IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2019, accepted August 2, 2019, date of publication August 20, 2019, date of current version September 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936248

Community Detection and Visualization in Complex Network by the Density-Canopy-Kmeans Algorithm and MDS Embedding

MANZHI LI[1,2], HONGTAO WANG[1], HAIXIA LONG[3], JU XIANG[4,5], BO WANG[6], JUNLIN XU[7], AND JIALIANG YANG[1,6]

[1]School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China
[2]Hainan Key Laboratory for Computational Science and Application, Hainan Normal University, Haikou 571158, China
[3]School of Information Science Technology, Hainan Normal University, Haikou 571158, China
[4]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[5]Neuroscience Research Center & Department of Basic Medical Sciences, Changsha Medical University, Changsha 410219, China
[6]Geneis Beijing Company Ltd., Beijing 100102, China
[7]College of Information Science and Engineering, Hunan University, Changsha 410082, China

Corresponding authors: Ju Xiang (xiang.ju@foxmail.com) and Jialiang Yang (yangjl@geneis.cn)

This work was supported in part by the Hainan Province Natural Science Foundation under Grant 118QN231 and Grant 618MS057, in part by the National Natural Science Foundation of China under Grant 61903106, Grant 61762034, and Grant 61702054, in part by the Project of Hainan Key Laboratory for Computational Science and Application, in part by the Training Program for Excellent Young Innovators of Changsha under Grant kq1802024, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2018JJ3568, and in part by the Scientific Research Fund of Education Department of Hunan Province under Grant 17A024.

ABSTRACT With the increasing availability of social networks and biological networks, detecting network community structure has become more and more important. However, most traditional methods for detecting community structure have limitations in dimension reduction or parameter optimization. In this paper, we propose a Density-Canopy-Kmeans clustering algorithm (DCK) to detect network community structure. Specifically, we define a novel distance metric, which integrates random distance and community structure coefficient based on the Jaccard distance. After applying the Multidimensional Scaling (MDS) dimension reduction, we cluster the nodes. KMEANS is combined with density clustering and canopy clustering to determine the optimal number of communities and the best initial seeds are determined to improve the accuracy and stability of the K-means algorithm. Compared with traditional community detection methods, our method has a higher classification accuracy and a better visualization effect. Thus, this method is effective for analyzing network communities.

INDEX TERMS Network, community detection, the Density-Canopy-Kmeans clustering algorithm, MDS.

I. INTRODUCTION

Complex networks, such as protein interaction networks, gene regulatory networks, social networks and cooperative networks, have received more and more attention recently, especially from interdisciplinary domains [1], [2]. It has been found that these complex networks have many common topological properties, among which community structure or modular structure is one of the most important and widely studied properties [3]. A community, module or cluster in a network is a set of nodes such that the nodes within the community are closely connected, while those in different communities are sparsely connected. The communities often correspond to functional units in real-world networks. Therefore, the analysis of community structure has a wide range of applications in the field of biology, physics, computer graphics and sociology [4], [5].

In order to effectively detect communities and visualize network structures, we design a new distance metric formula and constructs a Density-Canopy-Kmeans algorithm (DCK), which can infer the optimal number of communities and the best initial seeds for K-means. In addition, based on the distance metric, we adopt MDS to project a complex network into a two-dimensional space. The projection graph displays a more intuitive and clearer community structure while keeping important connections in the original network. Taking only

The associate editor coordinating the review of this article and approving it for publication was Weiguo Xia.

the adjacency matrix of a network as input, the algorithm can achieve a good community partition for visualization.

II. RELATED WORK

Community detection algorithms in complex networks can generally be divided into three categories including clustering-based algorithms, optimization-based algorithms and network dynamics-based algorithms. Hierarchical clustering-based algorithms detect communities according to the hierarchical clustering of nodes, e.g., agglomerative hierarchical clustering based on maximal clique (EAGLE) [6], cluster-overlap Newman Girvan (CONGA) [7] and the method based on the local optimization of a fitness function (MSCD-LFK) [8]. With the development of community detection algorithms, many evaluation parameters for community structure such as modularity and conductivity have been proposed [9]. As a result, there are many optimization-based algorithms proposed to optimize the objective functions based on these parameters, such as multi-objective genetic algorithm for community detection in networks (MOGA-NET) [10], heuristic artificial bee colony (HABC) [11], Order Statistics Local Optimization Method (OSLOM) [12], LouvainSprs [13], LouvainSgnf [14] and multi-objective discrete cuckoo search algorithm with local search (MDCL) [15]. Because of the intrinsic correlation between network structure and dynamical behaviors in the networks, many network dynamics-based algorithms utilize the dynamical characteristics of complex networks for community detection, e.g., random walks and diffusion on networks [16], maps of random walks on complex networks (Infomap) [17], the method using random walks (Walktrap) [18], the Label Propagation Algorithm (LPcopra) [19], and multiresolution community detection in large-scale networks (MSCD_HSLSW) [20]. The readers are referred to several reviews for a comprehensive summary on network community detection algorithms [2], [21].

The essence of complex network community detection is to cluster nodes in a network. Therefore, it is critical to define the similarity between nodes, based on which a clustering algorithm can be applied to divide the network into communities. For example, Cai et al. used the internal positive similarity of community as the node feature to calculate the distance between two nodes, and applied the clustering algorithm to cluster the network nodes [22]. However, the clustering results are very dependent on the distance metric and the parameters of the algorithm. The visualization of the networks is generally based on multidimensional scaling (MDS) [23], PCA [24], Laplacian Eigenmaps(LE) [25], T-SNE [26] and other dimensionality reduction methods. These methods usually map the nodes in a network to a low-dimensional Euclidean space, assign a reasonable k-dimensional coordinate (generally k=2 or 3) to each node, and draw a graph according to the coordinates for observing community structure. However, the visualization of community structure will be highly affected by the dimension reduction methods and distance metrics.

