which the links between nodes are densely connected to each other, while nodes within a community are sparsely connected with other communities[3]. Detecting community structure in a network has important practical applications and can help understanding and analyzing the network system. For example, on the World Wide Web, topically related pages may link more densely among themselves [4] and society is organized into groups, friendship circles, families, villages and associations [5].

* Corresponding author: asheikhahmadi@iausdj.ac.ir (Amir Sheikhahmadi).

Network communities can be categorized as disjoint and overlapping. Disjoint communities do not share any common nodes, while overlapping communities might have some nodes shared between communities. Majority of the existing community detection algorithms have been originally proposed for disjoint communities, where each node is assigned to only a single community [6-9] . Detecting overlapping communities is an important topic in complex network analysis and Overlapping nodes may play a special role in a network system [10, 11]. In recent years, many algorithms have been proposed for detecting overlapping communities, such as Clique Percolation Method (CPM), which is based on this idea that the internal links of a community tends to construct cliques because of their high density [12] , Link partitioning where the idea is partitioning the set of links rather than the set of nodes [13] , density peaks based algorithms[14], Local expansion and optimization algorithms [15, 16], Label(or degree) Propagation Approach (LPA) [17, 18] , methods based on fuzzy relations and theory[19] , and matrix factorization based methods [20].

Overlapping community detection can be classified to fuzzy and non-fuzzy (crisp) methods. Non-fuzzy methods assume that a node either belongs to a community or it does not. However, fuzzy methods assume that a node may participate in several communities with varying degrees. For example, in a collaboration network of researchers, the overlapping may be fuzzy because a researcher who belongs to several communities cannot be fully involved with all of them, as a result of limited time and resources. The result of fuzzy community detection methods is a stochastic membership matrix that describes the probability of belonging every node to different communities [21]. Fuzzy methods for finding overlapping communities can be divided into four categories, as:

a) Methods based on non-negative matrix factorization (NMF)[22]
b) Methods based on modularity optimization [23]
c) Fuzzy relation-based methods [24]
d) Combined methods based on fuzzy c-means [21, 25-27]

Of these four categories, the last category has been the subject of much research in recent years. Zhang et al [21] mapped the network in an Euclidean space by spectral mapping and used the FCM to detect communities. In [25] a fuzzy community detection algorithm has been proposed based on local random walk (LRW) and a new distance metric. This method first calculates the dissimilarity index between nodes using the new distance measurement and then the network structure is mapped into a low-dimensional space by multidimensional scaling (MDS). Finally, FCM is employed to find fuzzy communities in a network. Deng et al [26] proposed a new method based on label propagation and fuzzy c-means algorithm. Firstly, the labels are initialized using the neighborhood evaluation method. Secondly, the nodes with diversity degree in each community are selected to change their labels by fuzzy c-means. In this algorithm the fuzzy c-means parameters are updated during iteration. In [27], an ant-based algorithm together with FCM has been proposed. In this method FCM is used to fine-tune the solution in final step. However, all of these methods still suffer from the drawbacks of FCM algorithm, such as trapping in a local optima, dependence of results on initialization, sensitivity to the presence of noise, and inability to distinguish between equally highly likely and highly unlikely [28].

Neutrosophy theory was proposed by Smarandache in 1995. It is a new branch of philosophy dealing with the origin, nature and scope of neutralities, and their interactions ideational spectra [29] . Neutrosophic theory provides a powerful tool to deal with the indeterminacy, and has found practical applications in various fields, such as data clustering [28] , image segmentation [30], and semantic web services [31].

In this paper, a new community detection algorithm is proposed based on neutrosophic set; the algorithm is named Neutrosophic c-means Community Detection (NCD). This algorithm consists of three steps. Firstly, the node distance matrix is calculated based on a new distance criterion. The network nodes are then mapped in a Euclidean space by MDS algorithm. Finally, the Neuromorphic c-means algorithm is applied to the existing points in the Euclidean space and the membership matrix is obtained. By applying a threshold to membership matrix values, one can detect non-fuzzy overlapping communities. The main motivation of this work is to handle overlapping nodes by using indeterminacy set (I) and detect outlier nodes by falsity set (F). NCD algorithm calculates the degrees belonging to three sets T, I and F for each node of network. T, I and F are considered as the membership degrees to determinant, indeterminacy and outlier communities, respectively. Indeterminacy community allows us to consider the nodes that are lying near the community boundary and outlier community allows us to reject the nodes that are far from the center of communities. The membership degrees to the indeterminacy and outlier communities are learned during the iterations of the algorithm.

The main contributions are summarized as follows:

1- This paper proposes a novel method to uncover overlapping communities based on Neutrosophic theory. The traditional fuzzy overlapping community detection methods only describe the membership degree to every community. For some nodes in the boundary regions and the outliers, it is difficult determine which community is belongs to. Moreover, the membership degrees of such nodes make the centers of community inaccurate. The proposed method overcome this problem.

2- The proposed method uses neutrosophic theory in the third phase of the algorithm, which enables this method to detect overlapping nodes and outliers using indeterminacy and falsity sets.

3- The proposed algorithm consists of three steps: In the first step, a new formula proposed for calculating the distance matrix elements. In the second step, based on the obtained matrix, the input network is mapped to a Euclidean space, and in the third step, overlapping communities are obtained according to the membership values obtained from the neutrosophic c-means algorithm.

Compared to existing methods, NCD can effectively detect and handle atypical nodes such outliers and boundary, and the adverse effect of such nodes on the calculation of community centers is reduced.

The rest of the paper is organized as follows. FCM algorithm and NS theory are reviewed in section 2. The NCD algorithm is described in section 3. Experimental results of the proposed algorithm are illustrated in section 4. Finally, Section 5 concludes the paper.

2. Neutrosophic set and fuzzy c-means

The proposed method in this paper is derived from neutrosophic set and fuzzy c-means concepts. In this section, these concepts are introduced.

2.1 Neutrosophic set

Uncertainty is lack of certainty in a system or problem that can make it difficult, if not impossible, to describe the outcome. A number of theories have been introduced to deal with uncertainties and imprecision enclosed in real-world systems. Examples include probability theory, fuzzy set theory, intuitionistic fuzzy sets, and rough set theory. Such methods fail to deal with indeterminate and inconsistent information. To outcome this drawback, Neutrosophic set (NS) theory was proposed as a generalization of intuitionistic fuzzy set theory [29] . This mathematical tool is described by three membership functions namely truth-membership (T), falsity membership (F) and indeterminacy-membership (I). In this framework the quantification of indeterminacy is explicit and determined by the value of indeterminacy-membership. Generally, in a neutrosophic set S, each element e is expressed as $e(t, i, f)$ which means that it is t% true, i% indeterminacy, and f% false. In each application, the concept behind true, false and indeterminacy is proposed by domain experts [32].

