A New Local Centrality Measure for Detecting Community Cores in Social Networks

Jafar Sheykhzadeh[1], Bagher Zarei[2], and Farhad Soleimanian Gharehchopogh[3]
[1,3]Department of Computer Engineering, Ur.C., Islamic Azad University, Urmia, Iran
[2]Department of Computer Engineering, Shab.C., Islamic Azad University, Shabestar, Iran
Email: zarei.bagher@iau.ac.ir (corresponding author)
Receive Date: 10 Feb 2025 Revise Date: 6 March 2025 Accept Date: 28 April 2025

ABSTRACT

Community detection is one of the important challenges in social network analysis. This problem involves identifying the internal structures of the network and grouping nodes into communities with common characteristics. One of the effective approaches for community detection is to identify important nodes in the network and consider them as the initial cores of communities. In this study, first, a new local centrality criterion is introduced to determine the importance of nodes. Based on this criterion, important nodes are identified and considered as the cores of initial communities. Also, in this paper, a new algorithm called CDHC is presented for community detection. Modularity and NMI indices are used to evaluate the proposed algorithm. The results of the proposed algorithm on real and artificial networks show that the proposed algorithm is efficient compared to other algorithms.

Keywords: Social Network Analysis, Community Detection, Nodes centrality, Modularity

1. INTRODUCTION

Detecting communities in social networks is known as one of the main challenges in network analysis. In these networks, nodes and the relationships between them play an important role in understanding the structure and behavior of the network. Detecting communities can help to better understand these structures and have wide applications in areas such as identifying like-minded groups, analyzing social behavior, and optimizing networks [1, 2].

In many existing algorithms for detecting communities, high-importance nodes play a pivotal role. These nodes are usually considered as the cores of communities, because they usually have the greatest impact on network communications. Therefore, identifying important nodes is essential as the first step in the community detection process [3, 4].

The algorithm proposed in this paper is designed based on evaluating the local importance of nodes by a new centrality measure. After calculating the importance of nodes, these nodes are sorted according to their importance value and the important nodes are selected as the cores of initial communities. Then, using a prioritization strategy among the neighbors of the core nodes, initial communities are formed. Subsequently, weak communities with weak internal connections are identified and merged with stronger communities.

This algorithm is designed to improve the accuracy and efficiency of community detection using new criteria based on a more accurate assessment of nodes and their connections. Since social networks are constantly changing, this algorithm can be used dynamically and optimized to detect communities in such networks.

The rest of the paper is written as follows. In Section 2, some of the work done is reviewed. In Section 3, the proposed centrality criterion and the proposed algorithm are written. In Section 4, the results and experiments performed are presented.

2. Related Works

In recent years, community detection in social networks has become one of the most challenging and widely used problems in the field of complex network analysis. Communities are known as important infrastructures in networks where nodes with similar characteristics and behaviors gather. Identifying these communities plays a key role in analyzing network structure, predicting new relationships, optimizing information flow, and even identifying anomalous behaviors. Much work has been done in the field of community detection, which is generally divided into two categories: global algorithms and local algorithms.

Global algorithms: These algorithms try to identify communities in the best possible way by analyzing the entire network structure. Methods based on optimizing criteria such as modularity and methods based on hierarchical clustering are among the prominent examples of this category. The advantage of these algorithms is that they provide a comprehensive view of the network structure, but they often face limitations due to high computational complexity and inefficiency in large and dynamic networks [5-8].

Local algorithms: These methods focus on analyzing communities in smaller parts of the network. Since they do not need to analyze the entire network structure, these algorithms are faster and work much more efficiently in dynamic networks or with incomplete information. However, their main challenge is to achieve a comprehensive and optimal view in identifying communities [9-16].

In the following, we will review some of the most important research and previous achievements in this field.

The Girvan-Newman algorithm is one of the primary global algorithms for community detection in networks. It decomposes the network structure and identifies communities by gradually removing edges with the highest betweenness value. One of the advantages of this method is its ability to accurately separate communities in small networks. However, its high computational complexity reduces its efficiency in large networks [17, 18].

The Louvain algorithm is one of the most efficient global algorithms for community detection that follows hierarchical modularity optimization. This method assigns nodes to initial communities in two steps and then combines these communities as new units. Its high speed and good performance in large networks are the most important advantages of this algorithm. However, the possibility of getting stuck in local optima is one of its challenges [19].

The Infomap algorithm is based on the theory of information flow in networks. It seeks to compress information into possible routing paths of nodes to identify communities. The algorithm's excellent performance in large and complex networks is one of its important advantages. However, tuning the appropriate parameters to optimize its performance is a challenge [20].

The Label Propagation Algorithm is a simple and efficient local algorithm for community detection. It detects communities by iteratively propagating labels among network nodes. The advantages of this algorithm are its high execution speed and no need for initial parameters. However, its results can be unstable due to the random nature of the execution [3].

In [21], a new local algorithm called LCD-SN is introduced. This algorithm identifies communities using the information of the first- and second-degree neighbors of each node. The results show that LCD-SN performs well in detecting communities with high accuracy and reasonable efficiency.