III. METHODS

Assume that the network to be analyzed is an undirected and unweighted static network with n nodes. Let A be its adjacency matrix with $A_{ij} = 1$ if there is an edge between two node N_i and N_j, and $A_{ij} = 0$ otherwise. The algorithm framework consists of three parts. Firstly, we calculate the node similarity matrix D based on A. Secondly, the clustering algorithm is applied to D to generate network communities. Finally, the distance matrix D is projected by MDS for visualization.

A. COMPUTATION OF SIMILARITY DISTANCE

At present, there are various methods to calculate similarity between two nodes in a network. Ten similarity measures were compared in Lü *et al.* [27], which showed that the Jaccard similarity is suitable for measuring topological closeness. The *Jaccard* similarity of node N_i and N_j is defined as $s_{ij} = |N(i) \cap N(j)|/|N(i) \cup N(j)|$, where $N(i)$ is the neighbor set of node i and $|N(i)|$ represents the number of elements in the set $N(i)$ [28].

Each row of the adjacency matrix A indicates the neighbors of each node, where the elements are 1 for its neighbors and 0 for non-neighbors. Assume that L_{11} represents the number of corresponding bits the values of which are 1 in both row i and j; L_{10} denotes the total number of bits that are 1 in row i and 0 in row j; L_{01} denotes the total number of bits that are 0 in row i and 1 in row j; L_{00} is the total number of corresponding bits with a value of 0 in both rows, and thus the total number of nodes $L_{11}+L_{01}+L_{10}+L_{00} = n$. Then the *Jaccard* similarity is $s_{ij} = L_{11}/(L_{01}+L_{10}+L_{11})$, and the corresponding *Jaccard* distance is

$$dist_{ij} = (L_{01} + L_{10})/(L_{01} + L_{10} + L_{11}) \tag{1}$$

The *Jaccard* similarity will be 0 if node i and j does not share any neighbor.

For complex networks, these zero-similar nodes can't apply MDS methods to get their two-dimensional coordinates. For this problem, we generate a random symmetric matrix with the same dimension of the adjacency matrix as the random distance $dist\_random_{ij}$ of each node, where the random number range is [0, 1]. In this way, a new distance matrix D is formed, and the formula for calculating its elements d_{ij} is shown in formula (2)

$$d_{ij} = \gamma \cdot dist_{ij} + dist\_random_{ij}, \tag{2}$$

where γ is the community structure coefficient, and its value affects the community structure discovery during clustering and visualization.

B. THE MDS EMBEDDING

Multidimensional Scaling (MDS) is a multivariate data analysis technique that displays 'distance' data structures in low-dimensional space. As a typical representative of information visualization technology, MDS can be used to reveal the relationship between abstract objects, show the spatial clustering

of data, and help people explore and discover information in an effective and intuitive visual environment. MDS can keep distance information when more elements are mapped to Euclidean space [29]. MDS is a reasonable tool for mapping network nodes to Euclidean space. The process of MDS is as follows:

Firstly, calculate the distance matrix D from equation (2) whose elements d_{ij} are the distance between the nodes N_i and N_j.

Secondly, calculate matrix B, where,

$$b_{ij} = -\frac{1}{2} * (d_{ij}^2 - (\sum_{j=1}^{n} d_{ij}^2 - \sum_{i=1}^{n} d_{ij}^2)/n + (\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2)/n^2)$$

After eigenvalue decomposition of the matrix B, eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$ are sorted in descending order, the two largest eigenvalues λ_1, λ_2 are taken to form a diagonal matrix Λ, and the corresponding eigenvector matrix Ψ is extracted.

Finally, calculate the bivariate coordinate matrix of nodes $X = \Psi\sqrt{\Lambda}$.

C. THE DENSITY-CANOPY-KMEANS ALGORITHM

After the distance matrix of the node is calculated by (2), the nodes can be clustered after applying MDS for dimensionality reduction. K-means is a classic partitioned clustering algorithm [30], but the traditional K-means needs to specify the number K of communities in advance. In order to improve the accuracy and stability of the Kmeans algorithm, and solve the problem of determining the optimal number K of clustering and the optimal initial seeds, we construct the Density-Canopy-Kmeans clustering algorithm. Canopy belongs to the coarse clustering method and is often used in conjunction with Kmeans [31], [32].

The traditional canopy algorithm is: set a Canopy initial center point and a region radius T1 for a certain data set, and efficiently divide the data set into several overlapping subsets (i.e., Canopy), so that all objects fall within the range of Canopy coverage. For the objects falling in the same area, recalculate the new center point and re-divide the area to which the object belongs according to the distance between the object and the new center point; cyclically execute the process of 'dividing the Canopy calculation center point' until the positions of the k center points no longer change.

The canopy algorithm needs to determine two parameters, how to determine them is still unclear. In order to improve the stability and computational efficiency of the canopy, the principle of density clustering is applied further. In the process of Density-Canopy-Kmeans algorithm, the values of R and T directly affect the overlap rate and granularity of canopy: when R is too large, the sample will belong to multiple canopy, and the difference between canopies is not obvious; when T is too large, it will reduce the number of canopy, and when T is too small, it will increase the number of canopy and the calculation time.

The values of R and T are related to the distance matrix. The calculation formula is as follows:

$$R = d_{\min} + r \cdot (d_{\max} - d_{\min}) \quad (3)$$

$$T = d_{\min} + t \cdot (d_{\max} - d_{\min}) \quad (4)$$

where $d_{\min}$ is the average value of the minimum distance between each node and other nodes, $d_{\max}$ is the average value of the maximum distance between each node and other nodes, r is the density radius coefficient, and $t$is the distance threshold coefficient. We tested r&t and found that it is insensitive to the results, and finally we set $r = t = 1/6$.

