Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Motif-Based Community Detection: A Probabilistic Model Based on Repeating Patterns

H. Hajibabaei [1], V. Seydi [1,2,*], A. Koochari [2]

[1]*Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*

[2]*Centre for Applied Marine Sciences, School of Ocean Sciences, Bangor University, Menai Bridge, UK.*

Article Info

Article History:
Received 13 August 2023
Reviewed 15 October 2023
Revised 05 November 2023
Accepted 04 December 2023

Keywords:
Community detection
Motif
Complex networks
Probabilistic model

*Corresponding Author's Email Address: *V.seydi@bangor.ac.uk*

Abstract

Background and Objectives: The detection of community in networks is an important tool for revealing hidden data in network analysis. One of the signs that the community exists in the network is the neighborhood density between nodes. Also, the existence of a concept called a motif indicates that a community with a high edge density has a correlation between nodes that goes beyond their close neighbors. Motifs are repetitive edge patterns that are frequently seen in the network.

Methods: By estimating the triangular motif in the network, our proposed probabilistic motif-based community detection model (PMCD) helps to find the communities in the network. The idea of the proposed model is network analysis based on structural density between nodes and detecting communities by estimating motifs using probabilistic methods.

Results: The suggested model's output is the strength of each node's affiliation to the communities and detecting overlaps in communities. To evaluate the performance and accuracy of the proposed method, experiments are done on real-world and synthetic networks. The findings show that, compared to other algorithms, the proposed method is acting more accurately and densely in detecting communities.

Conclusion: The advantage of PMCD in using the probabilistic generative model is speeding up the computation of the hidden parameters and establishing the community based on the likelihood of triangular motifs. In fact, the proposed method proves there is a probabilistic correlation between the observation of two node pairs in different communities and the increased existence of motif structure in the network.

Introduction

For effective network component monitoring and recognition, network analysis is a key tool. A complex network [1] can consist of cells in biology [2], social networks with friendly communication [3], or a network of scientists doing joint scientific studies [4]. To put it another way, it can be any grid with nodes and edges that can be represented as a graph. One of the most effective methods and methodologies to analyze complex networks is community detection.

Community detection identifies subgraphs of a network whose relationship among their nodes is more robust and dense than those between other nodes of the network [5]. A community can represent an idea, a group, an interest, a focus on a particular topic, and so on. Communities can be used separately or together; the latter are referred to as overlap communities. The machine learning clustering topic's component,

community detection, has the potential to be applied in a variety of engineering fields, such as text classification, traffic network optimization, and social network analysis. The goal of community detection is to group the nodes of a network into different communities so that they are strongly connected or have similar node features [5]. A key problem in dynamical network research is the discovery of communities with the aim of revealing hidden features of a complex network, which are frequently densely coupled nodes [6].

Community detection in networks is an NP-hard issue that categorized from different perspectives. These categories include weighted [7], [8] and unweighted [9], [10], directed [11] and undirected [12], global [13] and local [12], overlapping [14], [15], and non-overlapping [16] community discovery techniques. Different community detection techniques were developed based on these criteria. Examples include model-based approaches [7], [14], [17], clique percolation methods [18], modularity-based methods [4], [19], [20], label propagation methods [11], [21]-[23], model-based methods [7], [14], methods for network embedding [24] and community detection methods with deep learning [25]-[27]. It can be seen from examining the different approaches used for community detection and the research on this topic that a straightforward analysis of node properties won't produce the accuracy needed for community detection in networks; rather, taking a deeper look at the networks' particulars and using the graphs' original characteristics, like motif structure, will produce better results.

The PMCD method employs a probabilistic relation to find communities in complex networks. We extend probabilistic model-based methods from edge creation to motif generation. Complex networks commonly contain "motifs", which are a type of small, linked sub-networks. Based on empirical studies, communities with similar nodes have related motifs. As a result, using motifs with lots of connections can be a useful strategy for finding communities and performing more accurate network analysis [28]. We show that the chance of a triangle motif existing between three nodes in shared communities grows with the observation of more nodes in such communities. In other words, we locate the hidden parameter of the probabilistic model and find the community by using the triangle motif. We define the triangular motif estimator function as a Bernoulli loss function over one node and two of that node's neighbors for the probabilistic motif generator's function. We also research how community overlap affects how motifs are generated.

Related Works

The problem of community detection in complex networks gets a lot of attention. Several research projects on different aspects of community detection have been performed over the past few years. The first methods of community detection employed traditional techniques and clustering-based algorithms. These methods presented key ideas for community detection and laid the groundwork for future developments. Traditional approaches include graph partitioning, hierarchical, spectral and partitional clustering [29].

The algorithms used to detect communities based on modularity have been extensively studied and used due to their simple tactics and clear outcomes. However, they also encounter difficulties, such as communities that are unstable and sensitive to seed node selection [30]. One of these, the Louvain technique [5], is frequently applied to weighted graphs. This approach provides a straightforward and quick methodology to detect distinct communities and maximizes modularity by clustering network nodes using the greedy approaches [31]. The Leiden method, however, corrects numerous flaws in the Louvain algorithm [32]. The objective is to change the community developed throughout the iteration cycle while simultaneously speeding up local mobility and transferring nodes to arbitrary neighbors.

The label propagation algorithm (LPA), a practical community detection approach, was initially introduced in [22]. Although its simple design and low complexity are widely respected, there are several downsides, such as the randomness of node selection and label updating. In the LPA technique, a node is selected at random, and through an iterative process, its label is updated with the most prevalent label nearby [33]. To handle the weaknesses in the LPA methodology, the Speaker-Listener Label Propagation Algorithm (SLPA) [34] and the COPRA [35] were created.

Cliques are one of the fundamental ideas in graph analysis and are utilized to detect communities in networks. The clique percolation algorithm (CPM) [18] and CFinder [36] were proposed as overlapping community detection algorithms based on the clique percolation method's search for local patterns.