3. The Proposed Contributions

$G = (V, E)$ is an undirected and weighted network with a set of vertices $V(G) = \{v_1, v_2, \dots, v_n\}$, and a set of edges $E(G)$. The number of nodes is $|v| = n$, and the number of edges is $|E| = m$. Each edge connects a pair of vertices (v_i, v_j). The neighborhood set $\Gamma(v_i)$ is all the nodes connected to a vertex v_i. The size $|\Gamma(v_i)|$ is called the degree of a vertex v_i. There are many low-degree nodes and few high-degree nodes in a community, which have many common neighbors [22].

The Jaccard Similarity measures the similarity between two sets by comparing the number of common elements to the total number of unique elements. The formula is as follows:

$$\mathrm{Jaccard(i,j)} = \frac{|\mathrm{N_i} \cap \mathrm{N_j}|}{|\mathrm{N_i} \cup \mathrm{N_j}|} \tag{1}$$

In relation (1), N_i is the first-degree neighbors of node i [23].

- **Harmonic Mean**

In statistics, the harmonic mean is a measure of central tendency. The harmonic mean is similar to a harmonic sequence in the form of equation (2), and for this reason it is called the harmonic mean [24].

$$\mathrm{H} = \frac{\mathrm{n}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\mathrm{x_i}}} = \frac{\mathrm{n}}{\frac{1}{\mathrm{x_1}} + \frac{1}{\mathrm{x_2}} + \dots + \frac{1}{\mathrm{x_n}}} \quad \mathrm{x_i} > 0 \text{, for all i.} \tag{2}$$

This average can be calculated for two values x_1 and x_2 as follows.

$$\mathrm{H} = \frac{2\mathrm{x_1 x_2}}{\mathrm{x_1 + x_2}} \tag{3}$$

Harmonic average is used when two different rates are involved in a common workload. In this study, we will use harmonic average to calculate the average impact of first and second order neighboring nodes.

3.1. Proposed centrality criterion

In this section, a new centrality measure based on the first- and second-degree neighbors of each node is presented. To calculate the importance of nodes, the structural features and local importance of nodes are used. In social networks, the connections between nodes are based on their importance. In these networks, the post of an individual who is more famous or important among the group members is discussed and republished more often. Usually, important people in a network do not republish tweets and posts of ordinary people. Therefore, to calculate the influence of individuals in a local network, we must pay attention to their neighbors. The less important the people in the neighborhood of an individual are, the more authoritative the person in question (node) will be in that community and the opinions published in the same local network will be of higher importance. Therefore, taking this into account, to obtain the local importance of a node, we consider the degrees of the nodes of the first- and second-degree neighbors of the node and present our centrality measure.

$$\mathrm{RGN}(\alpha)=\frac{2*\sum_{i\in\Gamma(\alpha)}\frac{1}{d(i)}*\sum_{j\in\Gamma(i)}\frac{1}{d(j)}}{\sum_{i\in\Gamma(\alpha)}\frac{1}{d(i)}+\sum_{j\in\Gamma(i)}\frac{1}{d(j)}} \quad (4)$$

In equation (4), the strength and importance of a node is calculated based on the harmonic mean of the degrees of the first and second order neighboring nodes of node α, and $\Gamma(\alpha)$ is the number of neighbors of node α, $\Gamma(i)$ is the number of second order neighbors of α, and $d(i)$ is the degree of node i. The obtained value will be the importance or strength of node α in equation (4).

- **Evaluate of proposed centrality**

For example, in the following network consisting of three clear communities, if the nodes selected as the centers of the initial communities are inappropriate, the extracted communities will be of low quality. The ranking of the nodes of the network Fig. 1 given in Table (1), as can be seen, the nodes selected by the proposed ranking method are appropriate and the important nodes of each community have been correctly identified. The importance of the proposed criterion for ranking nodes in a social network is due to the selection of the importance of the nodes locally, as you can see in Fig. 1 on the left, the three important nodes with the proposed ranking are 3, 11, 4, respectively. The important nodes within the local communities are appropriately selected and by considering them as the core of the initial communities, we will not lose any information about the relationships between the nodes of the communities. In most global centrality methods such as PageRank, Betweenness, and Eigenvector, three important nodes are usually placed in a community, which means that one of the communities cannot be identified. In the best-case scenario, the centrality criterion for this graph is Closeness, which will be in the form on the right. Due to the failure to select the cores of the initial communities, the extracted communities are two communities, which will lose the relationships between one community.

Table 1. The RNG value calculated for the network nodes is shown in Fig. 1.

Node	Score	Degree	Rank
4	2.1962	5	1
2	1.7959	4	2
8	1.7698	4	3
6	1.7008	4	4
9	1.6187	4	5
11	1.6187	4	6
12	1.5956	4	7
3	1.3407	3	8
7	1.1339	3	9
10	1.1186	3	10
1	0.8116	2	11
5	0.7252	2	12

In the first step, a value called RGN is calculated for each node. The RGN(α) index for node α indicates its local importance based on the degree of its first and second order neighbors α, meaning how much higher importance and influence does node α have among local nodes. The larger the value of this index, the higher its relative importance compared to its neighbors, and the highest value corresponds to the strongest node in the network. If a node has a lower value among its neighbors, it should not be selected as the initial community core. Implementation of the proposed algorithm in the Fig. 1 network:

