Information Sciences 632 (2023) 1–18

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Large-scale community detection based on core node and layer-by-layer label propagation

Weitong Zhang [a,b,*], Ronghua Shang [a], Licheng Jiao [a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi Province 710071, China
[b] The Guangzhou Institute of Technology, Xidian University, Guangzhou, Guangdong Province 510555, China

ARTICLE INFO

Keywords:
Community detection
Core node
Layer-by-layer label propagation
Overlapping community

ABSTRACT

With the vigorous development of big data, the task of large-scale community structure detection is more challenging. In this paper, a large-scale community detection method based on core node and layer-by-layer label propagation is proposed, which is further extended to the detection of overlapping community structures. First, the core nodes whose node degree is greater than the average degree in the graph are found to make effective use of the feature that the core node is the potential community center. This will also avoid the impact of nodes with low node degrees on community structure detection. Then, starting from the core node, label propagation is carried out layer-by-layer according to the node degree and node connection, which effectively improves the accuracy of community detection. The node labels after label propagation are calibrated according to the current attraction of the community to the nodes, which effectively improves the situation of misclassification in the early community detection. Finally, overlapping community detection is carried out based on the non-overlapping community structure. At this time, the overlapping community detection result is more accurate and interpretable. The community detection results on 3 synthetic networks and 12 real datasets show that the proposed algorithm has more advantages than four non-overlapping community detection methods and two overlapping community detection methods.

1. Introduction

A complex network is a graph structure abstracted from a complex system. Its expression originates from the description of individuals and relationships in graph theory [1]: nodes in the graph represent individuals, and edges (lines between nodes) in the graph represent relationships among individuals. The study of complex graphs is helpful to understand the characteristics of complex systems [2]. We can not only observe the overall interaction of the system from the perspective of the whole complex system, but also find the law of the interaction between individuals in the system [3]. There are three main characteristics of complex graphs that are often discussed. Small world [4], the average value of the shortest distance between all nodes increases logarithmically or very slowly with the increase of graph scale. Scale-free [5], the degree distribution of nodes in the graph obeys the characteristics of a power-law distribution. The degree orders of most individuals are very small, and the degree orders of a few individuals are very large. Community structure [6] is a subgraph structure with tight internal connections and sparse external connections. Among them, the study of

* Corresponding author.
E-mail address: wtzhang_1@xidian.edu.cn (W. Zhang).

https://doi.org/10.1016/j.ins.2023.02.090
Received 16 August 2022; Received in revised form 21 February 2023; Accepted 26 February 2023
Available online 3 March 2023

community structure information of complex graphs is helpful to further analyze the topology of graphs and the interaction between individuals. For example, the community structure analysis of a biological network diagrams can not only obtain several modules reflecting biological functions, but also infer the role of individuals as small as molecular structure in the process of biological evolution through the specific position of biological molecular individuals in these functional modules [7]. Individuals in the graph may have multiple attributes, so some individuals may belong to multiple subgraphs at the same time, forming overlapping community structures [8].

In recent years, community detection methods can be divided into three categories according to search rules: methods based on modularity degree function [9], methods based on dynamics, and methods based on graph topology. Newman et al. proposed a modularity function to test the advantages and disadvantages of community structure division in 2004. After that, many algorithms based on optimizing modularity functions have been proposed and improved. Guimera et al. first used the modularity function as the objective function of the simulated annealing method to search the community structure [10]. Because of its strong global search ability, the evolutionary algorithm is widely used in the field of modularity optimization [11]. Such as the genetic algorithm [12], ant colony algorithm [13], and particle swarm optimization algorithm [14]. Liu et al. combined the migration operator with the classical genetic operator and proposed an evolutionary clustering method to detect the evolutionary structure of a dynamic community [15]. Tian et al. proposed a fuzzy overlapping group detection method based on an evolutionary multi-objective optimization algorithm. The overlapping community detection results are obtained by calculating the membership degree between nodes and community centers [16]. However, the running speed of the evolutionary algorithm is slow, and it is not suitable for community detection of large-scale graphs. Aiming at the resolution problem of modularity function in the process of increasing data scale [17], some multi-objective optimization methods based on modularity density have been proposed [18]. Due to its low complexity, the community integration strategy is also commonly used to search for the maximum modularity degree or modularity density [19]. Ye et al. embedded nodes into low dimensional space through a transformation matrix, and proposed an adaptive affinity matrix learning method to improve the effectiveness of community detection from the perspective of node internal similarity [20]. Among the community structure search methods based on dynamics, the most common is the graph representation method based on random walk [21]. It is usually necessary to define the length of the walk step and probability transition matrix. Deepwalk [22] is the first node embedding method based on the random walk, which searches the random walk path of nodes in the network to obtain the node sequence. Then, the skip-gram model and hierarchical softmax model are used to model the probability of nodes in the sequence. The Node2vec method [23] is further extended based on Deepwalk. The random walk of nodes in Deepwalk is uniformly and randomly distributed. Node2vec controls the width and depth priority search in the generation of node sequence by introducing two search bias parameters. Width first search focuses on finding the local network structure, while the depth first search focuses on finding the similarity between nodes. The objective functions of Deepwalk and Node2vec are easy to fall into local optimization.