2.2 fuzzy c-means

Fuzzy c-means (FCM) [33] is one of the most widely used clustering algorithms where each data point has a certain membership degree of belonging to different clusters. Given $X = \{x_1, x_2, \dots, x_n\}$ be a dataset, and c $(2 \le c \le n)$ be the desired number of clusters, FCM returns a list of C cluster centers $C = \{c_1, c_2, \dots, c_c\}$, and a membership matrix $M = \{\mu_{ij} | \mu_{ij} \in [0,1], i = 1, 2, \dots, n, j = 1, 2, \dots, c\}$ where μ_{ij} denotes the membership degree of node i to cluster j. FCM minimizes an objective function J_m,

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \left\| x_i - c_j \right\|^2 \qquad (1)$$

where m is a hyper-parameter that controls the fuzziness level of the algorithm, and $x_i - c_j$ is the distance between the centers of the i-th and j-th clusters[33].

Fuzzy clustering is carried out through an iterative optimization of the objective function, and the membership μ_{ij} and cluster centers c_j are updated in each iteration by:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_k \right\|}{\left\| x_i - c_k \right\|}\right)^{2/m-1}} \qquad (2)$$

$$c_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^m \cdot x_i}{\sum_{i=1}^{n}(\mu_{ij})^m} \quad (3)$$

The iteration will not stop until $\max_{i,j}\{\left|\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}\right|\} < \varepsilon$, where ε is a small quantity, and k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m. Finally, each data point is assigned to a distinct cluster according to the membership degree[33].

3. Proposed method

In this section, the proposed algorithm, Neutrosophic c-means Community Detection (NCD), is introduced and discussed in detail. The algorithm consists of three phases: (i) computing the pairwise node similarity and construct a distance matrix of the network; (ii) embedding the nodes into a 2-D space by multi-dimensional scaling (MDS) method; and (iii) clustering the nodes by using Neutrosophic FCM. Fig. 1 shows flowchart of the NCD algorithm.

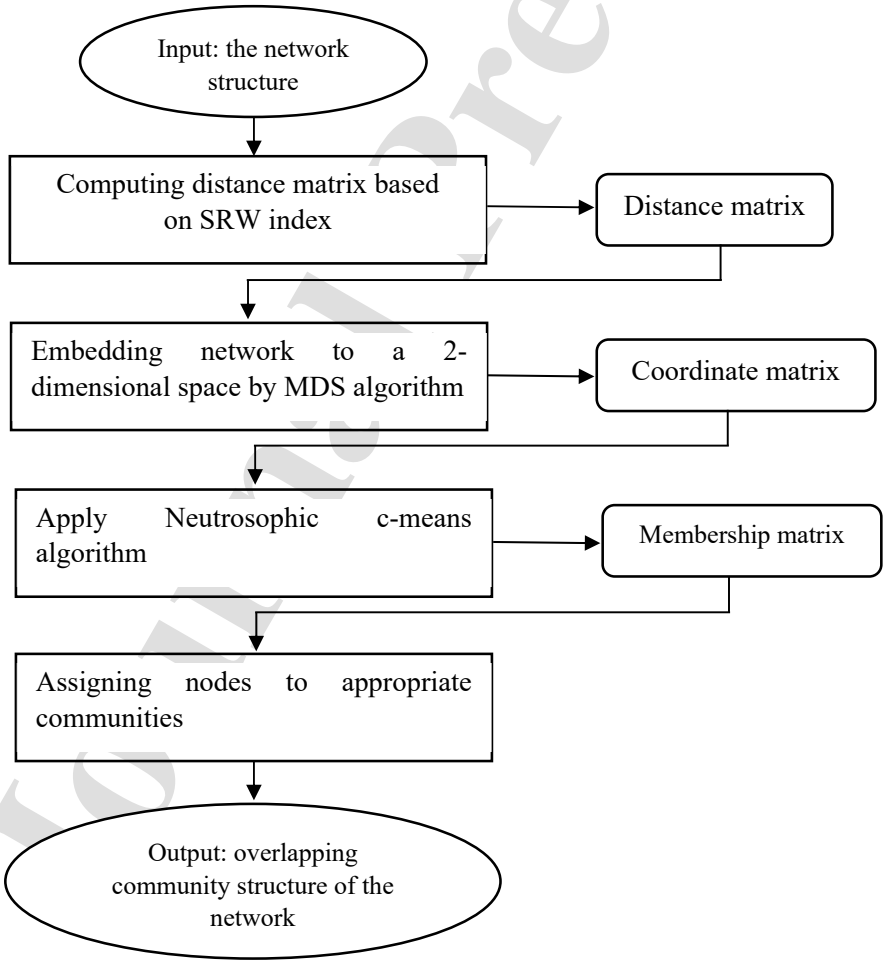


Figure 1: Flowchart of the proposed method

3.1 Computing distance matrix based on Local random walk

In this phase, we first calculate the similarity of each pair of nodes and then compute the distance between them by a simple subtracting operation. There are several approaches for computing node similarity. One group is node-dependent and require only local topology information like degree and the nearest neighborhood. Examples of such methods include Common Neighbors algorithm [34], Resource Allocation index and Local Path index [35]. Second group is path-dependent and exploits the global knowledge of the network topology. Examples include Average Commute Time (ACT) [36] and Random Walk with Restart (RWR) [37]. Third group comes under hybrid category and exploits the advantages of local and global features of the network. The hybrid methods do not require the complete network structure, and thus are less complex compared to the method requiring global information. Being more complex than local methods, hybrid methods give more accurate results. In this paper we use SRW [38] for calculating similarity of node pairs, because of its lower computational complexity and good accuracy. Given a random walker starting from node i let's denote $\pi_{ij}(t)$ as the probability that this walker locates at node j after t steps. The evolution equation can be presented as

$$\vec{\pi}_i(t) = P^T \vec{\pi}_i(t-1) \qquad (4)$$

where $\vec{\pi}_i(0) = \vec{e}_i$, P is transition probability matrix, and T is the matrix transposition[38]. LRW index at time step t is thus defined as

$$s_{ij}^{LRW}(t) = \frac{k_i}{2|E|} . \pi_{ij}(t) + \frac{k_i}{2|E|} . \pi_{ji}(t), \qquad (5)$$

where k_x denotes the degree of node x, and $|E|$ is the number of edges in the network. Superposed Random Walk (SRW) index assumes that the random walkers are continuously released from the starting point, resulting in a higher similarity between the target node and the nodes nearby. SRW is defined as[38]:

$$s_{ij}^{SRW}(t) = \sum_{l=1}^{t} s_{ij}^{LRW}(l) \qquad (6)$$

Using the above relationships, we define the distance matrix $D = d_{ij}$ of the network as:

$$d_{ij}(t) = \begin{cases} 1 - s_{ij}^{SRW}(t), & i \neq j; \\ 0, & i = j; \\ 3 & i \neq j,\ \sum_{k=1,k\neq j}^{n} s_{kj}^{SRW}(t) = 0 \ or \ \sum_{k=1,k\neq i}^{n} s_{ik}^{SRW}(t) = 0 \end{cases} \qquad (7)$$