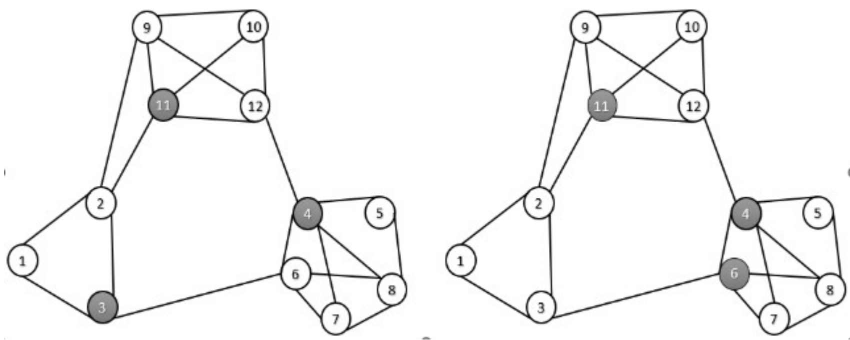

Fig. 1. How important nodes are selected by ranking algorithms - Figure on the right (by most algorithms) Figure on the left (by the proposed algorithm)

After ranking the nodes, node 4 is selected as the most important node and all its connected neighbors are placed in a community, and the assigned nodes are removed from the list of ranked nodes:
Ranked Node={4,2,8,6,9,11,12,3,7,10,1,5}

Selected Node = 4

C1= {4,5,6,7,8,12}

Ranked Node= {2,9,11,3,10,1}

Now we form node 2 of choice and the community around it.

Selected Node= 2

C2= {2, 1, 3, 9, 11}

Ranked Node= {10}

Now we form node 10 of choice and the community around it.

Selected Node=10

C3= {10, 9, 11, 12}

Ranked Node= {}

The first phase is complete and now we start the second phase, which is the assignment of overlapping nodes. In this step, using the modified Leicht-Holme-Newman similarity criterion, we assign the overlapping nodes to the appropriate community based on their similarity to the nodes in the neighboring communities. Nodes 9 and 11 overlap between clusters 2 and 3, and node 12 overlaps between clusters 1 and 3. Using the Leicht-Holme-Newman similarity criterion, nodes 9 and 11 are moved to cluster 3 and node 12 to cluster 3. At the end of this step, the communities will be as follows:

C1= {4, 5, 6, 7, 8}

C2= {2, 1, 3}

C3= {10, 9, 11, 12}

Due to the small size of the network, the output at the end of the merging stage will also be the same as the output of the second stage.

3.2. The proposed Algorithm

The proposed algorithm is implemented in three stages as follows:

Core selection: After calculating the importance of the nodes, they are sorted in descending order and a number of nodes are selected as cores based on a threshold. The threshold criterion is that nodes whose importance value is greater than the average importance of all nodes are identified as cores. This method automatically identifies high-importance nodes and considers them as the central points of the communities.

Initial community formation: In the initial community formation stage, instead of simply assigning neighboring nodes to the cores, a prioritization is first performed among the neighbors of the core nodes. In this stage, neighbors that have more connections to the core nodes will have a higher priority to join the initial communities. In other words, neighboring nodes that have a greater number of direct connections to the core nodes are selected as the main members of the communities. This prioritization is based on scoring the neighbors of the core nodes, so that each neighboring node is scored in proportion to the number of direct connections with its core nodes. Nodes that have several direct connections with core nodes receive a higher score and are more likely to join the corresponding community. This method allows the initial communities to be formed purposefully and according to the intensity of their internal connections and avoid communities with weak connections.

Community integration: In this step, scattered communities are identified based on the density of internal connections and merged with coherent communities. In this step, the communities of scattered communities are identified, and the density of internal connections of the communities is calculated. For each community, its internal connection density is calculated as the ratio of the number of internal edges (i.e., edges that exist within the community) to the number of possible edges (if all members of the community are connected). The density of internal connections is defined as:

$$C_{in}(C) = \frac{Number\ of\ internal\ edges\ within\ C}{\binom{|C|}{2}} \quad (5)$$

Where |C| is the number of members of community C and $\binom{|C|}{2}$ is the number of possible edges in community C (i.e., if all members of the community are connected).

Communities whose density of internal connections is less than a certain threshold (e.g., 0.4) are identified as weak communities. This threshold can be chosen based on experience or previous analyses. In general, the lower the density of internal connections, the weaker and more dispersed the community.

After identifying weak communities, these communities should be merged with stronger communities. In this step, a merging method is used that is based on criteria such as proximity to neighboring communities or common connections between members of weak and strong communities.

4. Results

In this section, the results of implementing the proposed algorithm on real and synthetic datasets will be compared with 5 well-known community detection algorithms in the literature based on their performance using NMI and modularity indices. In the following, after introducing the datasets provided and the indices used, you will see the results of the comparisons.

4.1. Datasets

In the comparison of the proposed algorithm, 13 real datasets and 4 synthetic datasets were used according to the procedure introduced by Lancichinetti et al. in [25] and the Girvan–Newman (GN) dataset [18]. Table 2 summarizes the information of these real datasets and Table 3 presents the parameters used to create the synthetic networks.