Label propagation algorithm is a fast and widely used strategy in community detection. Label propagation rules often depend on graph topology information such as node degree and commonly connected nodes [24]. Raghavan et al. used the label propagation algorithm for community structure information detection for the first time [25]. The traditional label propagation algorithm has strong randomness and weak robustness, so it is often easy to fall into local optimization and difficult to obtain more effective community detection results. Yu et al. proposed an overlapping community detection method based on DeepWalk and improved label propagation [26]. This method uses the DeepWalk model to learn the network topology information. After obtaining the low-dimensional vector representation of nodes, it carries out vector dot product operation and constructs the weight matrix. At the same time, a label propagation algorithm with a preference selection strategy is designed to detect overlapping community structures by exchanging information with fixed neighboring nodes. Lu et al. proposed an improved label propagation algorithm based on the influence of neighbor nodes for overlapping community detection [27]. The algorithm detects overlapping community structures by using a fixed label propagation sequence based on ascending node importance and a label update strategy based on neighbor node influence and historical label priority strategy. Some scholars set the node label update rule of the label propagation algorithm as the growth of modularity value, and carried out label propagation to optimize the modularity [28]. To improve the disadvantage that the label propagation algorithm is easy to fall into local optimization, Liu et al. added a multi-step greedy [29] fusion method while optimizing the modularity [30]. Although this method improves the final community detection results, it increases the complexity of the algorithm itself and sacrifices the advantages of fast running and low complexity of the label propagation algorithm. Yazdanparast et al. proposed an accelerated modularity gain method for community detection based on label propagation by analyzing Newman's modularity gain function of label propagation in the graph [31]. Because modularity itself contains the problem of resolution limitation, the label propagation algorithm based on modularity still has its limitations. Lin et al. proposed a community kernel label propagation algorithm (CKLPA) [32], which improves the randomness of the label propagation algorithm and reduces its complexity of the algorithm. However, the number of community cores of the algorithm needs to be given in advance, so the detection results of the network may become random due to the selection of the number of community cores. Zhang et al. proposed a new LPA algorithm combining multi-layer neighborhood overlap and historical label similarity [33]. The algorithm considers both node update order and label selection rules. The label entropy is used as the basis of node update order, and the multi-layer neighborhood overlap and historical label similarity used to calculate node preference are defined. Zhao et al. proposed a large-scale community detection method based on graph compression and label propagation algorithm [34]. They compressed the nodes with low node degrees in the graph, and detected the community structure according to the similarity and other information after finding the community center. However, the search results of the community center have a great impact on the final community detection results, which is not conducive to the improvement of the community detection effect.

Based on the advantages and disadvantages of the above methods, a large-scale community detection method based on core node and layer-by-layer label propagation is proposed. The overlapping community structure information based on the community structure

information is further searched. The core node is the most influential node in the complex graph, which is often used as the central node of the community structure in the process of community detection. The "key [35]" feature of the core node is used to find the node with the largest node degree in the neighborhood of the graph structure as the core node. To prevent nodes with too small degree in the sparse graph from interfering with the final community detection results, the node degree greater than the average node degree is added as an additional rule for core node search. After the core node search, the label propagation algorithm is used to update the node label. The proposed algorithm can improve the accuracy of the label propagation algorithm, while largely preserving its linear complexity. Specifically, starting from the neighbor node directly connected to the core node, the node label is updated layer-by-layer according to the joint connected nodes and node degree of the core node. The nodes whose labels change at each layer are added to the core node set. The label propagation is repeated until all nodes in the graph are traversed. In this paper, the effectiveness of label propagation process mainly depends on the search results of core nodes, node degree and connection between nodes. In order to further improve the accuracy of community detection results, based on the node label update results obtained by the above method, node labels are recalibrated according to the current attraction of the community to the node. Overlapping community structure, that is, the nodes in the graph belong to multiple community structures at the same time. Therefore, after obtaining the non-overlapping community structure, the overlapping membership of nodes in each community is calculated according to the node degree and the connection between nodes and communities to obtain the overlapping community structure. While obtaining the non-overlapping community structure in the network, it is more explanatory and practical to detect the possible overlapping nodes. The main contributions of this paper are as follows.