The complete algorithm is presented in Alg. 1 and time complexity is $O(n^3)$ where n is the number of nodes.

Algorithm 1 DCK

Input: The adjacency matrix A *of* the network $G = (V, E)$, the parameter $r\&t$

Output: the set of clusters

1: Calculate the Jaccared distance matrix J according to Eq 1.
2: Calculate the distance matrix D according to Eq 2.
3: Apply MDS to reduce the D *is* matrix to 2-dimensions and calculate the Euclidean distance matrix D.
4: **for** each node V_i in network **do**
5: Count the number of nodes with $d < R$ and set list sorts from large to small.
6: **for** each node V_i in list **do**
7: Select the first node V_1 in order in the List, the nodes of $d < T$ belong to the V_1 canopy, and remove these nodes from the List.
8: Repeat step 7 until the List is empty
9: **end for**
10: **end for**
11: Obtain canopy set. If there are fewer than five nodes in the canopy, these nodes are considered to be isolated points, and these canopy are invalid.
12: The number of valid canopy sets is the number of clusters k. Apply Kmeans clustering for nodes in valid canopy set to get the clustering labels of these nodes.
13: Isolated nodes are defined cluster labels based on their shortest distance with the clusters center.

IV. RESULTS

In this section, we will evaluate the performance of DCK on artificial and realistic networks. In order to test the effectiveness of the proposed algorithm, we mainly use modularity to measure the quality of community detection. The network modularity evaluation function Q is used to quantitatively describe the quality of the network community structures [33], which is defined as the difference between the actual number of connections in the communities and the expected value in the null model (i.e. random networks).

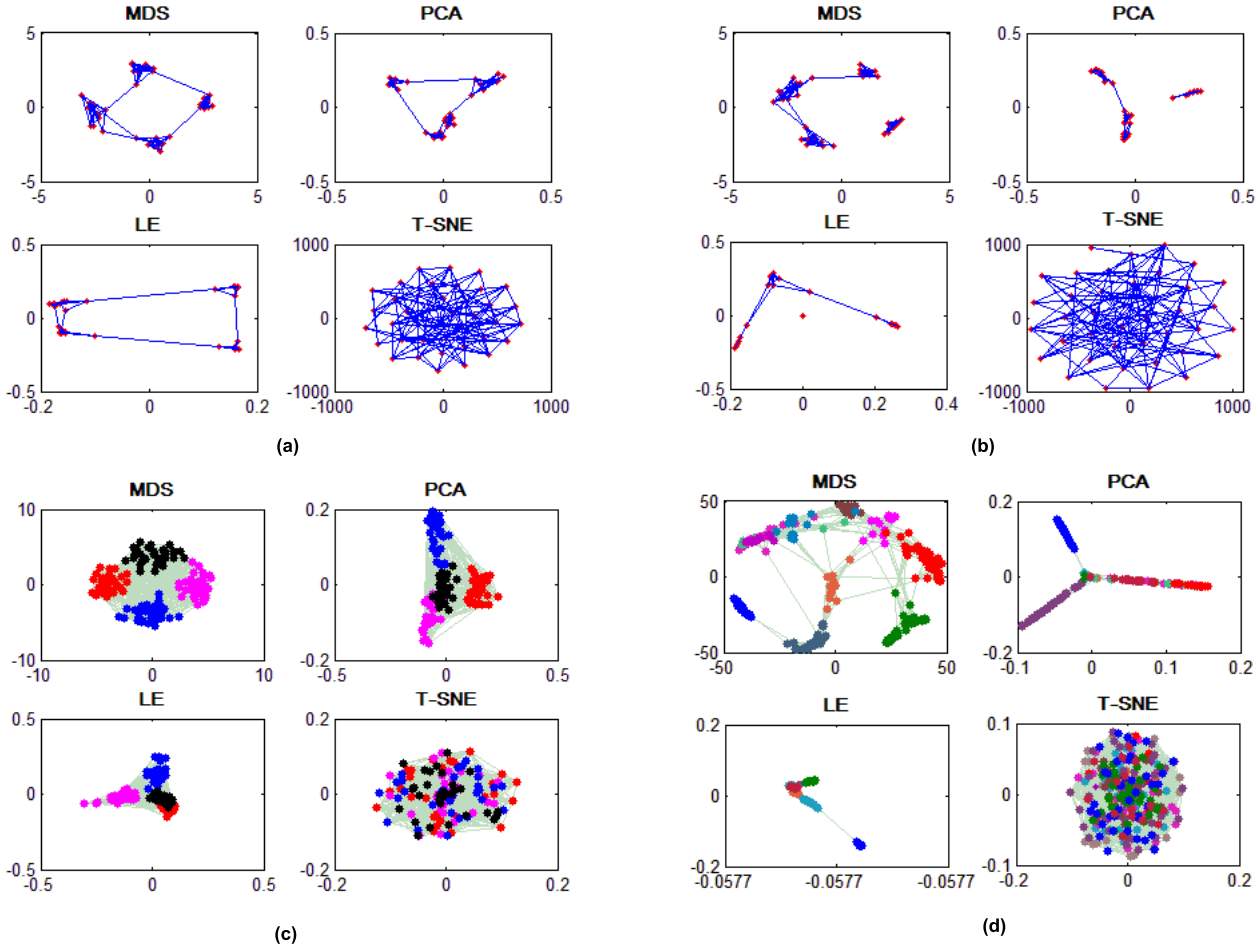


FIGURE 1. Comparison of different embedding methods in three types of networks: (a) Connected graph, (b) Mixed graph, (c) GN4 ($P_{in} = 0.69$, $k_{out} = 4$); (d) Pollen single cell.