- **LFR and GN synthetic networks**

For a more accurate comparison of the proposed algorithm, four LFR synthetic networks and a GN synthetic network with the parameters listed in Table 3 have also been used. If the input parameter values in the LFR dataset are assigned according to Table 3, the GN dataset will be generated. As can be seen in Table 3, the 128 nodes in this dataset are divided into four communities with 32 nodes. The edges between pairs of nodes are placed randomly with the probability P_{in} for nodes belonging to the same community and P_{out} for nodes belonging to other communities and $P_{in} > P_{out}$. The mixing parameter μ for values of 0.1 to 0.8 determines the ratio of the average internal degree of each node to the external degree of each node. The higher the value of μ, the less the network has a community structure and the more difficult it is to extract the community structure in these networks.

Table 2. Real datasets examined by the proposed algorithm

Network	N	M
Karate	34	78
Dolphins	62	318
Polbooks	105	882
Football	115	1232
Email	1133	5451
Power Grid	4941	5694
PGP	10680	24316
Internet	22963	96872
Cond-mat-2003	31163	240058
Email Enron	36692	183831
DBLP	317080	1049866
Amazon	334863	925872
Youtube	1134890	2987624

Table 3. Parameters used to generate LFR synthetic networks.

Network	N	k_n	k_{max}	μ	γ	β	C_{min}	C_{max}
LFR-1	5000	20	50	0.1-0.8	2	1	10	50
LFR-2	5000	20	50	0.1-0.8	2	1	20	100
LFR-3	10000	20	50	0.1-0.8	2	1	10	50
LFR-4	10000	20	50	0.1-0.8	2	1	20	100
GN	128	16	16	1-5	0	0	32	32

N is the number of nodes in the network, k_n is the average degree of each node, k_{max} is the maximum degree of each node, μ is the composition parameter, γ and β are the powers of the number of nodes and the size of the network, C_{min} is the minimum community size and C_{max} is the maximum community size.

4.2. Test and comparison indicators

The right algorithms should find the right separation of communities, but what is the definition of a right separation in communities? To compare a good and bad separation from each other, a series of simple and understandable indicators should be defined so that they can be used to compare different algorithms. The "quality function" is a function that takes a specific separation and produces a number in a specific range in the output, so that the quality of different separations can be compared.

We denote the quality function by Q and call it sociable if, for each member of its community $C \epsilon P$, we have a primary function $q(x)$ for which the following equation holds:

$$Q(P) = \sum_{C \epsilon P} q(C) \tag{6}$$

Although not required, most functions in the community detection literature satisfy this property. In the following, two well-known functions for evaluating the quality of the presented separations will be explained.

- **Modularity**

The most popular quality function is the modularity function, which was introduced by Newman and Girvan in [17]. The main idea behind the modularity function is that in random networks there is no expectation of social structure. Therefore, the probability of the existence of communities is obtained by comparing the true density of edges with the mathematical expectation of the density of edges in the absence of social structure in the network. The expected value of the edge density in networks without social structure is obtained from the null model. The definition of modularity can be written as follows:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i . C_j) \tag{7}$$

Where A is the adjacency matrix, m is the total number of edges, P_{ij} is the arithmetic expectation of the number of edges between node i and node j in the null model, and δ is a function with an output of 1 if two nodes i and j are in the same community ($C_i = C_j$), and with an output of 0 if the two nodes are not in the same community.

As is clear from the formula, if the entire network is considered as a single community, the resulting modularity will be equal to 0. The modularity value can grow under the influence of the network size, and for this reason, the modularity value of two networks of different sizes cannot be compared. Modularity is always less than one, but it can also take on negative values. For example, in a division where each node is considered as a community, a negative numerical modularity value will be produced.

Modularity is not only the most important indicator of the quality of partitioning, but also many methods based on modularity optimization have been proposed. However, the scalability limitation of modularity should be mentioned in the case of these methods. In [26], Lancichinetti et al. prove that modularity optimization-based methods are not able to detect communities smaller than a certain scale, i.e., if there is a fully connected but small subnet with respect to the entire network, modularity-based methods will not consider that subnet as a community.

- **Normalized Mutual Information (NMI)**

Normalized Mutual Information (NMI) is an information theory-based index that compares the quality of extracted communities in terms of their correspondence with real-world observations (the true state). Since NMI requires knowledge of the true state of communities, it cannot be used in datasets that do not provide such information [27].

Assuming the extracted communities $C = \{C_1 . C_2 . \dots . C_q\}$ and the ground-truth communities $C' = \{C'_1 . C'_2 . \dots . C'_k\}$, the NMI value can be calculated as follows:

$$\mathrm{NMI}(\mathrm{C}.\mathrm{C}') = \frac{2\mathrm{I}(\mathrm{C}.\mathrm{C}')}{\mathrm{H}(\mathrm{C}) + \mathrm{H}(\mathrm{C}')} \quad (8)$$

Where $I(X.Y)$ represents the mutual information and $H(X)$ represents the uncertainty factor or entropy. We can calculate each of the above as follows:

$$\mathrm{I}(\mathrm{C}.\mathrm{C}') = \mathrm{H}(\mathrm{C}) + \mathrm{H}(\mathrm{C}') - \mathrm{H}(\mathrm{C}.\mathrm{C}') \quad (9)$$

$$\mathrm{H}(\mathrm{C}) = -\sum_{\mathrm{i}=1}^{\mathrm{q}} \frac{|\mathrm{C_i}|}{\mathrm{n}} \log \frac{|\mathrm{C_i}|}{\mathrm{n}} \quad (10)$$