(1) According to the potential community center characteristics of core nodes, find the core nodes whose node degree is greater than the average node degree in the graph. Starting from the core node, layer-by-layer label propagation gives play to the advantage of fast label propagation, and effectively improves the accuracy of community detection.
(2) According to the attraction of the current community to the nodes, the node labels after label propagation are calibrated, which can avoid misclassification in the early community structure detection.
(3) Judging the overlap of nodes based on the non-overlapping community structure can obtain more accurate and interpretable overlapping community detection results.

2. Traditional label propagation algorithm

As the name suggests, label propagation refers to the propagation of labels in network nodes according to certain rules. When all

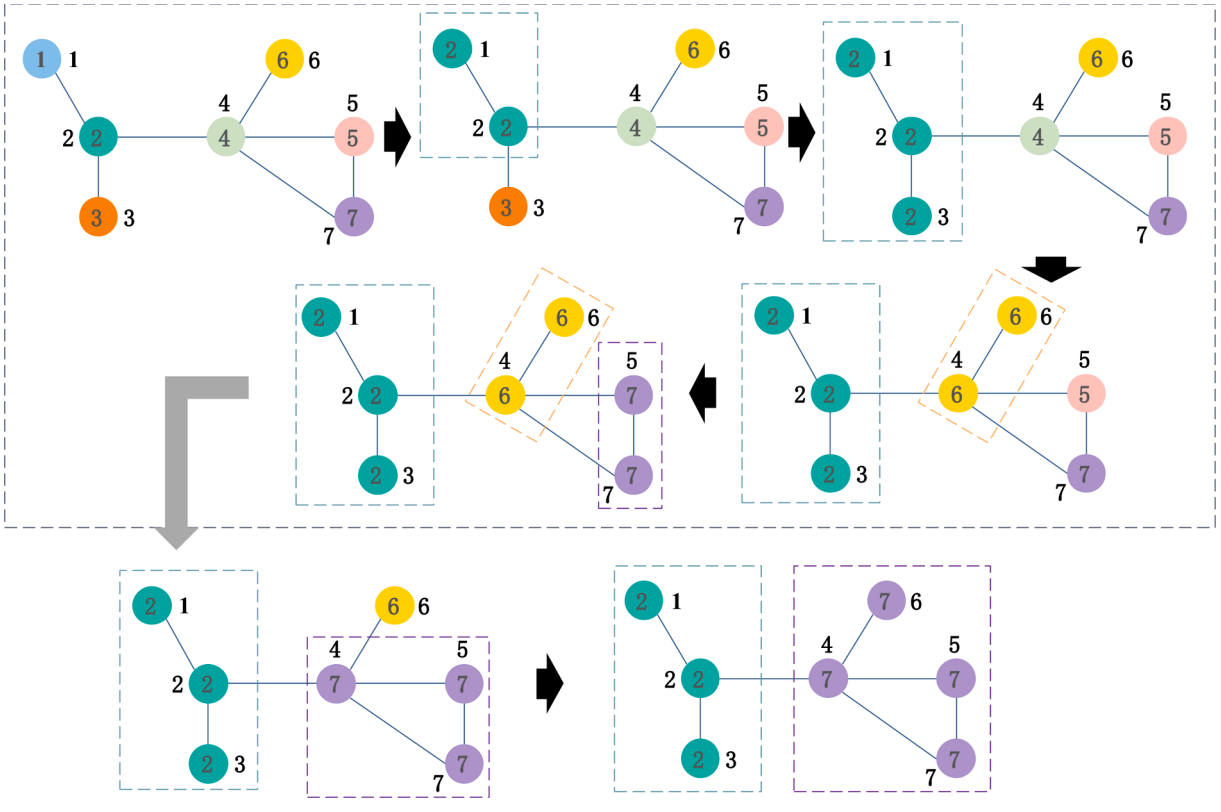

Fig. 1. Example of traditional label propagation process.

node labels reach a stable state and no longer change, the same label at this time indicates that the nodes belong to the same community. The common rule of the traditional label propagation algorithm is to update the node label according to the most labels in its neighborhood. This label propagation rule has low complexity, but low accuracy, and high requirements for the selection of initial nodes. The specific update rules are shown in Fig. 1.