The motif is another idea related to the clique. Small, linked sub-networks known as motifs frequently appear in complex networks and are one of the basic elements of the network [37]. In network analysis, motifs are used to detect communities and comprehend network structure [38]. Motifs demonstrate that a community with a high edge density will have relationships between nodes that go beyond their immediate neighbors. Although a few motif-based community detection methods have been proposed [38]-[40], when used on large-scale networks, they frequently encounter high computational complexity. It is still challenging to properly and economically combine lower-order and higher-order structural data into a unified framework for community detection.

The group of methods estimates the probabilistic model to detect communities, in difference to the techniques cited at the top that employ traditional methods to do so. This method creates a generative

sample of the graph and estimates the model parameters [14], [17]. The degree of node dependency on communities is a parameter in the generative model that is estimated using methods [14], [17], [41] using a matrix factorization-based model. The algorithm [42] presents a matrix factorization-based paradigm that makes it easy to add or delete edges. The non-negative matrix factorization model of the community detection issue is also described [43] and a transfer matrix is then used to control the dynamics of the network structure.

Proposed Algorithm Frameworks

In this research, a probabilistic motif-based community detection model (PMCD) is presented that uses the triangle motif and the affiliation graph model to detect community structures. The core idea of the proposed community detection method is that a robust community requires taking the node's structural model and relationship types into consideration. Two nodes observed in more shared communities are more likely to be connected, according to Yang and Leskovec's study [14] on the connection between edge (2-clique) likelihood and community overlap. In this paper, we examine the impact of community overlapping on the evolution of the triangle and 3-clique motifs. We show that by increasing the number of nodes observed in shared communities, the probability of the existence of a triangular motif between them increases. This result is in accordance with the fundamental principle that vertices situated in communities' overlaps are more densely connected than vertices within a single community. By using the optimized Bernoulli loss function for probabilistic estimation, we can therefore enhance and increase the AGM's ability [14], [17] to generate triangle motifs.

The PMCD model is different from other community detection approaches in that it considers additional properties that were not sufficiently considered in earlier methods, for instance:

- Using edge density to detect communities.
- Triangle motif estimation using a probabilistic approach.
- The conceptual link between community detection and the likelihood of the triangular motif being present or absent.
- Use of evolutionary approaches and maximum likelihood estimation in computations.

Fig. 1 presents an example of a simple graph to illustrate the concept of three-node motifs. Two different three-node motif types discovered in Fig. 1 are shown in Fig. 2. Depending on the characteristics of the networks, the triangle motifs observed in various types of networks can be interpreted in multiple ways. For instance, the 3-node motif (3, e) and the 4-node motif (4, e) are the most widely studied motifs in complex networks [44]. The proposed method uses the triangular motif to build the hidden parameter of the probabilistic model and detect the community.

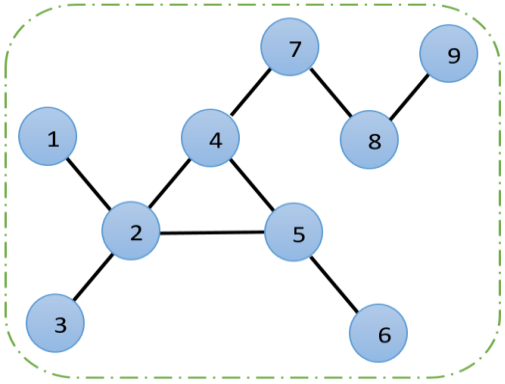

Fig. 1: Illustrated an example of a simple graph to illustrate the concept of three-node motifs.

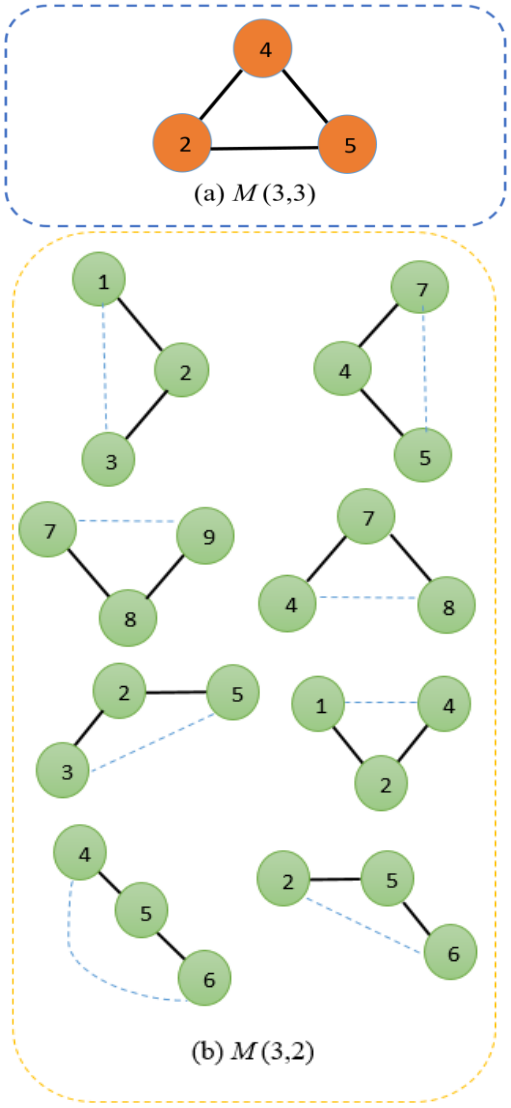

Fig. 2: Illustrated two types of three-node motifs that we use in the proposed model: (a) a 3-clique or closed triangle motif (denoted as M(3,3)-motif) with 3 nodes and 3 edges discovered from Fig. 1; (b) types of the opened triangle motif (denoted as M(3,2)-motif) with 3 nodes and 2 edges extracted from Fig. 1.

The PMCD model is based on a network G (N, E), where nodes and edges are referred to as N and E, respectively. We create M_{uc}, a nonnegative integer, to represent the strength of the node's affiliation with the community. (M_{uc} = 0 denotes u's non-membership in c.) The degree of reliance between each node and each community is thus shown in the M matrix.