$H(C.C')$ is also called Joint Entropy and is calculated as follows:

$$\mathrm{H}(\mathrm{C}.\mathrm{C}') = -\sum_{\mathrm{i}=1}^{\mathrm{q}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \frac{|\mathrm{C_i} \cap \mathrm{C'_j}|}{\mathrm{n}} \log \frac{|\mathrm{C_i} \cap \mathrm{C'_j}|}{\mathrm{n}} \quad (11)$$

5. Comparison and Analysis of Results

In this section, the results of the proposed algorithm are compared with 9 algorithms CNM, Infomap, Louvain, NIBLPA, LPA, Intimacy-LPA, LSMD, CLPR and CMA. The results of the observations show the quality and accuracy of the proposed algorithm compared to other algorithms. We selected basic and well-known algorithms for evaluation. The selected algorithms include global (based on random walk, greedy-based, modularity optimization, based on diffusion) and local (based on selecting important nodes and label diffusion) methods. The proposed method has a good performance similar to the best global algorithms presented. The results of the review of the presented algorithms show that no algorithm has high quality in all of them due to the different structure of the datasets (clustering coefficient, average degree of each node, network diameter, etc.), but the proposed method has the ability to extract high-quality communities close to the real discovered communities in most datasets. The multiplicity for community detection in different algorithms is shown in Table 4.It should be noted that the modularity criterion value is usually higher in global algorithms, and this does not necessarily mean that the proposed algorithm is weak. Some basic algorithms such as Louvain, CNM, which are global methods based on maximizing the modularity value, and the Infomap algorithm, a global method based on information theory and random steps, have a high modularity value for most datasets. The proposed algorithm, using a semi-local method based on ranking and node similarity, is able to extract communities close to real communities. This fact is evident in datasets with Ground Truth, and it may have a low modularity value but a higher NMI value, which indicates the accuracy of the proposed algorithm in correctly detecting communities in real and synthetic datasets. By examining Table 4, which compares the NMI index for datasets with ground-truth, it is observed that in datasets such as Zachary Karate Club, Dolphin, Pool Box, YouTube, Amazon, and DBLP, which have characteristics close to Real World networks, the proposed algorithm has the highest NMI. In Fig. 5 you can see the comparison of NMI values.

Algorithm	CNM	Infomap	Louvain	LPA	NIBLPA	LPA-Intimacy	CSLPR	CMA	DOSN
Karate	0.69	0.69	0.59	0.59	0.58	1	1	1	1
Dolphins	0.55	0.53	0.48	0.52	0.5	0.63	0.541	0.52	1
Polbooks	0.53	0.53	0.51	0.54	0.53	0.54	0.574	0.56	0.599
Football	0.75	0.95	0.88	0.92	0.72	0.86	0.828	0.81	0.896
DBLP	0.16	0.65	0.13	0.64	0.46	0.61	0.682	0.65	0.7145
Amazon	0.11	0.6	0.11	0.54	0.6	0.63	0.676	0.55	0.6912
YouTube	-	0.13	0.06	0.07	-	-	0.471	0.32	0.42

Table 4. Comparison of NMI of the proposed algorithm on real datasets

Table 5 shows the results of the modularity criterion test in real datasets. In the Karate dataset, the highest modularity value Q=0.4198 with 2 communities belongs to the CMA algorithm, which is not an advantage, but correct community detection in the real world is more important, which the proposed algorithm has achieved.

In the Dolphin dataset, the highest value of the modularity criterion belongs to the Infomap algorithm with a value of Q=0.527 and the number of extracted communities is 6, which in the real world is 2 communities and the actual modularity of the dataset is Q=0.3787. In the proposed algorithm, the number of extracted communities is 2 and the NMI value is 1 for the Dolphin dataset. In the case of the ca-GrQc dataset, the value of the modularity criterion is Q=0.827 and for the ca-HepTh dataset, it is Q=0.73, which is higher than all the algorithms.

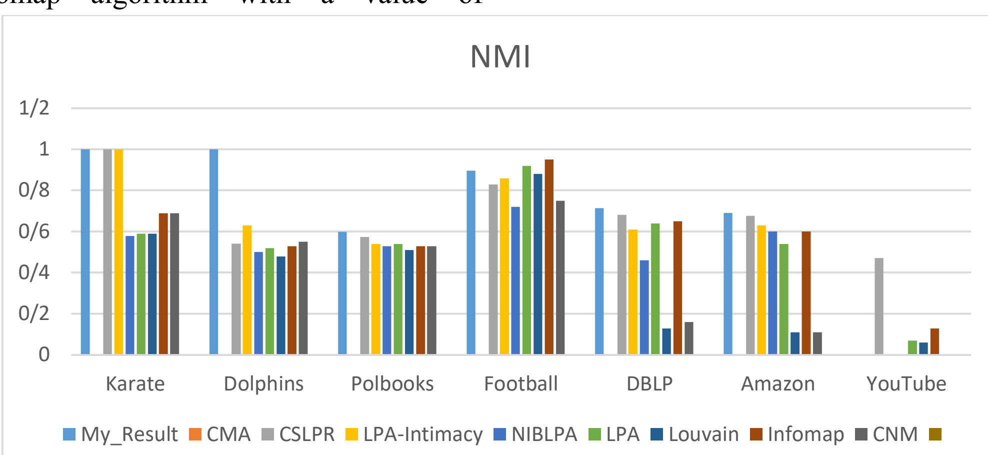

Fig. 2. Comparison of NMI and the proposed algorithm for datasets with Ground Truth