The calculation formula is as follows:

$$Q = \sum_{u=1}^{K} [(h_u/H) - (d_u/(2H))^2] \tag{5}$$

where K is the number of network communities, H is the total number of network connections, h_u is the total number of connections in the community u, and d_u is the sum of the degree of nodes in the community. There are three parameters γ, r, t in the model, and the optimal parameter is determined by the maximum Q value. In general, the greater the degree of modularity, the higher the quality of community partitioning in the network, but sometimes it does not fully conform to the community structure. Therefore, in addition to the modularity, we also use several widely used measurement indicators [21] with F-measure, Rand index (RI), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) to evaluate networks of known structures.

The algorithms that are used for comparison include MSCD-LFK [8], OSLOM [12], Lourvarinsprs [13], LouvainSgn [14], LPcopra [19], and Kmeans [30]. For the algorithm with randomness, we run 10 times and take the results corresponding to the optimal value of the module.

A. ARTIFICIAL NETWORKS

The artificial networks with known community structure are used to test the effectiveness of the community detection algorithm. In order to compare the visualization effects of the four projection methods and the test algorithm parameters and performance, we tested them in the synthetic network, GN network [33] and LFR network [34].

1) COMPARISON OF DIFFERENT PROJECTION METHODS

We use MDS to project several different networks onto a two-dimensional plane and compare it to the results of the other three projection methods: PCA, LE and t-SNE. A path between any two nodes in the network can reach each other as a connected graph such as Fig. 1(a). However, most of the networks have the characteristics of Fig. 1(b), including some connected communities and disconnected communities, called mixed graphs. These two networks are computer-synthesized. For more comprehensive testing, we tested it

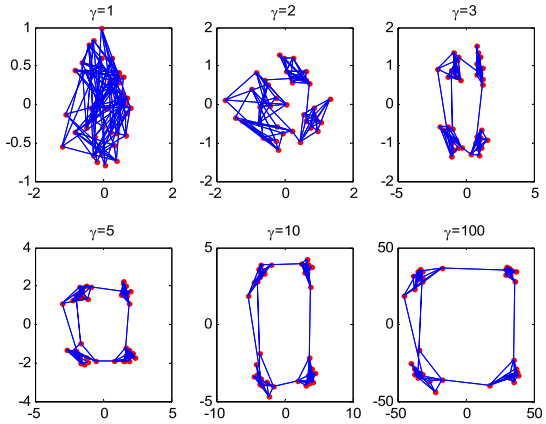


FIGURE 2. Comparison of community structure coefficient.

TABLE 1. LFR network parameter.

N	the number of nodes in the networks	----
k_{m}	the average degree	20
$k_{\max}$	the maximum degree	50
$c_{\min}/c_{\max}$	the minimum/ maximum community sizes	10/50
e_1	exponent for the degree distribution	2
μ	the mixing parameter which determines the ratio of the external degree on each node to the total degree of the node with respect to community	[0.1,0.7]

in the GN network ($P_{in} = 0.69$, $k_{out} = 4$), and finally we did a visual test in the Pollen single-cell data set which is described below. It can be seen that MDS have the best visualization effect. PCA has better visualization in the first three networks, but it does not work well in Pollen. LE can only effectively express the clustering of connected graphs. T-SNE plot cannot express communities. The reason is that the given distance metric is not applicable to T-SNE.

2) THE EFFECT OF PARAMETERS ON THE ALGORITHM

In equation (2), the value of the community structure coefficient γ has a direct impact on the clustering. Fig. 2 firstly displays the effects of different values on clustering visualization. It is found that the larger γ is, the better the cohesive effect of the communities is, but when γ reaches a certain level, the cohesive effect will not change any more. Therefore, we finally determined that $\gamma = 100$.

3) COMPARISON OF LFR NETWORKS

The definition of parameter of LFR network and their setting are shown in Tab. 1. By increasing the value of μ, the community structure will become increasingly blurred. When $\mu \leq 0.5$, the community structure is obvious. A view of the various μ values of the 300 nodes (other parameters are in Tab. 1) in the LFR network is shown in Fig.3. When $\mu \geq 0.6$, the community structure is invisible. This shows that DCK has a good view effect when applying MDS projection.

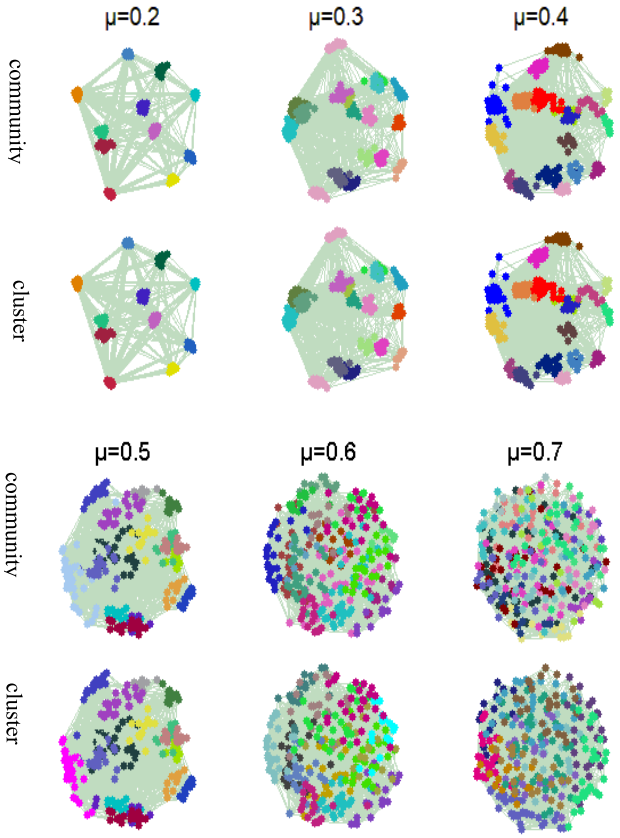


FIGURE 3. Comparison of different μ in LFR network.