As shown in Fig. 1, it is assumed that there are 7 nodes and 7 edges in the initial network. The initial network node label is 1–7. Starting from node 1, search for the largest number of labels in its neighbor nodes and update the labels of node 1. When the largest number of neighbor node labels is not unique, a label is randomly selected for updating. There is no case of multiple labels. Only one of the possible results of label propagation is given in Fig. 1. During the first node traversal, the label of node 1 is first updated to 2. At this time, the neighbor node labels of node 2 include nodes 2, 3, and 4 all with the number 1. In this example, it is assumed that label of node 2 is randomly selected. Therefore, the label of node 2 remains unchanged, and the label of node 3 is naturally updated to 2. Then, the neighbor labels of node 4 are 2, 5, 6 and 7, and the number is 1. In this example, it is assumed that label 6 is randomly selected as the updated label of node 4. Then the neighbor labels of node 5 are labels 6 and 7. In this example, it is assumed that label 7 is randomly selected as the new label of node 5. The label of node 6 remains unchanged refer to node 4. The neighbor labels of node 7 are labels 5 and 7. It is assumed that 7 is randomly selected as the label for node 7. At this time, the first traversal of label propagation ends, and the node label has not reached stability. The second label propagation traversal from node 1 is carried out. Following the above node label rules, it can be found that the labels of nodes 1 to 3 remain unchanged at the end of the first label propagation. The label of node 4 is updated to 7. The label of node 6 is thus updated to 7. The label of node 5 and 7 remains unchanged. Then the network node labels become stable, as shown in the network structure diagram in the bottom right-hand corner of Fig. 1. The labels of nodes 1, 2 and 3 are 2, and the labels of nodes 4, 5, 6 and 7 are 7. The same labels of nodes indicates that they belong to the same community. Therefore, in this example, the final result divides the network into two communities to complete the label propagation process.

3. Community detection method based on core nodes and layer-by-layer label propagation

Firstly, the non-overlapping community is detected layer by layer based on the core nodes and label propagation strategy.

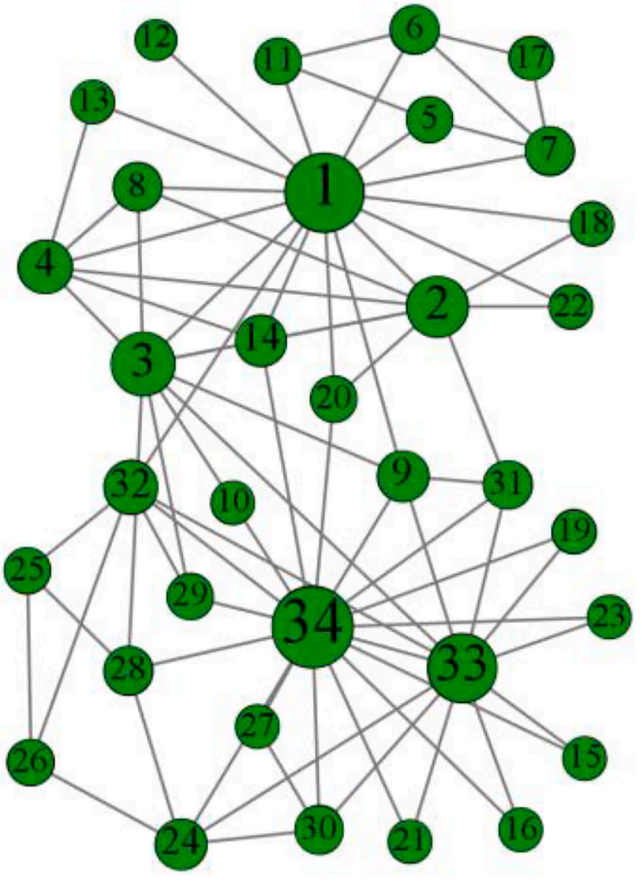

Fig. 2. Node degree of Karate dataset.

3.1. Searching core nodes

Firstly, the core node in the graph is searched, and the specific rule is the node with the largest node degree in its neighborhood. Fig. 2 shows the node degree distribution of the Karate dataset. The larger circle size corresponds to the larger node degree.

As can be seen from Fig. 2, nodes 1 and 34 have the largest node degree in the neighborhood and are selected as core nodes in this algorithm.

Fig. 3 shows the real community division of the Karate dataset. Nodes circled by dotted lines belong to the same community. As shown in Fig. 3, node 1 and node 34 are respectively located in the center of the two real community structures in the graph. It can be seen that compared with the method of randomly traversing the whole graph from one node to update the label, determining the core nodes first is more helpful to improve the accuracy of community detection.