The value of M in PMCD establishes whether or not a triangular motif between three nodes (u, v_1, and v_2) will appear in a community (c). Specifically, we presumed that three nodes, u, v_1, and v_2, are triangular motifs by taking into account the following likelihood. For the probabilistic motif generator's function, we define the triangular motif estimator function as a loss function over one node and two neighbors of that node, that is,

$$P_c(u,v_1,v_2) = \underset{u,v_1\in E}{P_c}(u,v_1).\underset{u,v_2\in E}{P_c}(u,v_2) = \left[1-\exp(-M_{uc}\cdot M_{v_1c}^T)\right]\cdot\left[1-\exp(-M_{uc}\cdot M_{v_2c}^T)\right] \quad (1)$$

Due to the generative probabilistic approach between two couples of nodes in a triangle motif, each couple of nodes is independently propagated by the Bernoulli model. Thus, per element of the adjacency matrix is formed on the following probabilistic method:

$$P(u,v_1,v_2) = \underset{v_1\in N(u)}{P}(u,v_1).\underset{v_2\in N(u)}{P}(u,v_2) = \left[1-\exp(-M_u\cdot M_{v_1}^T)\right]\cdot\left[1-\exp(-M_u\cdot M_{v_2}^T)\right] \quad (2)$$

$$A_{uv_1} \sim Bernoulli(P_{uv_1}) \odot A_{uv_2} \sim Bernoulli(P_{uv_2})$$

The framework of computation for (1) and (2), which describe a probabilistic generative model, is predicated on the following premises:

- In a community, a triangle motif can exist between two pairs of nodes (one node and two neighbors of that node).
- The probability of the existence triangle motif increases when two pairs of nodes are observed in multiple communities.
- Communities can overlap; communities that overlap have a higher density of triangle motifs.

Community detection by PMCD model

We defined the components of the PMCD model before illustrating how to utilize it for community detection in networks. The model parameter discussed in the preceding section is the degree of a node's community membership (M_{uc}). By maximizing the likelihood, we can get the optimum M as follows:

$$likelihood(M) = l(M) = log\ P(G\mid M)$$

$$\hat{M} = \arg\max l(M) = \arg\max_{M\geq 0} \prod_{(u,v)\in v} p(u,v_1,v_2) \prod_{(u,v)\notin v}(1-p(u,v_1,v_2)) = \arg\max_M \left[\prod_{(u,v)\in v} p(u,v_1)\cdot p(u,v_2)\right]\cdot \left[\prod_{(u,v)\notin v}(1-(p(u,v_1)\cdot p(u,v_2))\right] \quad (3)$$

After combining (2) and (3), a natural logarithm is needed to be computed on both sides to correct the multiplication to the aggregate and reduce the next computations. The logarithm is completely ascending; therefore, it won't interfere with the maximum likelihood estimation.

$$L(M) = \left[\sum_{(u,v_1)\in E}\log(1-\exp(-M_u.M_{v_1}^T)) - \sum_{(u,v_1)\notin E} M_u.M_{v_1}^T\right] + \left[\sum_{(u,v_2)\in E}\log(1-\exp(-M_u.M_{v_2}^T)) - \sum_{(u,v_2)\notin E} M_u.M_{v_2}^T\right] \quad (4)$$

Updating the Parameter

The non-linear likelihood function of (4), which contains the latent variable M, cannot be maximized by conventional optimization methods. We calculate the objective function in (4) using the *Block Coordinate Ascent* approach [45], which helps solve optimization problems with latent variables in machine learning. By maintaining fixed neighbors (M_v), we update M_u for each node u.

$$L(M_u) = \left[\sum_{v\in N(u)}\log(1-\exp(-M_u.M_{v_1}^T)) - \sum_{v\notin N(u)} M_u.M_{v_1}^T\right] + \left[\sum_{v\in N(u)}\log(1-\exp(-M_u.M_{v_2}^T)) - \sum_{v\notin N(u)} M_u.M_{v_2}^T\right] \quad (5)$$

In (5), *N(u)* is a set of neighbours of u. In order to calculate the maximum probability (the diagram's maximum point), we must find a location on the Figurative chart where the gradient equals 0. Thus, it is necessary to derive the partial derivation of the likelihood logarithmic in (5) than M_u.

$$\frac{\partial\ell(M_u)}{\partial M_u} = \left[\sum_{v_1\in N(u)} M_u \frac{\exp(-M_u M_{v_1})}{1-\exp(-M_u M_{v_1})} - \sum_{v_1\notin N(u)} M_{v_1}\right] + \left[\sum_{v_2\in N(u)} M_u \frac{\exp(-M_u M_{v_2})}{1-\exp(-M_u M_{v_2})} - \sum_{v_2\notin N(u)} M_{v_2}\right] \quad (6)$$

The gradient ascent algorithm will eventually update M_u values [46], [47]. A node's belonging strength to a community will be replaced with 0 if it detects it, as it is impossible for it to be negative.

$$M_u(t+1) = \max\left(0, M_u(t) - \eta\ \left(\frac{\partial\ell(M_u)}{\partial M_u}\right)\right) \quad (7)$$

In (5), η is a learning parameter, As long as the difference between the value from the last step and the current value is lower than the acceptable threshold, the process of updating each M_u at each stage of the algorithm iteration is repeated.

PMCD Algorithm

Algorithm 1 displays the proposed PMCD model (probabilistic motif-based community detection). A graph

(G) and the number of communities (k) are the method's inputs. The model also creates a matrix (M_{uc}) that displays the degree to which each node belongs to each community. When they are observed in different communities, the likelihood that there is an existing motif structure between two groups of nodes increases.

Since the hidden variable (M) is initialized (details of computing M addressed later), the method then begins an iterative process. After the difference among $M_u(t+1)$ and $M_u(t)$ was smaller than a predefined point (in this case, the stop threshold is 0.005), the iterations stop. In order to estimate the model's unknown parameter in the graph, this iterative method calculates the likelihood of the probabilistic model ($L(M_u)$). To derive the likelihood function's logarithm as close as possible to its maximum value (when the line's slope is 0), the likelihood function's logarithm is collected from each node u using the formula $D(L(M_u))$.