3.2 Network embedding by Multidimensional scaling (MDS) algorithm

Multidimensional scaling (MDS) [39] is a dimensionality reduction technique used to map a multidimensional data into a low dimension space such that the distance matrix in the original data is preserved. Here, we use MDS algorithm for placing each node of network into a 2-D space such that the distances between the nodes are preserved as much as possible. Given distance matrix $D \in \mathbb{R}^{n\times n}$ as the input of MDS algorithm, d_{ij} denotes the pairwise distance between nodes i and j. The steps of MDS algorithm are described as follows:

1. Compute the squared distance matrix $D^{(2)} = [d_{ij}^2]$
2. Apply double centering to $D^{(2)}$ as $B = -\frac{1}{2}JD^{(2)}J$ using the centering matrix $= I - \frac{1}{n}\vec{1}\vec{1}^T$, where n is the number of nodes, I is the identity matrix, and $\vec{1}$ is a n-dimensional column vector with each entry being 1.
3. Determine the m largest eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_m$ and corresponding eigenvectors $e_1, e_2, \dots, e_m$ of B (where m is the number of dimensions disired for the output, here $m = 2$)
4. The resulting coordinate matrix X can be obtained via $X = E_m \Lambda_m^{1/2}$, where E_m is the matrix of m eigenvectors of B.

The coordinate matrix X contains the coordinates of nodes in a 2-D space. In this way, the information of an m-by-m space is mapped onto an m-by-2 space (m is the number of nodes in the network).

3.3 Community detection by Neutrosophic c-means

In this phase, neutrosophic c-means algorithm is applied to the embedded vector to extract the overlapping communities. A new community detection method is proposed derived from NS and FCM. Here, we consider not only the degree by which nodes belong to determinate communities, but also the degree of belonging to the indeterminate communities. Given T be the degree of belonging to determinant communities, I be the degree to the boundary regions, and F be the degree belonging to outlier region, we use the objective function proposed in [28] :

$$J(T, I, F, C) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\varpi_1 T_{ij})^m \|x_i - c_j\|^2 + \sum_{i=1}^{N}(\varpi_2 I_i)^m \|x_i - \bar{c}_{i\,max}\|^2 + \sum_{i=1}^{N}\delta^2(\varpi_3 F_i)^m \quad (8)$$

$$\bar{c}_{i\,max} = \frac{c_{max1} + c_{max2}}{2} \quad (9)$$

$$p_i = \max_{j=1,2,\dots,c}(T_{ij}) \quad (10)$$

$$q_i = \max_{j \neq p_i \cap j=1,2,\dots,c}(T_{ij}) \quad (11)$$

where m is a constant real number, δ is used to control the number of outliers, ϖ_i is the weight factor, T_{ij} is the degree of membership of node i to the determinant community j, $max1$ and $max2$ are the community indexes of the biggest and second biggest value of T_{ij}. $\bar{c}_{i\,max}$ is the average of the centers of two communities c_{max1} and c_{max2}, which is a constant number for each node i, I_i and F_i are the membership degrees belonging to overlapping regions and outlier set, $0 < T_{ij}, I_i, F_i < 1$, and satisfy with following formula:

$$\sum_{j=1}^{c} T_{ij} + I_i + F_i = 1 \quad (12)$$

After constructing the Lagrange objective function and using some operations to minimize it, the following relationships are obtained [28]:

$$c_j = \frac{\sum_{i=1}^{N}(\varpi_1 T_{ij})^m x_i}{\sum_{i=1}^{N}(\varpi_1 T_{ij})^m} \quad (13)$$

$$T_{ij} = \frac{K}{\varpi_1}(x_i - c_j)^{-(2/m-1)} \quad (14)$$

$$I_i = \frac{K}{\varpi_2}(x_i - \bar{c}_{i\,max})^{-(2/m-1)} \quad (15)$$

$$F_i = \frac{K}{\varpi_3}\delta^{-(2/m-1)} \quad (16)$$

The clustering is carried out through an iterative optimization of the objective function, and the membership T_{ij} , I_i, F_i and the community centers c_j are updated by Eqs. 13-16 at each iteration. The iteration will not stop until $\left|T_{ij}^{(k+1)} - T_{ij}^{(k)}\right| < \varepsilon$, where ε is a termination criterion between 0 and 1, and k is iteration step. Finally, each node is assigned into the community with the biggest $TM = [T, I, F]$ value.

According to the above equations, the third phase (neutrosophic c-means) can be summarized in the following steps:

1- Initialize k, m, δ, ε, ϖ_1, ϖ_2, ϖ_3 parameters
2- Initialize T, I, and F in time 0 randomly
3- Calculate the centers of communities using Eq. 13
4- Compute $\bar{c}_{i\,max}$ according to indexes of the largest and second largest value of T
5- Compute $T_{k+1}, I_{k+1}, F_{k+1}$ using Eq. 14, 15, 16
6- If $|T_{k+1} - T_k| < \varepsilon$ then stop; otherwise return to step 3.
7- Assign each data into the class with biggest $TM = [T, I, F]$ value.

It is important to mention that the parameterization of the methods is a fundamental step in the execution of this method, the main parameters of NCD algorithm, which are used in three phases, are listed in table 1. The first parameter is t that defines the number of random walk steps in phase 1. The experiments on synthetic and real-world networks used in this work suggest that acceptable results can be obtained despite having the number of random walk steps smaller than the diameter of the network. Thus, we set the diameter of the network as the number of random walk steps t in the experiments. The embedding dimension p is taken as 2 for reducing the computational complexity in phase 2. To set the value for the rest of the parameters, we adopt the strategy of maximizing the value of the modularity to determine the optimal value for parameters.

Table 1: The important parameters of the proposed algorithm

Phase	Parameter name	Description
1	t	The number of random walk steps
2	p	Embedding dimension
3	k	The number of communities
	m	The fuzzification constant

	ε	Termination condition parameter
	δ	Controls the number of outliers
	ϖ_1	Weight parameter for T set
	ϖ_2	Weight parameter for I set
	ϖ_3	Weight parameter for F set

4. Experiments

In this section, we study the performance of NCD algorithm, and compare its performance with state-of-the-art overlapping community detection algorithms. We first demonstrate the performance of the NCD algorithm using two simple networks and then evaluate the performance of the NCD on both synthetic and real-world networks. We also compare NCD algorithm with the baseline algorithms in terms of running time. All experiments have been implemented with python in a system 2.4 GHz cpu 16 Gb RAM. The details of the test settings are described in each section separately.

4.1 Simple test networks

In this section, we use two simple networks to illustrate the performance of NCD. The first network (net1) contains 10 nodes and two communities, node 5 is shared between two communities and node 10 is not connected to the rest of the network, i.e., an outlier node (Fig. 2).