We show a composite comparison of various sizes and μ in LFR network in Tab. 2. By increasing the value of μ, the community structure will become more and more blurred. When $\mu \leq 0.5$, NMI/Fmeasure/Accuracy/RI/ARI are well, and the number of community modules is basically the same as the number of real clusters. When $\mu \geq 0.6$, NMI/Fmeasure/RI/ARI is greatly reduced. There are also large deviations between the number of original communities and the number of DCK clusters. This shows that the algorithm has high accuracy in estimating the number of clusters, and other indicators are also very high.

B. REAL-WORLD NETWORK

The importance of community discovery lies in the community structure discovery in real networks. The effectiveness of the proposed algorithm has been verified by artificial networks. To further illustrate the effectiveness of the proposed algorithm, we compared the DCK algorithm with other algorithms on real-world networks. These networks include a widely used social networks for testing community detection algorithms (Zachary karate club Network [35]), dolphins [36] and several biological networks constructed by Pollen single-cell RNA-seq data [37] and Treutlin single-cell RNA-seq data [38]. We have downloaded the gene expression data of cells from the dataset. Note that Pollen data

TABLE 2. LFR network test results.

N	μ	Community number	Cluster number	Modularity	NMI	F-Measure	RI	ARI
300	0.1	10.0	10.0	0.759	1.000	1.000	1.000	1.000
300	0.2	13.8	13.8	0.671	1.000	1.000	1.000	1.000
300	0.3	15.8	15.8	0.580	1.000	1.000	1.000	1.000
300	0.4	15.8	15.8	0.487	0.996	0.997	0.999	0.993
300	0.5	15.8	15.8	0.388	0.950	0.964	0.987	0.908
300	0.6	15.8	14.6	0.245	0.755	0.736	0.938	0.554
300	0.7	15.8	10.8	0.069	0.334	0.328	0.857	0.113
1000	0.1	41.8	41.8	0.852	0.996	0.987	0.999	0.989
1000	0.2	51.4	51.2	0.748	0.992	0.972	0.999	0.973
1000	0.3	54.6	54.4	0.637	0.986	0.954	0.998	0.951
1000	0.4	54.6	54.6	0.545	0.981	0.945	0.997	0.931
1000	0.5	54.6	53.2	0.433	0.966	0.900	0.994	0.870
1000	0.6	54.6	45.4	0.322	0.905	0.791	0.988	0.737
1000	0.7	54.6	41.0	0.148	0.633	0.476	0.968	0.296
2000	0.1	85.8	84.4	0.858	0.992	0.973	0.999	0.972
2000	0.2	100.2	96.2	0.760	0.988	0.954	0.999	0.959
2000	0.3	101.4	98.4	0.661	0.989	0.960	0.999	0.964
2000	0.4	101.4	95.8	0.560	0.980	0.930	0.998	0.926
2000	0.5	101.4	91.0	0.447	0.965	0.880	0.997	0.865
2000	0.6	101.4	73.6	0.356	0.926	0.798	0.994	0.771
2000	0.7	101.4	65.6	0.145	0.644	0.429	0.980	0.267
5000	0.1	230.4	181.0	0.881	0.976	0.887	0.999	0.890
5000	0.2	253.8	171.6	0.787	0.965	0.835	0.998	0.843
5000	0.3	256.2	174.8	0.688	0.964	0.835	0.998	0.841
5000	0.4	256.2	171.4	0.586	0.962	0.824	0.998	0.830
5000	0.5	256.2	155.0	0.486	0.949	0.778	0.997	0.778
5000	0.6	256.2	126.4	0.374	0.908	0.671	0.995	0.646
5000	0.7	256.2	109.0	0.153	0.659	0.351	0.989	0.233

has 300 cells and 11 clusters, and Treutlin data has 80 cells and 5 clusters. They are not network structures. We construct two kinds of networks by KNN with K=5 and K=10. The community structure of these real networks are known, so we can use NMI, ARI, etc. to measure the performance of algorithms.

We tested the accuracy of the algorithm when determining the number of clusters. In Table 3, we list the original tag categories of several real networks and the number of clusters obtained by each algorithm, and mark the red for the exact same set, where the number of clusters of KMEANS is also provided by DCK. It can be seen that DCK can get the correct number of clusters in most cases. In addition, the NMI and other evaluation indicators of each algorithm are compared. DCK is the average of 10 iterations, and DCK-MAX indicates the best among the 10 groups. A set of results. For the best results we also marked the red. For the two- single-cell data Pollen&Treulin, the clustering results are related to the parameters of the KNN network, and the effect of k=10 is better than k=5.

Now we consider the famous Zachary Karate Club network, which has become a common workbench for community search algorithms [39], [40]. The club network is divided into two parts by DCK, which are exactly the same as the original labels, as shown in Fig 4(a). Dolphin networks are also considered. The original labels are of two types. DCK and KMEANS are of three types, in which the open symbols represent categories that do not match the original labels, as shown in Fig 4(b). Similarly, we also compared two sets of single cell data, as shown in Fig 4(c,d). The results show that under the same settings, DCK better identifies the community better than KMEANS.