Algorithm 1: Probabilistic motif-based community detection (PMCD)

1: Inputs: Graph G = (N; E);
Number of communities (k);
2: Output: M_{uc} belonging of each node *u* Community c
3: $t \leftarrow 0$
4: M = *local_maximun_neighborhood()*
5: while $|M_u(t+1) - M_u(t)| <= 0.005$ do
6: t ← t + 1
7: **for** i = 1 to |V| do
8: $L(M) = log\ p(G|M)$
9: $D(L(M_u))$ = *Derivation_finder_L*(M_u)
10: Update: $M_u\ (t+1) =$
Gradient_ascent $(D(L(M_u)); M_u(t))$
11: **end for**
12: end while
13: for i = 1 to |V| do
14: **for** j = 1 to k do
15: **if** $M_{uc} > threshold$ then
16: Add: $c_j \leftarrow u_i$
17: **end if**
18: **end for**
19: end for

In line 10 of Algorithm 1**,** we chose the ascending gradient approach [46], [47] to optimize the probability because the computations were complex. This method is used to update the latent variable of the model (Mu) at each iteration of the algorithm. After the M value has been fixed, each node's ability to contribute per community is assessed. Since comparing this value to a testing point (such as the median of M), it may be defined as either belonging to or not belonging to the communities, and the output of the model will then be realized.

Computational Complexity

The count of communities and dense motifs affect the computational complexity of the PMCD method. The core concept of Algorithm 1, as shown in its iteration phases, is the rate of depending on the community, which is updated using (6) and (7). In this case, whether or not two nodes have neighbors who are members of one or more communities determines whether or not those nodes share a theme. Because of this, the computational complexity will depend on the number of communities present and the order of each node's neighbors (N(u)); in the worst case, this complexity will be O(2k.|E|).

Initialize

The matrix of depending strengths for the nodes communities can start in a variety of ways. The first option, which also appears to be the simplest, involves filling in the values at random. The algorithm's major drawback, however, is that it repeats the steps more frequently, increasing computing complexity as it advances to the model stability phases.

The other choice is the local minim neighborhood approach [48], which has been shown through studies to be an excellent starting point for community discovery algorithms. Using this method has the added benefit of being able to estimate the initial number of communities to start the proposed model's community detection phase, in addition to minimizing iteration steps and starting the process in a stable state.

Experiments

The proposed PMCD method has been implemented in the Spyder environment using the Python programming language. We used five real-world data sets (Table 2) and sixteen synthetic networks (Table 5), respectively, to evaluate the results. Additionally, the statistics include the node's "ground-truth" community memberships. In these datasets, the proposed method is compared with fundamental algorithms like Louvain [49], Leiden [32], Bigclam [14], [17], CPM [18], Label propagation [35], and SLPA [34]. Table 1 lists these algorithms in brief.

Evaluation Metrics

We evaluate the community detection algorithms' effectiveness and accuracy using three standard evaluation metrics. Modularity [50] is an internal metric for assessing community quality, whereas the F1_Score and NMI are external metrics for assessing community accuracy by comparing them to ground-truth communities [6]. The modularity measure in internal metrics, a popular benchmark for estimating the density in the community is derived from Girvan-Newman [50].

Table 1: The employed methods for PMCD evaluation

Method Name	Description
Louvain	Louvain amplifies the modularity value of communities
Leiden	The Leiden method is an advancement of the Louvain
Bigclam	The probabilistic community detection method that scales to large networks
CPM	Find k-clique communities in a graph using the percolation method
LPA	The label propagation algorithm detects communities by network structure
SLPA	SLPA is an overlapping community discovery that extends the LPA

By dividing the projected community edges by the expected community edges, the modularity value is calculated. The identified community will perform better if there are more nodes inside the community and if the modularity score of the community is around 1. When comparing the frequency of properly recognizing the nodes in each community using the supplied ground truth data, the F1_Score is a well-known evaluation statistic used in community detection methods. The other outsider statistic is NMI, or mutual information, about the connection found among the recognized groups and the real world.

Real-World Datasets

Five real-world datasets are used in the experiments. Zachary's karate club network [51] is the first dataset, containing 34 nodes, 78 connecting edges between them, and 2 ground-truth communities. This dataset contains social ties among university karate club members collected by Wayne Zachary in 1977. Dolphins' online social network [52] is the second dataset, which contains 62 nodes, 159 connecting edges, and two ground-truth communities containing a list of all the links, where a link represents frequent associations between dolphins. The third dataset [53], with 105 nodes, 441 connecting edges, and 3 ground-truth communities, is based on data from the network of books about US politics published around the time of the 2004 presidential election. Edges between books represent frequent co-purchasing of books by the same buyers. The fourth dataset is the American football [4], with 116 nodes, 613 connecting edges, and 12 ground-truth communities. This network contains American football games between Division IA colleges during the fall of 2000. The fifth dataset is a large network generated using email data from a large European research institution [54], [55]. This network contains 1005 members of the institution as nodes, and 25571 edges contain emails sent between members of the institution and people outside of the institution. The dataset also assumes departments at the research institute as the nodes' ground-truth community memberships. Each individual belongs to exactly one of the 42 departments at the research institute.

The real-world datasets analyzed during the current study are shown in Table 2, where N is the number of nodes, E is the number of edges, and K is the number of ground truths. These datasets are available in the network repository[1] [56], the KONECT project[2] [53], and the Stanford Network Analysis Project[3] [54] (SNAP).

Table 2: The specifics of the real-world dataset used

Dataset Name	N (#Nodes)	E (#Edges)	K (#Ground_truth)
Karate	34	78	2
Dolphin	62	159	2
Pol-Book	105	441	3
Football	115	613	12
Email-EU	1005	25571	42

[1] https://networkrepository.com/
[2] http://konect.cc/
[3] https://snap.stanford.edu/

Experimental on Real Datasets

We evaluate the PMCD by four kinds of community detection models, such as modularity optimization, label propagation, probabilistic estimation, and clique percolation, in order to assessment the efficacy and accuracy of PMCD in community detection. In the sections before, several of these methods were briefly discussed. The suggested approach is assessed using six algorithms using internal evaluation criteria (modularity and community number) as well as external evaluation metrics (NMI and F1_Score).

The findings in Table 3 demonstrate that our method has more accuracy than other methods in terms of the internal metrics (modularity maximum and accuracy in the number of communities).