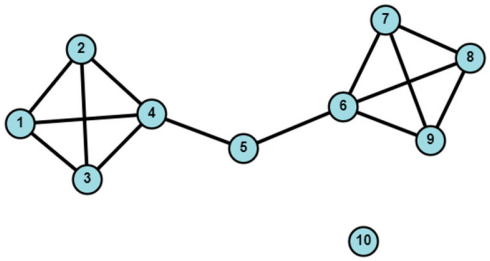


Figure 2: A sample network with node 5 in the overlap and node 10 as outlier

The values of T, I and F memberships are shown in Table 2, and the neutrosophic assignments for nodes are presented in the last column. These assignments are obtained according to the maximum value in each row. It can be seen that the two natural communities (c_1, c_2) are correctly detected. Nodes 1, 2, 3 and 4 are assigned to community c_1 and nodes 6, 7, 8 and 9 are assigned to c_2 . Node 5 is assigned to the indeterminacy set, which indicates that this node is an overlapping node between c_1 and c_2. Node 10 is assigned to falsity set. It can be considered as an outlier node.

Table 2: Membership values obtained from applying NCD on net.1

point	$T(c_1)$	$T(c_2)$	I	F	Community
1	0.9561	0.0047	0.0347	0.0043	c_1
2	0.9837	0.0017	0.0130	0.0014	c_1
3	0.8934	0.0107	0.0872	0.0085	c_1
4	0.5600	0.0448	0.3635	0.0314	c_1

5	0.1100	0.1100	0.7108	0.0691	Overlapping node
6	0.0448	0.5600	0.3635	0.0314	c_2
7	0.0047	0.9561	0.0347	0.0043	c_2
8	0.0017	0.9837	0.0130	0.0014	c_2
9	0.0107	0.8934	0.0872	0.0085	c_2
10	0.0822	0.0822	0.1101	0.7253	Outlier node

We compare the results obtained by NCD with those of FCM [20], NMF [22] and LRW [25] algorithms in terms of the degree of membership of nodes to communities. The membership values generated by these algorithms are presented in Fig. 3. Every algorithm finds a reasonable community structure of the network, but the overlapping and outlier nodes are assigned to the communities. However, NCD is the only algorithm that is able to detect these conditions.

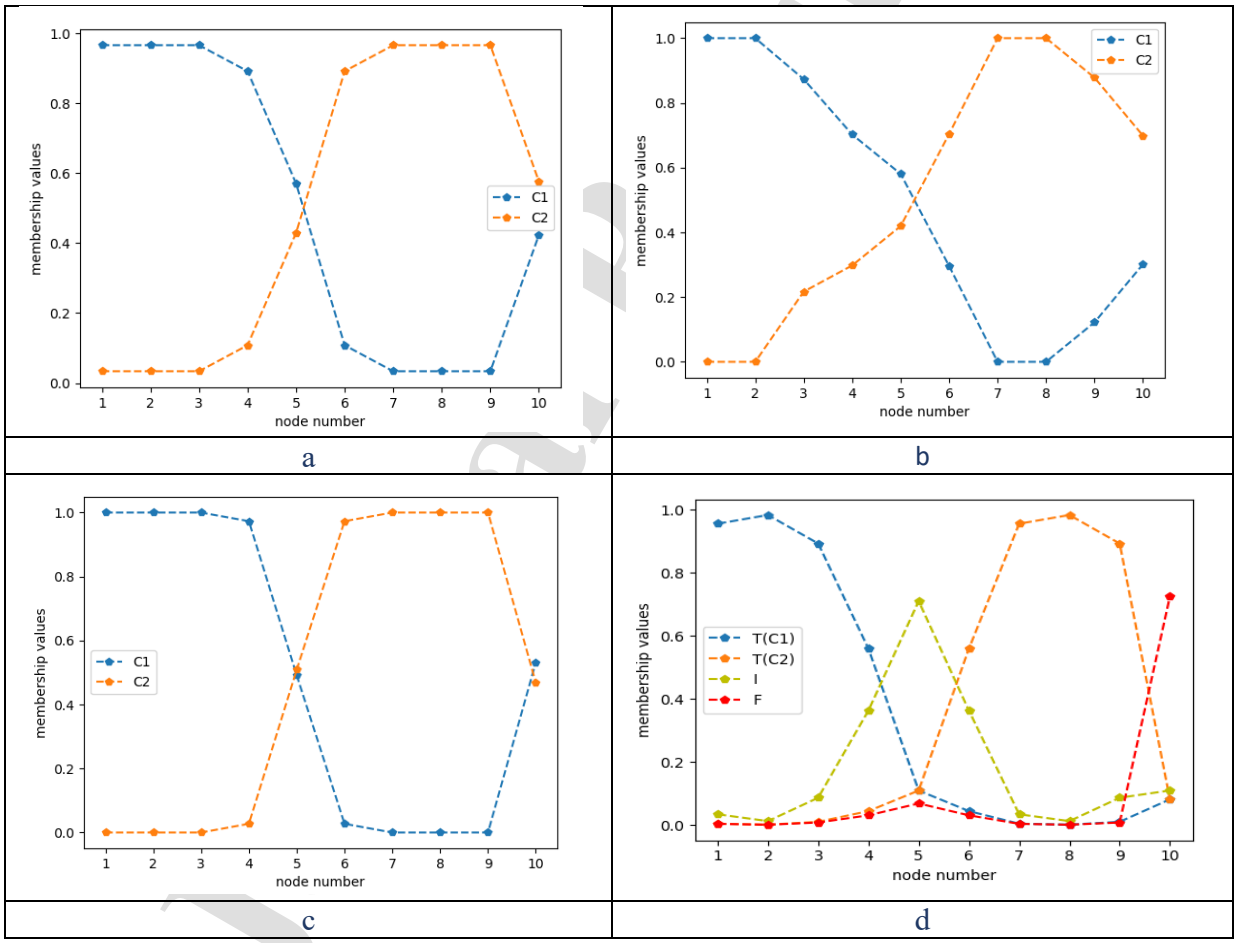


Figure 3: Community detection results on the simple network: (a) result of FCM; (b) result of NMF; (c) result of LRW; (d) result of NCD

The second network(net2) used here is a more complicated network composed of 16 nodes. It contains three communities, two overlapping nodes (5,10) and two outlier nodes (16, 17). It is shown in Fig. 4.

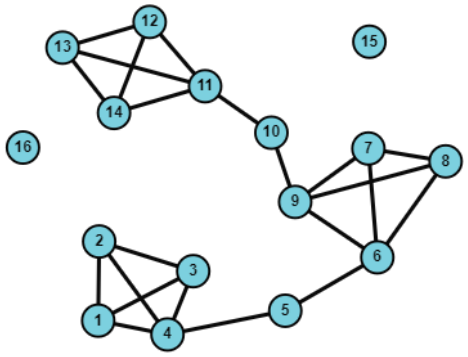


Figure 4: A sample network with node 5 in the overlap and node 10 as outlier

The results of applying NCD algorithm on net2 is shown in Table 3, indicating that NCD correctly detected the three natural communities. Nodes 5 and 10 are assigned to indeterminacy set, which indicates the overlapping nodes, and nodes 15 and 16 are assigned to falsity set as outlier nodes. The membership values obtained by FCM, NMF, LRW and NCD algorithms are shown in Fig. 5. It can be seen that NCD is the only algorithm that can detect overlapping and outlier nodes correctly. Table 3 Shows the T, I and F membership values and the last column shows neutrosophic community assignments for nodes.