TABLE 3. Real-work network test results.

data set	Method	Community number	Cluster number	NMI	F-measure	RI	ARI
karate (34 node)	MSCD_LFK	2	1	0	0.668	0.487	0
	OSLOM	2	1.8	0.363	0.786	0.658	0.323
	LPcopra	2	14.8	0.439	0.576	0.64	0.265
	LouvainSprs	2	4.3	0.659	0.821	0.782	0.559
	LouvainSgnf	2	5.9	0.57	0.74	0.73	0.452
	KMEANS	**2**	**2**	**1**	**1**	**1**	**1**
	DCK	**2**	**2**	**1**	**1**	**1**	**1**
	DCK-MAX	**2**	**2**	**1**	**1**	**1**	**1**
dolphin (62 node)	MSCD_LFK	2	1	0	0.705	0.556	0
	OSLOM	2	1.8	0.524	0.878	0.791	0.558
	LPcopra	2	6	0.652	0.871	0.836	0.681
	LouvainSprs	2	13.6	0.409	0.518	0.553	0.176
	LouvainSgnf	2	16.4	0.381	0.462	0.523	0.128
	KMEANS	2	3.2	0.631	0.808	0.761	0.539
	DCK	2	3.2	0.648	0.815	0.772	0.56
	DCK-MAX	2	3	**0.718**	**0.94**	**0.913**	**0.826**
Pollen KNN=5 (300 node)	MSCD_LFK	11	8	**0.892**	**0.845**	0.95	**0.781**
	OSLOM	11	16	0.839	0.719	0.935	0.611
	LPcopra	11	21.8	0.836	0.724	0.935	0.608
	LouvainSprs	11	18.4	0.822	0.718	0.936	0.592
	LouvainSgnf	11	27.6	0.786	0.628	0.923	0.463
	KMEANS	11	16	0.815	0.693	0.929	0.593
	DCK	11	16	0.833	0.719	0.938	0.633
	DCK-MAX	11	16	0.855	0.783	**0.955**	0.742
Pollen KNN=10 (300 node)	MSCD_LFK	11	4	0.681	0.572	0.75	0.358
	OSLOM	11	13.7	0.824	0.736	0.94	0.65
	LPcopra	11	12.5	0.879	0.798	0.952	0.741
	LouvainSprs	11	12.6	0.841	0.763	0.945	0.691
	LouvainSgnf	11	19.6	0.822	0.722	0.934	0.585
	KMEANS	**11**	**11**	0.832	0.765	0.945	0.711
	DCK	**11**	**11**	0.848	0.782	0.949	0.737
	DCK-MAX	**11**	**11**	**0.889**	**0.88**	**0.977**	**0.883**
Treutlin KNN=5 (80 node)	MSCD_LFK	5	2	0.52	0.627	0.616	0.33
	OSLOM	5	4	**0.875**	**0.92**	**0.961**	**0.913**
	LPcopra	5	3.5	0.775	0.828	0.887	0.762
	LouvainSprs	5	7.6	0.707	0.716	0.785	0.428
	LouvainSgnf	5	10.6	0.68	0.655	0.764	0.347
	KMEANS	5	6	0.681	0.706	0.79	0.452
	DCK	5	5.8	0.719	0.767	0.833	0.572
	DCK-MAX	**5**	**5**	0.823	0.913	0.929	0.836
Treutlin KNN=10 (80 node)	MSCD_LFK	5	2	0.52	0.627	0.616	0.33
	OSLOM	5	3.2	0.758	0.802	0.858	0.708
	LPcopra	5	3.5	0.725	0.762	0.831	0.639
	LouvainSprs	5	7.3	0.707	0.731	0.793	0.455
	LouvainSgnf	5	9.6	0.661	0.685	0.768	0.374
	KMEANS	5	5.6	0.779	0.772	0.845	0.602
	DCK	5	5.8	0.791	0.791	0.857	0.634
	DCK-MAX	**5**	**5**	**0.906**	**0.932**	**0.976**	**0.944**