Additionally, Fig. 3 and 4 demonstrate that our suggested method has absolute superiority over probabilistic estimation and clique percolation methods and relative superiority over modularity optimization and label propagation methods in the external assessment criteria (NMI and F1_Score).

Table 3: Experimental results on real-world networks by the modularity metric (Q) and community number (CN)

Methods	Louvain		Leiden		Bigclam		CPM		LPA		SLPA		PMCD	
Datasets	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN
Karate	0.415	4	0.116	5	0.204	4	0.215	4	0.354	3	0.371	3	0.397	3
Dolphin	0.518	5	0.134	7	0.185	6	0.321	5	0.456	4	0.470	3	0.522	3
Pol-Book	0.526	4	0.279	9	0.347	8	0.271	9	0.481	5	0.493	5	0.543	4
Football	0.604	10	0.257	19	0.381	16	0.283	18	0.552	14	0.596	13	0.632	11
Email-EU	0.432	27	0.226	58	0.207	65	0.162	74	0.274	55	0.303	52	0.507	46

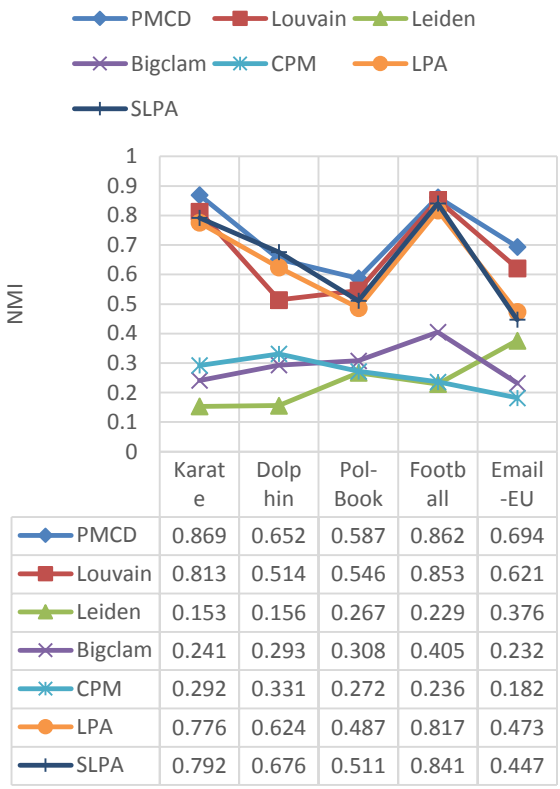

Fig. 3: NMI assessment chart, compare PMCD by community detection models on five real-world data sets.

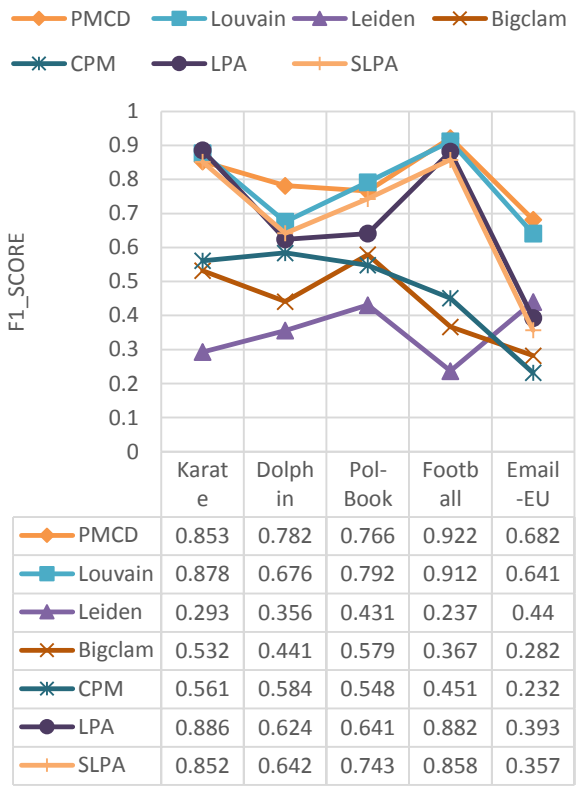

Fig. 4: F1_Score assessment chart, compare PMCD by community detection models on five real-world data sets.

Artificial Datasets

Utilizing an artificial network for evaluating community detection methods makes sense. Different methods can be used to generate artificial networks.

One of the most famous and often used strategies is the LFR benchmark [57]. The LFR benchmark builds types of artificial graphs with ground truth communities using density and dimension of communities. The network and community parameters can be set up before using LFR to simulate networks. The mixing parameter (μ) is one of the essential LFR parameters. This variable controls how various communities interact. A high mixing parameter value (μ), as indicated in Table 5, will reduce the network's level of modularity (Q_{GT}). As a result, the LFR-generated datasets are divided into two categories based on the modularity measure and mixing parameter: sparse communities and dense communities. The average degree is another crucial element that might be raised to encourage more intercommunal interaction.

Table 4 displays the key characteristics of the LFR artificial networks. Table 5 details the dataset that was created using our LFR approach.

Table 4: Parameters of LFR synthetic datasets [57]

Parameter	Description
N	Node number
K	Average degree
Min K	Minimum nodes degree
Max K	Maximum nodes degree
μ	Mixing parameter for the structure
Min C	Minimum for the community sizes
Max C	Maximum for the community sizes
tau1(γ)	The degree distribution
tau2(β)	The community size distribution

Table 5: The LFR artificial network properties

Graph Name	*N*	*k*	γ	β	μ	Q_{GT}
LFR-1	1000	20	3	1.5	0.05	0.895
LFR-2	1000	20	3	1.5	0.10	0.844
LFR-3	1000	20	3	1.5	0.15	0.800
LFR-4	1000	20	3	1.5	0.20	0.739
LFR-5	1000	20	3	1.5	0.25	0.699
LFR-6	1000	20	3	1.5	0.30	0.647
LFR-7	1000	20	3	1.5	0.35	0.603
LFR-8	1000	20	3	1.5	0.40	0.560
LFR-9	1000	20	3	1.5	0.45	0.504
LFR-10	1000	20	3	1.5	0.50	0.460
LFR-11	1000	20	3	1.5	0.55	0.407
LFR-12	1000	20	3	1.5	0.60	0.364
LFR-13	1000	20	3	1.5	0.65	0.321
LFR-14	1000	20	3	1.5	0.70	0.275
LFR-15	1000	20	3	1.5	0.75	0.229
LFR-16	1000	20	3	1.5	0.80	0.182