Table 3: Membership values obtained from applying NCD on net.2

point	$T(c_1)$	$T(c_2)$	$T(c_3)$	I	F	Community
1	0.9108	0.0109	0.0135	0.0458	0.0187	c_1
2	0.7233	0.0479	0.0441	0.1489	0.0356	c_1
3	0.6526	0.0459	0.0589	0.1890	0.0534	c_1
4	0.5175	0.0906	0.0864	0.2077	0.0705	c_1
5	0.0990	0.0712	0.0373	0.6783	0.0840	I
6	0.1148	0.5754	0.1031	0.1361	0.0703	c_2
7	0.0352	0.8233	0.0371	0.0897	0.0145	c_2
8	0.0314	0.7788	0.0320	0.1039	0.0536	c_2
9	0.1066	0.5953	0.1151	0.1562	0.0772	c_2
10	0.0346	0.1139	0.1019	0.6605	0.0888	I
11	0.0962	0.1077	0.5360	0.1916	0.0682	c_3
12	0.0585	0.0450	0.6302	0.2104	0.0467	c_3
13	0.0505	0.0465	0.7446	0.1382	0.0200	c_3
14	0.0313	0.0180	0.8683	0.0730	0.0730	c_3
15	0.0837	0.0605	0.0919	0.1650	0.5988	F
16	0.0633	0.0678	0.0729	0.1174	0.6783	F

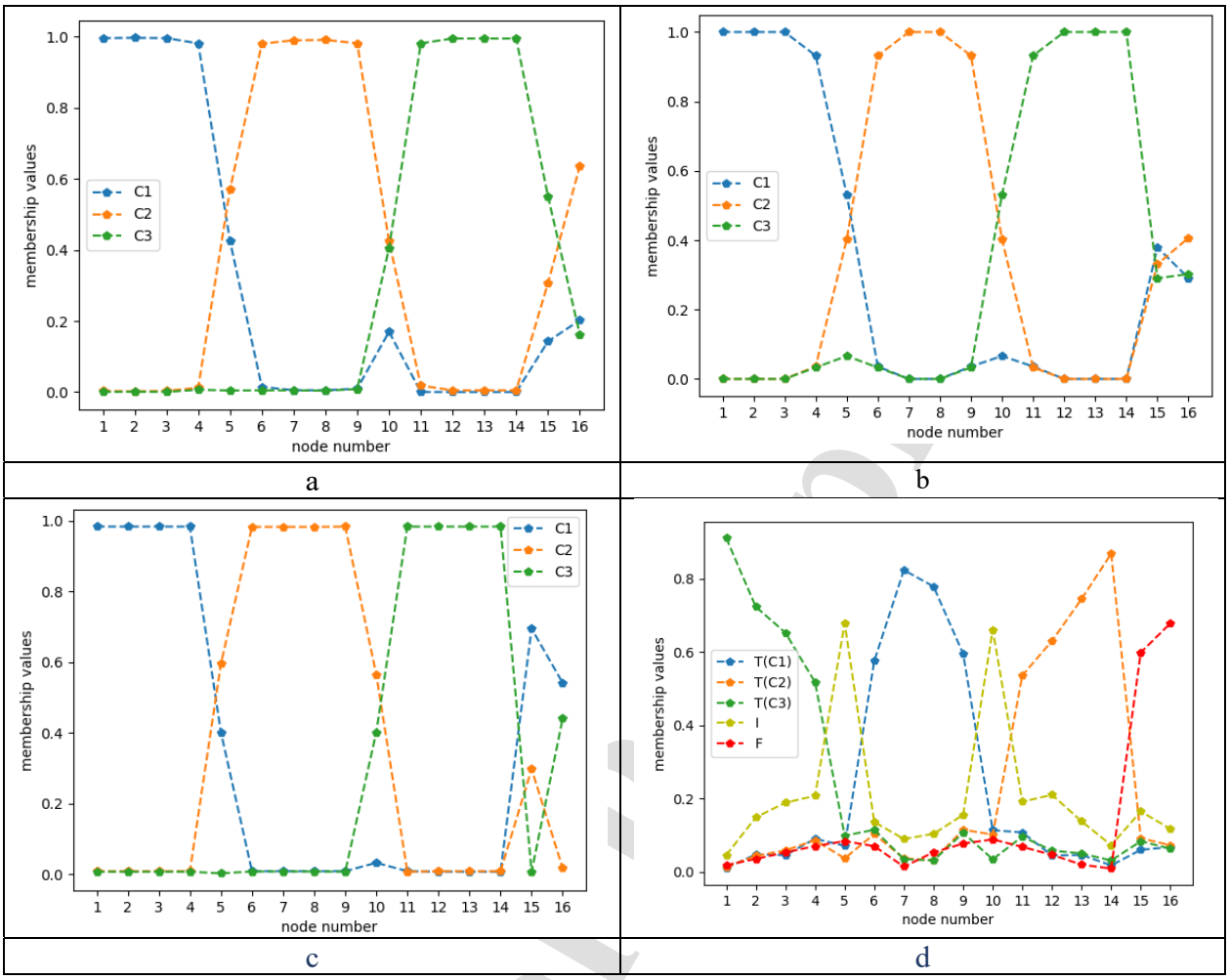


Figure 5: Community detection results on the simple network: (a) result of FCM ; (b) result of NMF ; (c) result of LRW; (d) result of NCD

4.2 Synthetic networks

In this section, the performance of the algorithms is tested in two artificially constructed networks obtained from the public LFR benchmark [40]. LFR benchmark model provides high flexibility to control the structure of network by tuning different parameters, including the number of nodes n, the average degree $\bar{k}$, the minimum and maximum community size $|C|_{min}$ and $|C|_{max}$, the topological mixing parameter μ, the number of nodes belonging to multiple communities O_n. We generate two benchmark networks with two values of the mixing parameter as $\mu = 0.1\,, 0.3$. The first network contains 500 nodes and the second contains 1000 nodes. Table 4 lists the value of other parameters we have used for generating the LFR networks.

We compare the NCD method with other approaches including the NMF [22], FCM [21], Local Random Walk (LRW) [25], EADP[14] and MDPA[17] . The parameters of these algorithms are set according to the reference of each. To specify the optimal number of communities We conduct experiments using different numbers of communities and select the optimal number according to the modularity.

To compare the detected overlapping community structure with the ground-truth, we use the overlapping normalized mutual information ($oNMI$) measure[41] which is an information-theoretic measure of the agreement between two overlapping community structures. The $oNMI$ between a ground truth A and the community structure B obtained by an particular approach can be defined as:

$$oNMI(A,B) = \frac{-2\sum_{i=1}^{k_A}\sum_{j=1}^{k_B} N_{ij}\log\left(\frac{nN_{ij}}{N_i N_{.j}}\right)}{\sum_{i=1}^{k_A} N_{i.}\log\left(\frac{N_{i.}}{n}\right) + \sum_{j=1}^{k_B} N_{.j}\log\left(\frac{N_{.j}}{n}\right)}$$

Where k_A and k_B denote the number of communities in A and B respectively, $N \in \mathbb{R}^{k_A \times k_B}$ is a matrix with N_{ij} being the number of nodes in the i-th community of A that appear in the j-th community of B, $N_{i.}$ is the sum over the i-th row of N, and $N_{.j}$ is the sum over the j-th column of N. $oNMI$ ranges from 0 to 1 and equals 1 in the perfect matching between detected overlapping community structure and the ground-truth.