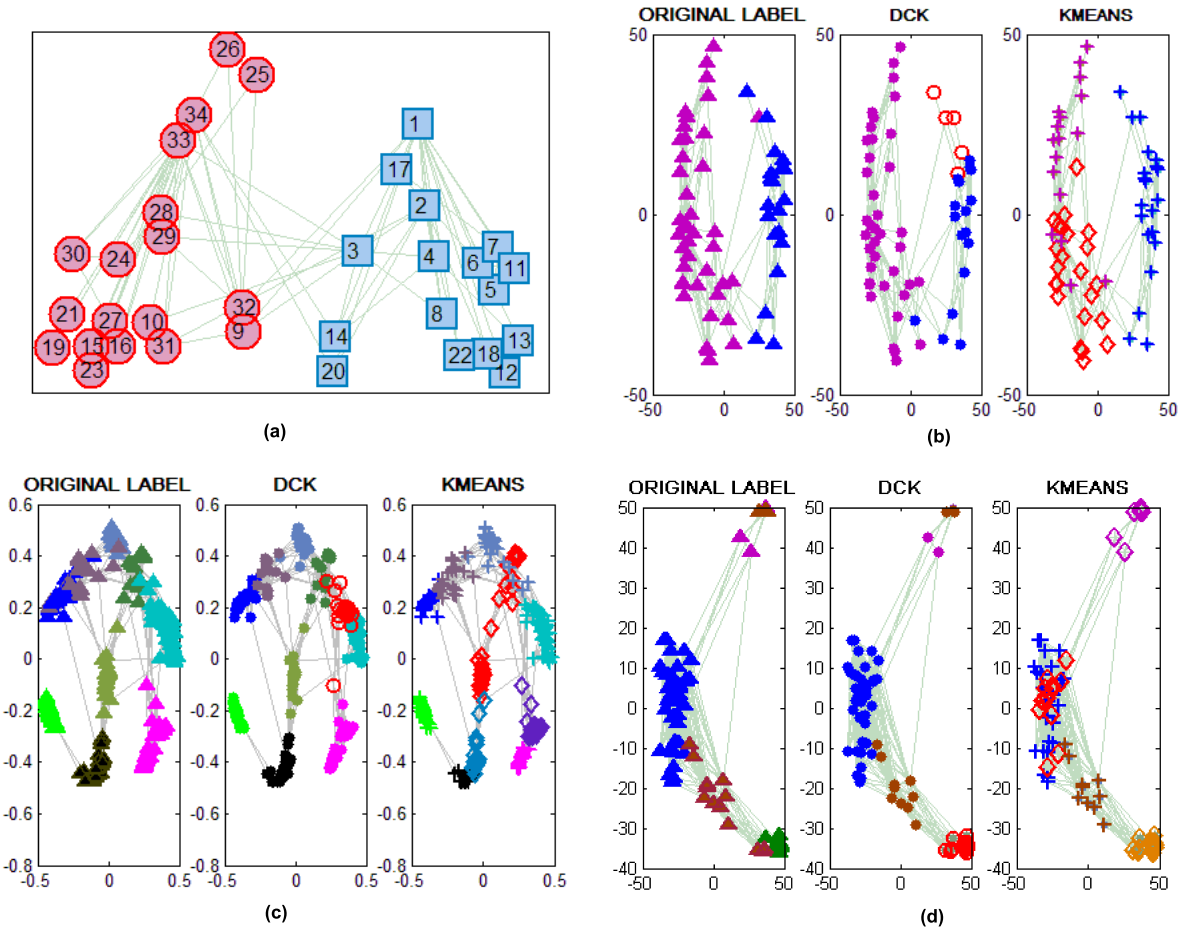


FIGURE 4. The effect of Real-world networks. (a) Visualization of Zachary karate club network and community structure detected by DCK. (b) Comparison of the original label of dolphin network and the clustering result of DCK algorithm and KMEANS algorithm. (c) Comparison of the original label of Pollen single cell (k=10) network and the clustering result of DCK algorithm and KMEANS algorithm (d) Comparison of the original label of Treutlin single cell (k=10) network and the clustering result of DCK algorithm and KMEANS algorithm, where the open symbol indicates a category that does not match the original label.

V. CONCLUSION

In this paper, we propose a Density-Canopy-Kmeans algorithm based on node adjacency matrix to detect community structures in networks. The algorithm is based on an intuitive idea that nodes within the same community are more closely connected than those between communities. For node distance metric, we first construct the Jaccard distance according to the node adjacency matrix, and add random distance and community structure coefficient to get a new distance metric that is easier for detecting the community. Then, combined with density clustering and canopy clustering, the Kmeans clustering process is improved, and thus the DCK algorithm is constructed. The algorithm obtains proper canopy sets by adjusting the parameters $r\&t$, then directly determine the number k of the clusters and the initial seeds from the number and average of the canopy sets, and finally detects the community structure in networks.

There are a few advantages of this method. First of all, the algorithm is very simple. Second, it does not require additional prior knowledge about the community structure, as long as the adjacency matrix of the network nodes is sufficient. And, a set of optimal parameters is given to effectively detect communities in networks.

To visualize the networks more effectively, the MDS projection is performed on the basis of the distance matrix, and the projection point map of the network in the two-dimensional plane is drawn. The MDS projection based on the Jaccard distance can display a more intuitive and clearer community structure while keeping important connections in the original network. Experimental results on computer synthesis and real-world networks have demonstrated the superiority of the MDS projection algorithm.

Finally, compared with traditional methods, our community detection algorithm and network visualization method have a higher classification accuracy and a better visualization effect. Our method can provide a useful tool for analyzing network structures. However, current version of our method is relatively slow especially for large networks. Moreover, some meaningful extensions of the MDS projection can be constructed in the future, and improvements can

be used for the identification and analysis of overlapping communities.

ACKNOWLEDGMENT

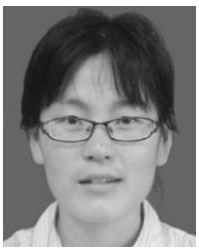
The authors appreciate Lihong Peng, Jujuan Zhuang and others for useful discussion.