Table 6: Experimental results on sixteen LFR artificial networks by the modularity metric

Mixing Parameter (μ)	Louvain	Leiden	Bigclam	CPM	LPA	SLPA	PMCD
0.05	1.00	0.64	0.89	0.84	0.99	1.00	1.00
0.10	0.99	0.51	0.86	0.81	0.97	0.98	0.97
0.15	0.96	0.42	0.77	0.72	0.93	0.95	0.98
0.20	0.93	0.41	0.73	0.66	0.88	0.87	0.91
0.25	0.89	0.37	0.69	0.57	0.83	0.85	0.91
0.30	0.86	0.23	0.59	0.53	0.76	0.79	0.84
0.35	0.82	0.22	0.57	0.43	0.69	0.72	0.81
0.40	0.79	0.19	0.47	0.31	0.52	0.64	0.72
0.45	0.70	0.14	0.31	0.29	0.43	0.56	0.73
0.50	0.53	0.12	0.25	0.24	0.41	0.48	0.66
0.55	0.50	0.09	0.19	0.22	0.36	0.33	0.52
0.60	0.44	0.05	0.12	0.14	0.28	0.29	0.39
0.65	0.37	0.04	0.08	0.07	0.24	0.18	0.38
0.70	0.29	0.01	0.05	0.03	0.15	0.14	0.30
0.75	0.21	0.00	0.03	0.01	0.11	0.09	0.24
0.80	0.15	0.00	0.01	0.00	0.08	0.05	0.18

Experimental on Artificial Networks

In addition to actual graphs, we have also examined LFR artificial networks. We contrast the PMCD by the famous community detection model in Table 1 to demonstrate the result of the optimized loss function for a probabilistic estimate on the community detection utilizing modularity, F1_Score, and NMI measure. For this, sixteen LFR artificial networks are developed with various configures of mixing parameters (μ) ranging from 0.05 to 0.8, as indicated in Table 5. These networks are created in accordance with the attributes of synthetic networks listed in Table 4. Table 6's experimental findings demonstrate that the communities are dense for low mixing parameter range (e.g., 0.05≤μ≤0.4) and that the compared methods are almost correct in this situation.

However, the major contrast between the methods becomes more apparent when the mixing parameter's (μ) value rises (e.g., 0.4<μ≤0.8) and the communities are sparse, making it difficult to identify communities since the edges between communities' rise.

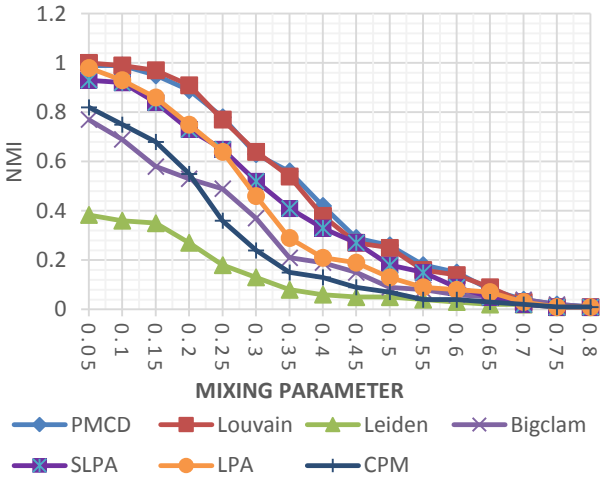

Fig. 5: NMI assessment graph on sixteen LFR datasets, comparing PMCD with six community detection methods.

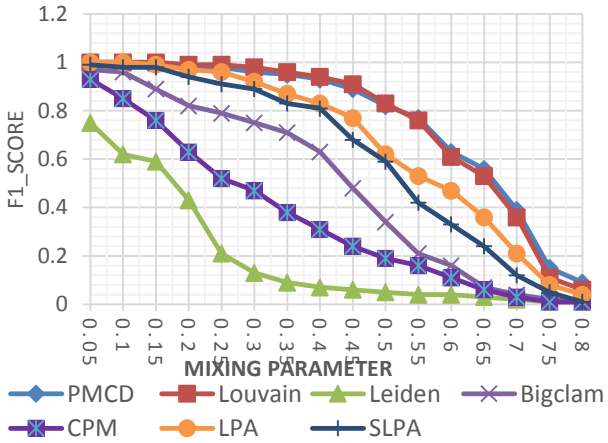

Fig. 6: On sixteen LFR datasets, the F1_score assessment chart compares PMCD with six community detection methods.

As can be seen in Figs 5 and 6, when the mixing parameter value increases, certain algorithms have NMI and F1Score values that are equivalent to zero. The majority of frequently used approaches in the range of 0.5 to 0.8 are outperformed by the suggested method.

Conclusion

We proposed a probabilistic motif-based method for detecting communities in complex networks. Due to the complexity of combining probabilistic approaches in motif structure, recent community detection methods have given the latent variable of the probabilistic model less consideration. However, the proposed approach leverages the intensity of the node's participation in the community and the relationship of at least two linked edges between three nodes (triangular motif structure) to estimate the hidden variable of the probabilistic model. The research maximized the likelihood function and extracted the model's latent parameters using the well-known block coordinate ascent technique. The association between node membership in communities and edge density is another aspect that helps in the examination of newly detected communities; three nodes are more likely to create a motif structure when seen in various communities. Overlapping in the identification of communities is another benefit of PMCD; according to the findings, communities that overlap have a greater density of triangular motifs. We employed 16 artificial graphs and 5 real graphs to evaluate the performance of the suggested method. In comparison to the other six methods, PMCD was able to achieve a sufficient quorum on real-world networks and surpass them in terms of internal and external assessment criteria. Synthetic network assessments further show that the suggested strategy performs better in sparse datasets than other approaches. Furthermore, a review of the complexity of the execution time reveals that the suggested method outperforms previous approaches. Future research can develop PMCD. A probabilistic generative model can be used to estimate edge weight while taking a latent parameter into account. Also, the suggested method can be enhanced by utilizing network node properties in order to give a more precise description of the found communities.