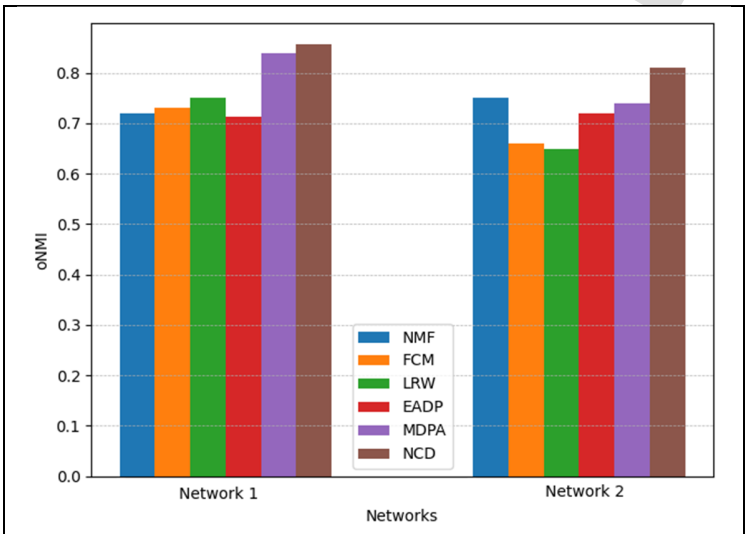
The $oNMI$ obtained by proposed method and five baseline methods are presented in table 5 and the bar plot of the $oNMI$ values is shown in Fig. 6. It can be observed that the $oNMI$ values achieved by all algorithms decrease with the increase of the parameter μ. Because with the parameter μ increases, the networks are changing to highly overlapping and connection inside communities are weaker for larger μ. On the other hand, by increasing network size from 500 to 1000 typically results in slightly lower performance. We also found the NCD method performs better than other methods in detecting overlapping communities in the tested networks. The reason of this superiority is that NCD can detect overlapping nodes and outlier nodes with higher accuracy by applying neutrosophic theory.

Table 4: The value of the parameters for generating LFR networks

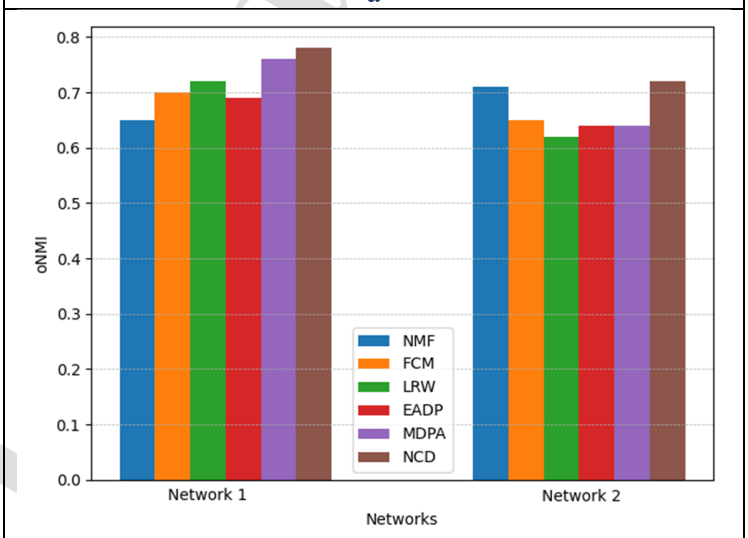
Parameter	Description	Network 1	Network 2
n	Number of nodes	500	1000
μ	Mixing parameter	0.1, 0.3	0.1, 0.3
$\bar{k}$	Average degree	10	10
$\|C\|_{min}$	Minimum community size	5	5
$\|C\|_{max}$	Maximum community size	15	15
τ_1	Node degree distribution exp.	2	2
τ_2	Community size distribution exp.	1	1
o_n	Number of nodes belonging to multiple communities	50	100
o_m	Maximum number of communities that a node belongs to	4	4

Table 5: oNMI values obtained from NCD and baseline algorithms

	$\mu = 0.1$		$\mu = 0.3$	
	Network1($n = 500$)	Network2($n = 1000$)	Network1($n = 500$)	Network2($n = 1000$)
NMF	0.7220	0.7560	0.6567	0.7125
FCM	0.7345	0.6632	0.7040	0.6509
LRW	0.7512	0.6576	0.7234	0.6234
EADP	0.7134	0.7200	0.6903	0.6495
MDPA	0.8390	0.7451	0.7671	0.6456
NCD	0.8567	0.8113	0.7890	0.7209



a



b

Figure 6: The oNMI values for two synthetic networks with μ=0.1(a) and μ= 0.3(b)

4.3 Real-world networks

To further test our proposed method, we consider 7 real-world networks including Zachary's Karate Club network[42], the Bottlenose Dolphins network[43], the Jazz musician network[44], the American College Football network and Pol. Books [3], the GR-QC (General Relativity and Quantum Cosmology) collaboration network, and the HEP-PH (High Energy Physics - Phenomenology) collaboration network[45]. Detailed information on these networks is listed in table 6.

The baseline algorithms are the same as those tested in the previous section. To evaluate the proposed method and compare it with baseline algorithms, in this section we use extended modularity measure to overlapping communities (Q_{ov})[46]. Q_{ov} evaluates the quality of overlapping communities from structure perspective, which is used for networks that their ground-truth communities are unknown. Q_{ov} value varies from 0 to 1, and higher values of it indicate stronger community structure. We ran every algorithm 50 times on each network and the average value of Q_{ov} is recorded.

Table 6: Detailed information of the real benchmark datasets

Network	Nodes	Edges	Description
Karate[1]	34	78	Social network of friendships between 34 members of a karate club at a US university in the 1970s.
Dolphin[1]	62	159	between 62 dolphins in a community living off Doubtful Sound, New Zealand.
Pol. Books[1]	105	441	A network of books about US politics sold by the online bookseller. Edges between books represent frequent co-purchasing of books by the same buyers.
Football[1]	115	613	Network of American football games between Division IA colleges during regular season Fall 2000
Jazz[1]	198	2742	List of edges of the network of Jazz musicians
CA-GrQc[2]	5242	14496	Collaboration network of Arxiv General Relativity category. There is an edge if authors coauthored at least one paper.
CA-HepPh[2]	12008	118521	Collaboration network of Arxiv High Energy Physics Theory. There is an edge if authors coauthored at least one paper

1. http://konect.cc/networks/
2. https://snap.stanford.edu/data/

Table 7 presents the average Q_{ov} value obtained by NCD and other methods on real-world networks and in each row, the best case is bolded. According to Table 6, although NCD can't obtain the highest Q_{ov} on all the real-world networks, it outperforms the baseline approaches in most cases. As can be seen, NCD and MDPA exhibit relatively good performance among the five baseline approaches. The results obtained by these two algorithms are close to each other in most networks that have been tested. In the case of Network Pol. books, the result obtained by MDPA is slightly better than the NCD algorithm. But in the case of the Jazz network, the difference between the results is greater. This is due to the high average degree in the Jazz network, which makes the network structure more complex in terms of community overlap. The bar plot of the Q_{ov} values is shown in fig 7.