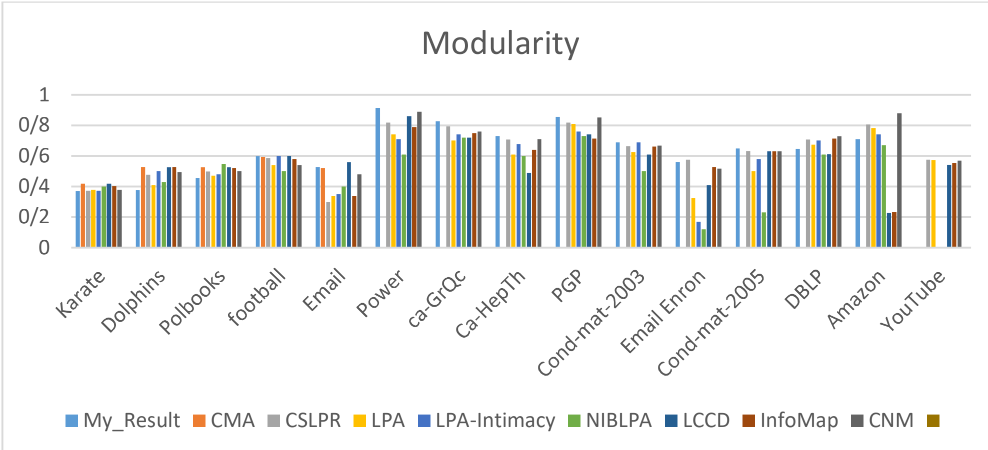

Fig. 3. Comparison of modularity and the proposed algorithm for datasets with Ground Truth

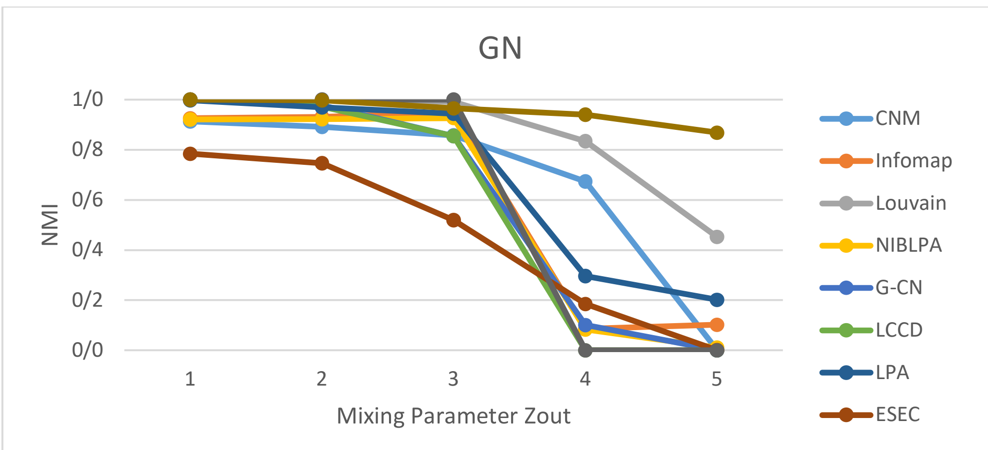

Fig. 4. Testing the proposed algorithm based on the obtained NMI for the synthetic GN network

In the case of the Power dataset, the value of the modularity criterion is Q=0.915, which is higher than the global algorithms CNM and Infomap and is also higher than all the local algorithms. In the PGP, Cond-mat-2003, Email Enron, Cond-mat-2005 datasets, the value of the modularity criterion is higher than all the algorithms under study and is only behind the Louvain algorithm. A noteworthy and important point about the proposed algorithm is that it has the best Q and NMI values among local algorithms.

The experimental results for the synthetic GN dataset are given in Fig. 7, which shows that the proposed algorithm accurately detects communities for $1 \leq Z_{out} \leq 5$, which is the best result among the compared algorithms and indicates good performance among the investigated algorithms.

The experimental results for the synthetic LFR datasets are shown in Fig. 5, where each point represents the average performance of different algorithms on the LFR benchmark networks with 10 runs. It is observed that for all μ, the proposed algorithm has acceptable NMI values.