Author Contributions

H. Hajibabaei, V. Seydi, and A. Koochari contributed to the research design and implementation, the analysis of the results, and the writing of the manuscript.

Acknowledgment

We thank the editor and all the anonymous reviewers.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy, have been completely observed by the authors.

Abbreviations

PMCD	probabilistic motif-based community detection
NMI	Normalized Mutual Information
LFR	Lancichinetti–Fortunato–Radicchi Benchmark
CPM	Clique Percolation Method
LPA	Label Propagation Algorithm
SLPA	Speaker-Listener Label Propagation Algorithm
Q_{GT}	Ground Truth Modularity
SNAP	Stanford Network Analysis Project

References

[1] J. Sia, E. Jonckheere, P. Bogdan, "Ollivier-ricci curvature-based method to community detection in complex networks," Sci. Rep., 9(1): 1-12, 2019.

[2] Y. Y. Ahn, J. P. Bagrow, S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, 466(7307): 761-764, 2010.

[3] J. J. McAuley, J. Leskovec, "Learning to discover social circles in ego networks," in Proc. NIPS: 548-556, 2012.

[4] M. Girvan, M. E. Newman, "Community structure in social and biological networks," PNAS, 99(12): 7821-7826, 2002.

[5] W. Wu, S. Kwong, Y. Zhou, Y. Jia, W. Gao, "Nonnegative matrix factorization with mixed hypergraph regularization for community detection," Inf. Sci., 435: 263-281, 2018.

[6] S. Fortunato, D. Hric, "Community detection in networks: A user guide," Phys. Rep., 659: 1-44, 2016.

[7] H. Hajibabaei, V. Seydi, A. Koochari, "Community detection in weighted networks using probabilistic generative model," J. Intell. Inf. Syst., 60: 119- 136, 2023.

[8] T. S. Wang, H. T. Lin, P. Wang, "Weighted-spectral clustering algorithm for detecting community structures in complex networks," Artif. Intell. Rev., 47(4): 463-483, 2017.

[9] X. Chen, J. Li, "Community detection in complex networks using edge-deleting with restrictions," Physica A, 519: 181-194, 2019.

[10] F. D. Zarandi, M. K. Rafsanjani, "Community detection in complex networks using structural similarity," Physica A, 503: 882-891, 2018.

[11] B. D. Le, H. Shen, H. Nguyen, N. Falkner, "Improved network community detection using meta-heuristic based label propagation," Appl. Intell., 49(4): 1451-1466, 2019.

[12] C. Lyu, Y. Shi, L. Sun, "A novel local community detection method using evolutionary computation," IEEE Trans. Cybern., 51(6): 3348-3360, 2019.

[13] W. Zhou, X. Wang, C. Zhang, R. Li, C. Wang, "Community detection by enhancing community structure in bipartite networks," Mod. Phys. Lett. B, 33(7): 1950076, 2019.

[14] J. Yang, J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in Proc. the Sixth ACM International Conference on Web Search and Data Mining: 587-596, 2013.

[15] T. Ma *et al.*, "LED: A fast overlapping communities detection algorithm based on structural clustering," Neurocomputing, 207: 488-500, 2016.

[16] F. Liu, D. Choi, L. Xie, K. Roeder, "Global spectral clustering in dynamic networks," PNAS, 115(5): 927-932, 2018.

[17] J. Yang, J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in Proc. 2012 IEEE 12th International Conference on Data Mining: 1170-1175, 2012.

[18] G. Palla, I. Derényi, I. Farkas, T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," nature, 435(7043): 814-818, 2005.

[19] X. Zhou, K. Yang, Y. Xie, C. Yang, T. Huang, "A novel modularity-based discrete state transition algorithm for community detection in networks," Neurocomputing, 334: 89-99, 2019.

[20] M. E. Newman, M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, 69(2): 026113, 2004.

[21] K. Berahmand, A. Bouyer, "A link-based similarity for improving community detection based on label propagation algorithm," J. Syst. Sci. Complexity, 32(3): 737-758, 2019.

[22] U. N. Raghavan, R. Albert, S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Phys. Rev. E, 76(3): 036106, 2007.

[23] M. Zarezade, E. Nourani, A. Bouyer, "Community detection using a new node scoring and synchronous label updating of boundary nodes in social networks," J. AI Data Min., 8(2): 201-212, 2020.

[24] S. Kumar, B. Panda, D. Aggarwal, "Community detection in complex networks using network embedding and gravitational search algorithm," J. Intell. Inf. Syst., 57: 51-72, 2021.

[25] A. Torkaman, K. Badie, A. Salajegheh, M. H. Bokaei, S. F. Fatemi, "A hybrid deep network representation model for detecting researchers' communities," J. AI Data Min., 10(2): 233-243, 2022.

[26] X. Su *et al.*, "A comprehensive survey on community detection with deep learning," IEEE Trans. Neural Networks Learn. Syst., 2022.

[27] M. Ali, M. Hassan, K. Kifayat, J. Y. Kim, S. Hakak, M. K. Khan, "Social media content classification and community detection using deep learning and graph analytics," Technol. Forecasting Social Change, 188: 122252, 2023.

[28] C. Li, Y. Tang, Z. Tang, J. Cao, Y. Zhang, "Motif-based embedding label propagation algorithm for community detection," Int. J. Intell. Syst., 37(3): 1880-1902, 2022.

[29] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, A. Baig, "Community detection in networks: A multidisciplinary review," J. Network Comput. Appl., 108: 87-111, 2018.

[30] K. Guo, X. Huang, L. Wu, Y. Chen, "Local community detection algorithm based on local modularity density," Appl. Intell., 52(2): 1238-1253, 2022.

[31] J. Sánchez-Oro, A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks," Future Gener. Comput. Syst., 88: 785-791, 2018.