Table 7: Q_{ov} values obtained from applying each algorithm on real-world networks

Network	NMF	FCM	LRW	EADP	MDPA	NCD
Karate	0.7155	0.6344	0.6364	0.7529	0.8302	**0.8360**
Dolphin	0.7462	0.7116	0.7169	0.7395	0.7459	**0.7631**
Pol. Books	0.5023	0.6900	0.6233	0.8342	**0.8416**	0.8320
Football	0.6521	0.6052	0.5206	0.6953	0.7350	**0.7532**
Jazz	0.7235	0.6455	0.7225	0.7322	**0.7418**	0.5309
CA-GrQc	0.5317	0.4808	0.4532	0.5020	0.5617	**0.5814**
CA-HepPh	0.2317	0.1841	0.2071	0.2560	0.2900	**0.2914**

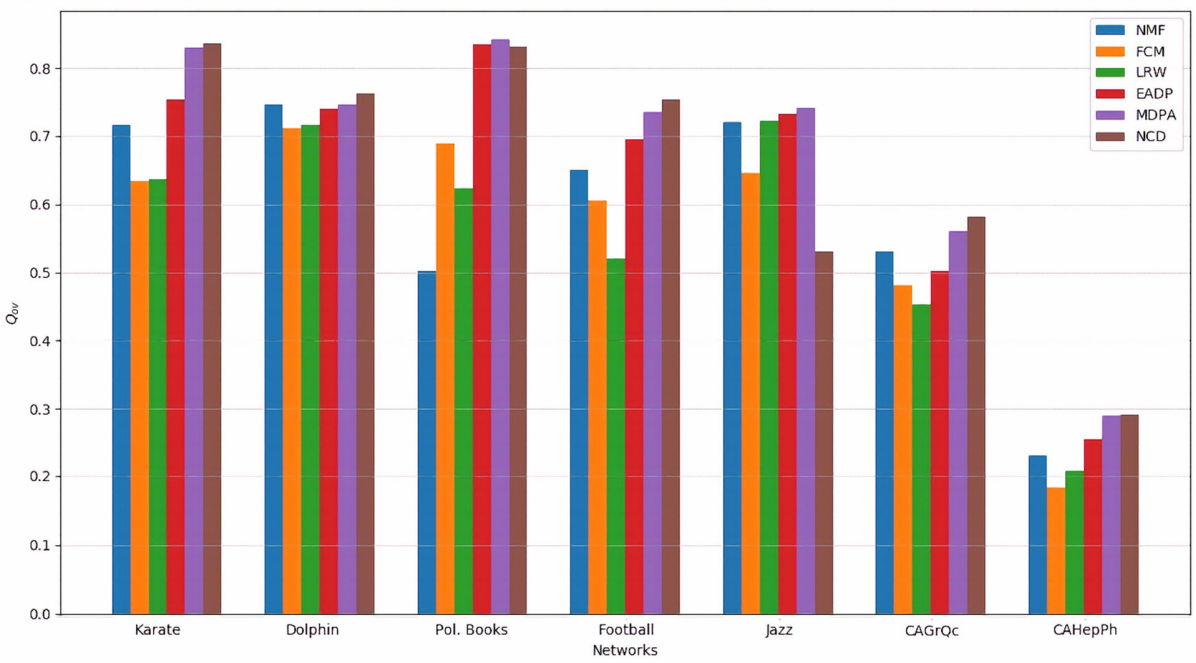


Figure 7: The Q_{ov} values for real networks

4.4 Running time

In this section, we compare NCD algorithm with the baseline algorithms in terms of running time. We use the real datasets as discussed in the previous section and the experimental results are shown in table 8. The main factor affecting the running time of the algorithms is the network size. As we can see in table 8, the running time of algorithms increases as the network size increases. FCM and NCD showed higher execution time compared to other algorithms. The main reason is that FCM and NCD repeat the center computing process many times.

Table 8: Comparison of Running Time with Baseline Algorithms (in seconds)

Network	NMF	FCM	LRW	EADP	MDPA	NCD
Karate	3	5	3	2	2	5
Dolphin	3	6	3	3	2	7
Pol. Books	6	8	6	5	4	8

Football	10	12	10	8	8	13
Jazz	15	16	13	15	14	20
CA-GrQc	152	210	123	140	198	216
CA-HepPh	1623	1876	1670	1535	1658	1700

5. Conclusions

In this paper, an efficient community detection algorithm, Neutrosophic c-means Community Detection algorithm (NCD), was introduced to overlapping community detection in complex network. The main idea is to handle boundary and outlier nodes by using I and F sets in NS domain. In this way, the effect of such nodes on reducing the accuracy of calculating the centers of the clusters is avoided. NCD detects overlapping communities in three phases, firstly it calculates the node distance matrix, in the second phase, MDS method is used to map each node in the network into lower-dimensional space. Finally, the nodes are clustered into communities by using neutrosophic c-means. In contrast with previous similar methods, NCD has three advantages: (i) it is based on a new idea about network analysis, (ii) the algorithm can effectively detect and handle atypical nodes such outliers and boundary nodes, and (iii) the algorithm can obtain reasonable results, The efficiency of the proposed NCD algorithm is tested on some synthetic and real-world networks. Experimental results show that the performance of our algorithm is more efficient than performances other state of the art algorithms. The NCD algorithm can be extended to directed and/or weighted networks. In addition, we plan to apply the NCM to the signed networks in our future work.

References

1. Newman, M.E., *The structure and function of complex networks.* SIAM review, 2003. **45**(2): p. 167-256.
2. Li, H.-J., et al., *The dynamics of epidemic spreading on signed networks.* Chaos, Solitons & Fractals, 2021. **151**: p. 111294.
3. Girvan, M. and M.E. Newman, *Community structure in social and biological networks.* Proceedings of the national academy of sciences, 2002. **99**(12): p. 7821-7826.
4. Flake, G.W., S. Lawrence, and C.L. Giles. *Efficient identification of web communities*. in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2000.
5. Feld, S.L., *The focused organization of social ties.* American journal of sociology, 1981. **86**(5): p. 1015-1035.
6. Fortunato, S., *Community detection in graphs.* Physics reports, 2010. **486**(3-5): p. 75-174.
7. Mohammadi, M., P. Moradi, and M. Jalili, *SCE: Subspace-based core expansion method for community detection in complex networks.* Physica A: Statistical Mechanics and its Applications, 2019. **527**: p. 121084.
8. Zare, H., M. Hajiabadi, and M. Jalili, *Detection of Community Structures in Networks With Nodal Features based on Generative Probabilistic Approach.* IEEE Transactions on Knowledge and Data Engineering, 2021. **33**(7): p. 2863-2874.