REFERENCES

[1] M. Zitnik, R. Sosič, and J. Leskovec, "Prioritizing network communities," *Nature Commun.*, vol. 9, no. 1, Jun. 2018, Art. no. 2544.
[2] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
[3] L. F. Robinson, L. Y. Atlas, and T. D. Wager, "Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia," *NeuroImage*, vol. 108, pp. 274–291, Mar. 2015.
[4] R. F. Betzel, J. D. Medaglia, and D. S. Bassett, "Diversity of meso-scale architecture in human and non-human connectomes," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 346.
[5] C. Nicolini, C. Bordier, and A. Bifone, "Community detection in weighted brain connectivity networks beyond the resolution limit," *NeuroImage*, vol. 146, pp. 28–39, Feb. 2017.
[6] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Phys. A, Stat. Mech. Appl.*, vol. 388, no. 8, pp. 1706–1712, Apr. 2009.
[7] S. Gregory, "An algorithm to find overlapping community structure in networks," in *Proc. Eur. Conf. Princ. Pract. Knowl. Discovery Databases*, 2007, pp. 91–102.
[8] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, Mar. 2009, Art. no. 033015.
[9] X. Wang, G. Liu, and J. Li, "Overlapping community detection based on structural centrality in complex networks," *IEEE Access*, vol. 5, pp. 25258–25269, 2017.
[10] C. Pizzuti, "A multi-objective genetic algorithm for community detection in networks," in *Proc. 21st IEEE Int. Conf. Tools Artif. Intell.*, Nov. 2009, pp. 379–386.
[11] Y. Guo, X. Li, Y. Tang, and J. Li, "Heuristic artificial bee colony algorithm for uncovering community in complex networks," *Math. Problems Eng.*, vol. 2017, Jan. 2017, Art. no. 4143638.
[12] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS ONE*, vol. 6, no. 4, 2011, Art. no. e18961.
[13] J. Xiang, Y. Zhang, J.-M. Li, H.-J. Li, and M. Li, "Identifying multi-scale communities in networks by asymptotic surprise," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 3, Mar. 2019, Art. no. 033403.
[14] J. Xiang, Z.-Z. Wang, H.-J. Li, Y. Zhang, F. Li, L.-P. Dong, J.-M. Li, and L.-J. Guo, "Community detection based on significance optimization in complex networks," *J. Stat. Mech., Theory Exp.*, vol. 2017, no. 5, 2017, Art. no. 053213.
[15] X. Zhou, Y. Liu, and B. Li, "A multi-objective discrete cuckoo search algorithm with local search for community detection in complex networks," *Mod. Phys. Lett. B*, vol. 30, no. 7, 2016, Art. no. 1650080.
[16] N. Masuda, M. A. Porter, and R. Lambiotte, "Random walks and diffusion on networks," *Phys. Rep.*, vols. 716–717, pp. 1–58, Nov. 2017.
[17] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, 2008.
[18] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Proc. Int. Symp. Comput. Inf. Sci.*, 2005, pp. 284–293.
[19] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, Oct. 2010, Art. no. 103018.
[20] J. Huang, H. Sun, Y. Liu, Q. Song, and T. Weninger, "Towards online multiresolution community detection in large-scale networks," *PLoS ONE*, vol. 6, no. 8, 2011, Art. no. e23829.
[21] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[22] Q. Cai, L. Ma, M. Gong, and D. Tian, "A survey on network community detection based on evolutionary computation," *Int. J. Bio-Inspired Comput.*, vol. 8, no. 2, pp. 84–98, 2016.
[23] L. Qiao, M. Efatmaneshnik, M. Ryan, and S. Shoval, "Product modular analysis with design structure matrix using a hybrid approach based on MDS and clustering," *J. Eng. Des.*, vol. 28, no. 6, pp. 433–456, 2017.
[24] J. Xue, C. Lee, S. G. Wakeham, and R. A. Armstrong, "Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean," *Organic Geochem.*, vol. 42, no. 4, pp. 356–367, 2011.
[25] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, Dec. 2001, no. 6, pp. 585–591.
[26] S. Xu, X. Hua, J. Xu, X. Xu, J. Gao, and J. An, "Cluster ensemble approach based on T-distributed stochastic neighbor embedding," *J. Electron. Inf. Technol.*, vol. 40, no. 6, pp. 1316–1322, Jun. 2018.
[27] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.
[28] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula," *Inf. Process. Manage.*, vol. 25, no. 3, pp. 315–318, 1989.
[29] D. J. Hand, "Principles of data mining," *Publications Amer. Stat. Assoc.*, vol. 98, no. 461, pp. 252–253, 2007.
[30] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[31] G. Zhang, C. Zhang, and H. Zhang, "Improved K-means algorithm based on density Canopy," *Knowl.-Based Syst.*, vol. 145, pp. 289–297, Apr. 2018.
[32] M. Dianhui, "Improved canopy-k means algorithm based on MapReduce," *Comput. Eng. Appl.*, vol. 48, no. 27, pp. 22–26, 2012.
[33] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, 2004, Art. no. 026113.
[34] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, 2008, Art. no. 046110.
[35] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropolog. Res.*, vol. 33, no. 4, pp. 452–473, 1977.
[36] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behav. Ecol. Sociobiol.*, vol. 54, no. 4, pp. 396–405 2003.
[37] A. A. Pollen, "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature Biotechnol.*, vol. 32, no. 10, pp. 1053–1058, Aug. 2014.
[38] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq," *Nature*, vol. 509, no. 7500, pp. 371–375, May 2014.
[39] H. Shakeri, P. Poggi-Corradini, N. Albin, and C. Scoglio, "Network clustering and community detection using modulus of families of loops," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 1, Jan. 2017, Art. no. 012316.
[40] Y. Chen, X. Wang, J. Bu, B. Tang, and X. Xiang, "Network structure exploration in networks with node attributes," *Phys. A, Stat. Mech. Appl.*, vol. 449, pp. 240–253, May 2016.

MANZHI LI received the master's degree in computational mathematics from Northwestern Polytechnical University. She is currently an Associate Professor with the School of Mathematics and Statistics, Hainan Normal University. His research interests include bioinformatics, complex networks, and clustering algorithms.

HONGTAO WANG received the master's degree in applied mathematics from Hainan Normal University, where he is currently an Associate Professor with the School of Mathematics and Statistics. His research interests include mathematical modeling and computer numerical simulation.

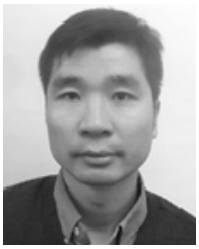


HAIXIA LONG received the Ph.D. degree in light industry information technology and engineering from Jiangnan University. She is currently an Associate Professor with the School of Information Science and Technology, Hainan Normal University. His research interests include swarm intelligence algorithm, deep learning, and bioinformatics.

JU XIANG received the B.S. and M.S. degrees from Xiangtan University, in 2005 and 2008, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Central South University, Changsha, China. He is also an Associate Professor with Changsha Medical University, China. His research interests include complex networks, bioinformatics, machine learning, and deep learning.

BO WANG was born in 1993. He received the B.S. degree from the Department of Biology, Shandong University, in 2014, and the M.S. degree from the Institute of Microbiology, Chinese Academy of Sciences, in 2017. He is currently a Bioinformatics Scientist with Geneis Beijing Company Ltd. His research interests include bioinformatics and oncology.

JUNLIN XU was born in 1993. He is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Hunan University, China. His research interests include bioinformatics, machine learning, biochemical research method, disease-associated non-coding RNA, and single cell.



JIALIANG YANG received the Ph.D. degree from the Department of Mathematics, National University of Singapore, in 2009. From 2010 to 2011, he was an Assistant Professor with the CAS-MPG Partner Institute for Computational Biology, China. After that, he moved to USA and became a Postdoctoral Fellow at the Department of Basic Sciences, Mississippi State University. In 2013, he had become a Postdoctoral Fellow and a Senior Scientist at the Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai. He is currently the Vice President of Geneis Beijing Company Ltd., and a Visiting Professor with Hainan Normal University. He has published more than 60 peer-reviewed articles. His main research areas include bioinformatics, machine learning, oncology, aging, and evolutionary biology.