[32] V. A. Traag, L. Waltman, N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Sci. Rep., 9(1): 1-12, 2019.

[33] C. Li, H. Chen, T. Li, X. Yang, "A stable community detection approach for complex network based on density peak clustering and label propagation," Appl. Intell., 52(2): 1188-1208, 2022.

[34] J. Xie, B. K. Szymanski, X. Liu, "Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in Proc. 2011 ieee 11th International Conference on Data Mining Workshops: 344-349, 2011.

[35] S. Gregory, "Finding overlapping communities in networks by label propagation," New J. Phys., 12(10): 103018, 2010.

[36] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," Bioinformatics, 22(8): 1021-1023, 2006.

[37] P. Bloem, S. de Rooij, "Large-scale network motif analysis using compression," Data Min. Knowl. Discovery, 34: 1421-1453, 2020.

[38] A. Arenas, A. Fernandez, S. Fortunato, S. Gomez, "Motif-based communities in complex networks," J. Phys. A: Math. Theor., 41(22): 224001, 2008.

[39] C. E. Tsourakakis, J. Pachocki, M. Mitzenmacher, "Scalable motif-aware graph clustering," in Proc. the 26th International Conference on World Wide Web: 1451-1460, 2017.

[40] L. Huang, H. Y. Chao, Q. Xie, "MuMod: A micro-unit connection approach for hybrid-order community detection," in Proc. the AAAI conference on artificial intelligence, 34(01): 107-114, 2020.

[41] J. Yang, J. McAuley, J. Leskovec, "Community detection in networks with node attributes," in Proc. IEEE 13th International Conference on Data Mining: 1151-1156, 2013.

[42] K. Yang, Q. Guo, J. G. Liu, "Community detection via measuring the strength between nodes for dynamic networks," Physica A, 509: 256-264, 2018.

[43] W. Yu, W. Wang, P. Jiao, X. Li, "Evolutionary clustering via graph regularized nonnegative matrix factorization for exploring temporal networks," Knowledge-Based Syst., 167: 1-10, 2019.

[44] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, "Network motifs: simple building blocks of complex networks," Science, 298(5594): 824-827, 2002.

[45] Y. Xu, W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," SIAM J. Imag. Sci., 6(3): 1758-1789, 2013.

[46] C. J. Hsieh, I. S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," in Proc. the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 1064-1072, 2011.

[47] C. J. Lin, "Projected gradient methods for nonnegative matrix factorization," Neural Comput., 19(10): 2756-2779, 2007.

[48] D. F. Gleich, C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in Proc, the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 597-605, 2012.

[49] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech: Theory Exp., 2008(10): P10008, 2008.

[50] A. Clauset, M. E. Newman, C. Moore, "Finding community structure in very large networks," Phys. Review E, 70(6): 066111, 2004.

[51] W. W. Zachary, "An information flow model for conflict and fission in small groups," J. Anthropol. Res., 33(4): 452-473, 1977.

[52] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behav. Ecol. Sociobiol., 54(4): 396-405, 2003.

[53] J. Kunegis, "Konect: the koblenz network collection," in Proc. the 22nd International Conference on World Wide Web: 1343-1350, 2013.

[54] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Trans. Knowl. Discovery Data (TKDD), 1(1): 2-es, 2007.

[55] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 555-564.

[56] R. Rossi, N. Ahmed, "The network data repository with interactive graph analytics and visualization," in Proc. Twenty-ninth AAAI Conference on Artificial Intelligence, 2015.

[57] A. Lancichinetti, S. Fortunato, F. Radicchi, "Benchmark graphs for testing community detection algorithms," Phys. Review E, 78(4): 046110, 2008.

Biographies

Hossein Hajibabaei received his M.Sc. in Software Engineering in 2014. Currently, he is a Ph.D. candidate in Computer Engineering majoring in artificial intelligence, and is teaching as a teaching assistant at the Islamic Azad University of Science and Research Branch. His areas of interest are social network analysis and deep learning.

- Email: h.hajibabaei@srbiau.ac.ir
- ORCID: 0009-0005-8063-981X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Vahid Seydi is a Senior Research Fellow in the School of Ocean Science at Bangor University in Data Science (DS) and Machine Learning (ML). Before Bangor, He was an Assistant Professor at the Department of Al at Azad University South Tehran Branch (Feb 2014 - Sep 2020) and an award-winning lecturer (Oct 2010 Feb 2014). Vahid received a B.Sc. (2005) in software engineering, and M.Sc. (2007), and Ph.D. (2014) in Al, from the Department of Computer Science at Azad University, Science and Research Branch, Tehran Iran. He has been awarded a Global Talen endorsement from the UK Royal Society (2023), His current research fellowship (2020), and a merit-based scholarship for attending the school of Al, Rome, Italy (2019), Also, He has achieved a full scholarship Award from Azad University (2010-2014), KNTU ISLAB Research Fellowship (2007-2010). He secured the first rank among the graduates from 2004-2005. His current research focuses on dedicating machine learning methods to analyze data associated with digital oceanography, especially in the offshore renewable energy section.

- Email: vahidseydi@gmail.com
- ORCID: 0000-0001-5702-2209
- Web of Science Researcher ID: NA
- Scopus Author ID: 23490316700
- Homepage: https://vahidseydi.github.io

Abbas Koochari received his Ph.D. in Computer Engineering majoring in artificial intelligence. He is currently an assistant professor and a member of the scientific staff of Islamic Azad University, Science and Research Branch. His areas of interest are image processing, machine vision, speech and natural language processing, and deep learning.

- Email: koochari@srbiau.ac.ir
- ORCID: 0000-0003-0584-6470
- Web of Science Researcher ID: NA
- Scopus Author ID: 36005396600
- Homepage: https://srb.iau.ir/faculty/a-koochari/en

How to cite this paper:
H. Hajibabaei, V. Seydi, A. Koochari, "Motif-based community detection: A probabilistic model based on repeating patterns," J. Electr. Comput. Eng. Innovations, 12(1): 247-258, 2024.

DOI: 10.22061/jecei.2023.9931.663

URL: https://jecei.sru.ac.ir/article_2013.html