Table 5. Comparing modularity in real datasets

Algorithm	CNM	InfomMap	Louvain	LCCD	NIBLPA	LPA-Intimacy	LPA	LSMD	CSLPR	CMA	CDHC
Karate	0.38	0.401	0.41	0.419	0.4	0.3715	0.38	0.3715	0.3715	0.4198	0.371
Dolphins	0.495	0.527	0.51	0.525	0.43	0.5	0.41	0.3787	0.478	0.5269	0.3787
Pollbooks	0.501	0.522	0.52	0.525	0.55	0.48	0.47	0.446	0.499	0.5248	0.457
football	0.54	0.58	0.60	0.60	0.5	0.60	0.54	0.58	0.586	0.5951	0.599
Email	0.48	0.34	0.54	0.56	0.40	0.35	0.34	0.42	0.299	0.5210	0.528
Power	0.89	0.79	0.94	0.86	0.61	0.71	0.74	0.793	0.819	0.77	0.915
ca-GrQc	0.76	0.75	0.76	0.72	0.72	0.74	0.7	0.771	0.794	0.79	**0.827**
Ca-HepTh	0.71	0.64	0.7	0.49	0.6	0.68	0.61	0.63	0.706	0.701	**0.73**
PGP	0.852	0.715	0.87	0.74	0.73	0.76	0.81	0.59	0.819	0.80	**0.857**
Cond-mat-2003	0.668	0.661	0.72	0.61	0.5	0.69	0.627	0.57	0.664	0.621	0.689
Email Enron	0.517	0.527	0.55	0.41	0.12	0.17	0.324	0.39	0.575	0.46	0.561
Cond-mat-2005	0.63	0.63	0.71	0.63	0.23	0.58	0.5	0.44	0.632	0.64	0.649
DBLP	0.728	0.714	0.81	0.612	0.61	0.7	0.674	0.65	0.707	0.59	0.648
Amazon	0.879	0.232	0.9	0.229	0.67	0.74	0.783	0.68	0.806	0.65	0.708
YouTube	0.569	0.556	---	0.544	---	---	0.573	0.42	0.575	0.49	0.51

6. Conclusion

In this paper, a new method for local detection of communities based on semi-local ranking of nodes in social networks was designed and presented. In the first phase, the method, after ranking the nodes, starts from the most important node of the network and places all their direct neighbors in a community and performs the same process for other important nodes, observing the condition that if two important nodes are directly connected to each other, the next important node will be a member of the community of the more important node and will be removed from the list of important nodes. In the second phase, we determine the assignment of common (overlapping) nodes in the initial communities with the modified Leicht-Holme-Newman similarity criterion. In the third phase, we examine the extracted communities and assign the communities whose number of nodes is less than 3 again as in the second phase, and after this phase, we merge the weak communities together to increase the quality of the final communities. One of the advantages of the proposed method is the formation of initial communities using the structural features of the network that have the least spatial and temporal complexity and reduce the number of uses of the similarity index between nodes. The most important characteristics of the proposed algorithm are stability, determinism, high quality, and accuracy in correctly extracting communities in real and synthetic networks. In future research, we will try to reduce the time required to search for suitable communities for integration by using crowd intelligence or metaheuristic methods.