9. Li, H.-J., et al., *Optimization of identifiability for efficient community detection.* New Journal of Physics, 2020. **22**(6): p. 063035.
10. Xie, J., S. Kelley, and B.K. Szymanski, *Overlapping community detection in networks: The state-of-the-art and comparative study.* Acm computing surveys (csur), 2013. **45**(4): p. 1-35.
11. Yang, J. and J. Leskovec, *Overlapping communities explain core–periphery organization of networks.* Proceedings of the IEEE, 2014. **102**(12): p. 1892-1902.
12. Palla, G., et al., *Uncovering the overlapping community structure of complex networks in nature and society.* nature, 2005. **435**(7043): p. 814-818.
13. Ahn, Y.-Y., J.P. Bagrow, and S. Lehmann, *Link communities reveal multiscale complexity in networks.* nature, 2010. **466**(7307): p. 761-764.
14. Xu, M., et al., *EADP: An extended adaptive density peaks clustering for overlapping community detection in social networks.* Neurocomputing, 2019. **337**: p. 287-302.
15. Shen, H.-W., *Detecting the overlapping and hierarchical community structure in networks*, in *Community Structure of Complex Networks*. 2013, Springer. p. 19-44.
16. Huang, L., C.-D. Wang, and H.-Y. Chao, *oComm: Overlapping community detection in multi-view brain network.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019. **18**(4): p. 1582-1595.
17. Gao, R., et al., *Overlapping Community Detection Based on Membership Degree Propagation.* Entropy, 2021. **23**(1): p. 15.
18. Gregory, S., *Finding overlapping communities in networks by label propagation.* New journal of Physics, 2010. **12**(10): p. 103018.
19. Wang, X., et al., *Uncovering fuzzy communities in networks with structural similarity.* Neurocomputing, 2016. **210**: p. 26-33.
20. Li, Z., et al., *An efficient semi-supervised community detection framework in social networks.* PloS one, 2017. **12**(5): p. e0178046.
21. Zhang, S., R.-S. Wang, and X.-S. Zhang, *Uncovering fuzzy community structure in complex networks.* Physical Review E, 2007. **76**(4): p. 046103.
22. Binesh, N. and M. Rezghi, *Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria.* Applied Soft Computing, 2018. **69**: p. 689-703.
23. Nepusz, T., et al., *Fuzzy communities and the concept of bridgeness in complex networks.* Physical Review E, 2008. **77**(1): p. 016107.
24. Sun, P.G., L. Gao, and S.S. Han, *Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks.* Information Sciences, 2011. **181**(6): p. 1060-1071.
25. Wang, W., et al., *Fuzzy overlapping community detection based on local random walk and multidimensional scaling.* Physica A: Statistical Mechanics and its Applications, 2013. **392**(24): p. 6578-6586.
26. Deng, Z.-H., et al., *A complex network community detection algorithm based on label propagation and fuzzy C-means.* Physica A: Statistical Mechanics and its Applications, 2019. **519**: p. 217-226.
27. Noveiri, E., M. Naderan, and S.E. Alavi, *ACFC: ant colony with fuzzy clustering algorithm for community detection in social networks.* International Journal of Ad Hoc and Ubiquitous Computing, 2019. **31**(1): p. 36-48.
28. Guo, Y. and A. Sengur, *NCM: Neutrosophic c-means clustering algorithm.* Pattern Recognition, 2015. **48**(8): p. 2710-2724.
29. Smarandache, F., *A unifying field in logics: neutrosophic logic. Neutrosophy, neutrosophic set, neutrosophic probability: neutrsophic logic. Neutrosophy, neutrosophic set, neutrosophic probability*. 2005: Infinite Study.

30. Heshmati, A., M. Gholami, and A. Rashno, *Scheme for unsupervised colour–texture image segmentation using neutrosophic set and non-subsampled contourlet transform.* IET Image Processing, 2016. **10**(6): p. 464-473.
31. Kandasamy, W.V. and F. Smarandache, *Some neutrosophic algebraic structures and neutrosophic n-algebraic structures*. 2006: Infinite Study.
32. Guo, Y. and H.-D. Cheng, *New neutrosophic approach to image segmentation.* Pattern Recognition, 2009. **42**(5): p. 587-595.
33. Dunn, J.C., *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.* 1973.
34. Lorrain, F. and H.C. White, *Structural equivalence of individuals in social networks.* The Journal of mathematical sociology, 1971. **1**(1): p. 49-80.
35. Lü, L., C.-H. Jin, and T. Zhou, *Similarity index based on local paths for link prediction of complex networks.* Physical Review E, 2009. **80**(4): p. 046122.
36. Klein, D.J. and M. Randić, *Resistance distance.* Journal of mathematical chemistry, 1993. **12**(1): p. 81-95.
37. Tong, H., C. Faloutsos, and J.-Y. Pan. *Fast random walk with restart and its applications*. in *Sixth international conference on data mining (ICDM'06)*. 2006. IEEE.
38. Liu, W. and L. Lü, *Link prediction based on local random walk.* EPL (Europhysics Letters), 2010. **89**(5): p. 58007.
39. Jain, A.K., R.P.W. Duin, and J. Mao, *Statistical pattern recognition: A review.* IEEE Transactions on pattern analysis and machine intelligence, 2000. **22**(1): p. 4-37.
40. Lancichinetti, A., S. Fortunato, and F. Radicchi, *Benchmark graphs for testing community detection algorithms.* Physical review E, 2008. **78**(4): p. 046110.
41. Lancichinetti, A., S. Fortunato, and J. Kertész, *Detecting the overlapping and hierarchical community structure in complex networks.* New journal of physics, 2009. **11**(3): p. 033015.
42. Zachary, W.W., *An information flow model for conflict and fission in small groups.* Journal of anthropological research, 1977. **33**(4): p. 452-473.
43. Lusseau, D., *The emergent properties of a dolphin social network.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 2003. **270**(suppl_2): p. S186-S188.
44. Gleiser, P.M. and L. Danon, *Community structure in jazz.* Advances in complex systems, 2003. **6**(04): p. 565-573.
45. Leskovec, J., J. Kleinberg, and C. Faloutsos, *Graph evolution: Densification and shrinking diameters.* ACM transactions on Knowledge Discovery from Data (TKDD), 2007. **1**(1): p. 2-es.
46. Nicosia, V., et al., *Extending the definition of modularity to directed graphs with overlapping communities.* Journal of Statistical Mechanics: Theory and Experiment, 2009. **2009**(03): p. P03024.

CRediT author statement

Maryam Gholami: Conceptualization, Methodology, Software, Writing- Original draft preparation. **Amir Sheikhahmadi**: Supervision, Methodology, Writing- Original draft preparation, Investigation, Validation, Reviewing. **Keyhan Khamforoosh**: Validation, Reviewing. **Mahdi Jalili:** Writing- Reviewing and Editing,

Declaration of Interest Statement

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.