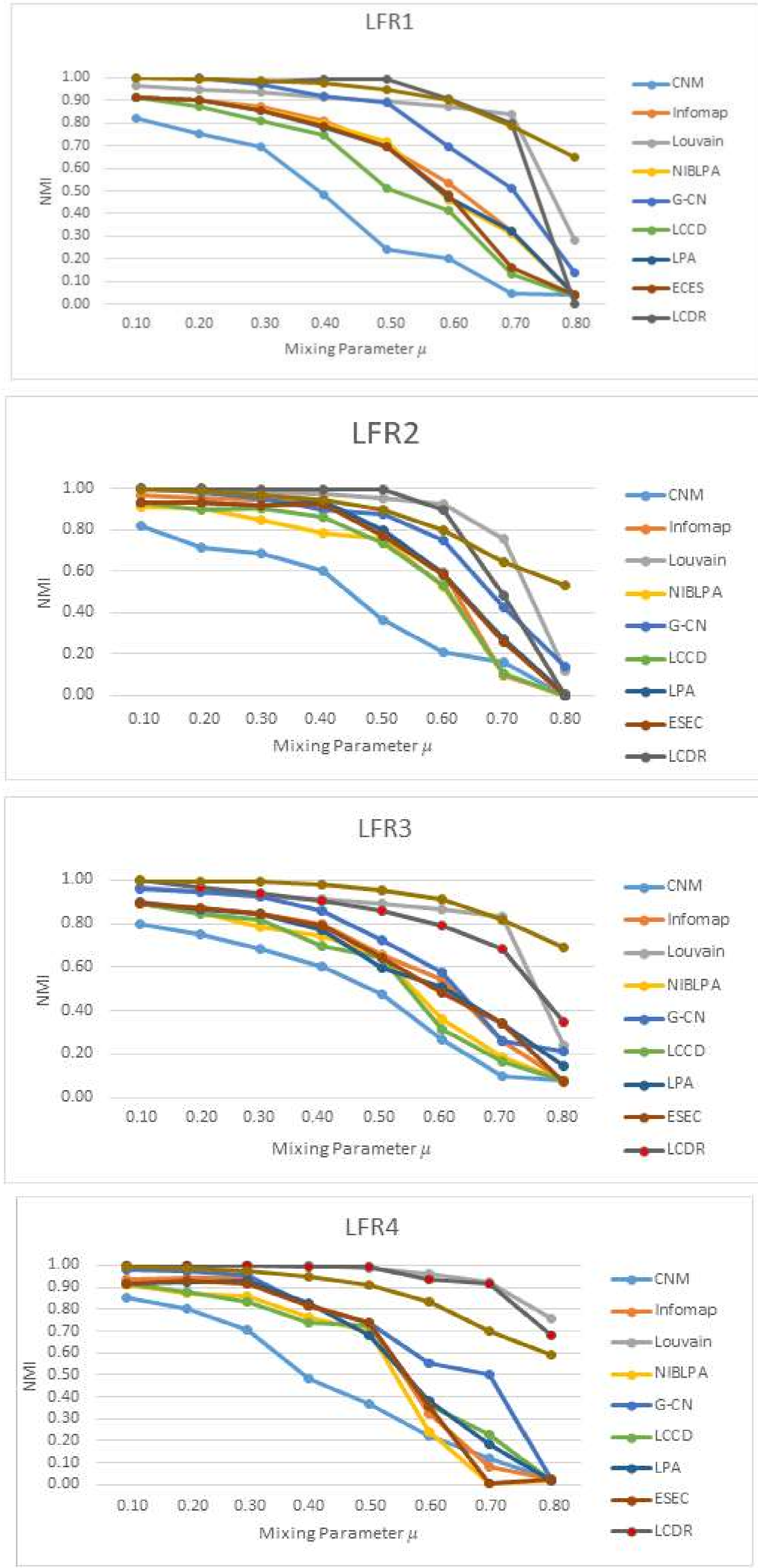

Fig. 5. Testing the proposed algorithm based on the obtained NMI for the synthetic LFR network

REFERENCES

[1] B. H. Good, Y.-A. De Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, vol. 81, no. 4, p. 046106, 2010.
[2] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," Proceedings of the national academy of sciences, vol. 107, no. 29, pp. 12755-12760, 2010.
[3] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Physical review E, vol. 76, no. 3, p. 036106, 2007.
[4] H. Zhou and R. Lipowsky, "Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities," in Computational Science-ICCS 2004: 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part III 4, 2004: Springer, pp. 1062-1069.
[5] K. Berahmand, M. Mohammadi, R. Sheikhpour, Y. Li, and Y. Xu, "WSNMF: Weighted symmetric nonnegative matrix factorization for attributed graph clustering," Neurocomputing, vol. 566, p. 127041, 2024.
[6] K. Berahmand, Y. Li, and Y. Xu, "DAC-HPP: deep attributed clustering with high-order proximity preserve," Neural Computing and Applications, vol. 35, no. 34, pp. 24493-24511, 2023.
[7] B. Zarei, M. R. Meybodi, and B. Masoumi, "A new evolutionary model based on cellular learning automata and chaos theory," New Generation Computing, vol. 40, no. 1, pp. 285-310, 2022.
[8] B. Zarei, M. R. Meybodi, and B. Masoumi, "Chaotic memetic algorithm and its application for detecting community structure in complex networks," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 30, no. 1, 2020.
[9] J. Sheykhzadeh, Zarei, B., & Gharehchopogh, F. S., "Community detection in social networks using a local approach based on node ranking," IEEE Access, 2024.
[10] H. Li et al., "LMFLS: A new fast local multi-factor node scoring and label selection-based algorithm for community detection," Chaos, Solitons & Fractals, vol. 185, p. 115126, 2024.
[11] A. Bouyer, P. Shahgholi, B. Arasteh, and E. B. Tirkolaee, "Local core expanding-based label diffusion and local deep embedding for fast community detection algorithm in social networks," Computers and Electrical Engineering, vol. 119, A, 2024.
[12] S. Wang, J. Yang, X. Ding, J. Zhang, and M. Zhao, "A local community detection algorithm based on potential community exploration," Frontiers in Physics, vol. 11, p. 1114296, 2023.
[13] W. Louafi and F. Titouna, "PCMeans: community detection using local PageRank, clustering, and K-means," Social Network Analysis and Mining, vol. 13, no. 1, p. 103, 2023.
[14] H. Roghani and A. Bouyer, "A fast local balanced label diffusion algorithm for community detection in social networks," IEEE Transactions on Knowledge and Data Engineering, 2022.
[15] S. Aghaalizadeh, S. T. Afshord, A. Bouyer, and B. Anari, "A three-stage algorithm for local community detection based on the high node importance ranking in social networks," Physica A: Statistical Mechanics and its Applications, vol. 563, p. 125420, 2021.
[16] X. Ding, J. Zhang, and J. Yang, "A robust two-stage algorithm for local community detection," Knowledge-Based Systems, vol. 152, pp. 188-199, 2018.
[17] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, no. 2, p. 026113, 2004.
[18] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821-7826, 2002.
[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.
[20] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proceedings of the national academy of sciences, vol. 105, no. 4, pp. 1118-1123, 2008.
[21] F. D. Zarandi and M. K. Rafsanjani, "Community detection in complex networks using structural similarity," Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 882-891, 2018.
[22] J. L. Gross, J. Yellen, and M. Anderson, Graph theory and its applications. Chapman and Hall/CRC, 2018.
[23] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," Physical Review E, vol. 73, no. 2, p. 026120, 2006.
[24] J. Gow, A history of Greek mathematics. Cambridge University Press, 1981.
[25] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical review E, vol. 78, no. 4, p. 046110, 2008.

[26] A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," Physical review E, vol. 84, no. 6, p. 066122, 2011.

[27] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," Journal of statistical mechanics: Theory and experiment, vol. 2005, no. 09, p. P09008, 2005.