However, the structure of many graphs is sparse, and the degree of many nodes in the graph is low. If all the nodes with the largest node degree in the neighborhood are taken as the core nodes, it may cause excessive subdivision of the community structure and reduce the accuracy of community detection. Therefore, a core node search rule is added to the proposed algorithm. That is, only the nodes whose node degree is greater than the average node degree and has the largest node degree in the neighborhood are regarded as the core nodes. Taking PGP dataset as an example, there are 10,680 nodes in the graph, including the main elements of the PGP algorithm user network. The average node degree is 4.55, with 24,340 edges in total. Find 454 nodes with the largest node degree in the neighborhood in the PGP graph, and the corresponding modularity value is 0.6653. After adding the search rule that the node degree of the core nodes must be greater than the average node degree, the number of core nodes is 133 and the corresponding modularity value is 0.7378.

3.2. Layer-by-layer label propagation algorithm based on core nodes and topology information

After obtaining the core node search results, the nodes in the graph are propagated layer by layer according to the graph topology information. Specifically, all core nodes obtained in Section 3.1 are regarded as independent initial communities and assigned unique labels. Start the node label propagation of the first layer. Firstly, find the nodes in the neighborhood of the core node that is connected to multiple core nodes at the same time. The label of such a "common neighbor" node cannot be determined simply according to the node connection. This paper proposes an updated formula of the "common neighbor" node label as follows.

$$S(i,c) = l_{ic} + \sum_{j \in N_{ic}} \frac{d_j}{d_{ave}} \tag{1}$$

where $S(i, c)$ represents the similarity between node i and community c, l_{ic} represents the number of connected edges between node i and community c, N_{ic} represents the node set connected with node i and community c at the same time, d_j represents the node degree of

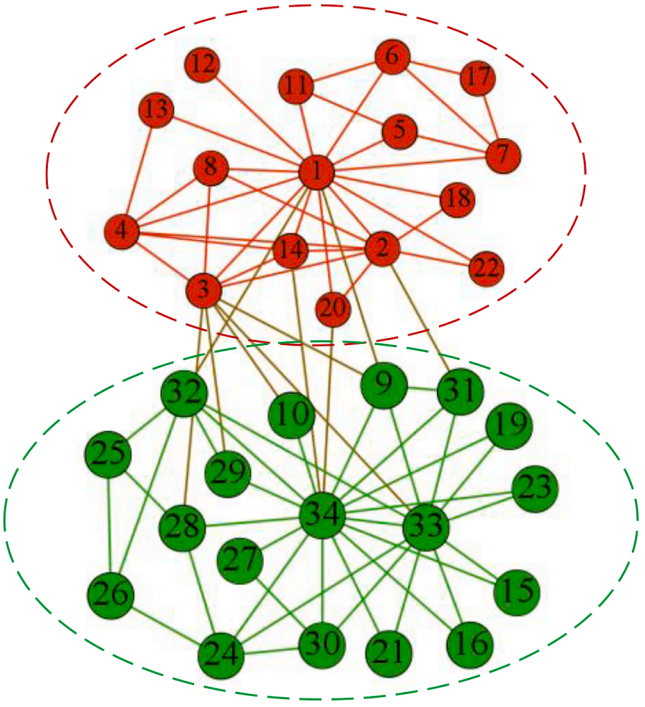

Fig. 3. Real community division of Karate dataset.

node j, and d_{ave} represents the average node degree of nodes in the graph.

Update the label of the "common neighbor" into the label with the highest similarity according to the formula (1). When the maximum similarity is not unique, a community label is randomly selected. After all the labels of "common neighbor" in the connected nodes of the first layer are updated, the remaining connected node labels can be obtained according to the connected core node labels. The nodes with updated labels are added to the initial independent communities to form a new temporary community structure. Then, the second layer of node label propagation is carried out.

The community neighbor node in the second layer is the neighbor of the community neighbor node in the first layer. Find the neighbor nodes outside each community that are connected to multiple communities at the same time. Assign a label to the "common neighbor" node according to Eq. (1) again. The labels of the remaining neighbor nodes on layer two are updated to the labels of connected communities, until all nodes are traversed. The process of layer-by-layer label propagation is shown in Fig. 4.

A simple example of layer-by-layer label propagation is shown in Fig. 4. There are 20 nodes in the example network. The arrangement of nodes is related to the connection relationship between nodes, and the connection is not given. The search results of core nodes in the network are shown in Fig. 4 (a), which are nodes 1 and 2. The first layer neighbor refers to all nodes directly connected to node 1 and node 2 in the network. The common neighbors of nodes 1 and 2 are nodes 3 and 4.

The second layer neighbor refers to all nodes in the network directly connected to the neighbor nodes of core nodes 1 and 2. The same as the traditional label propagation algorithm, all nodes in the network are labeled 1–20. The first layer of label propagation is

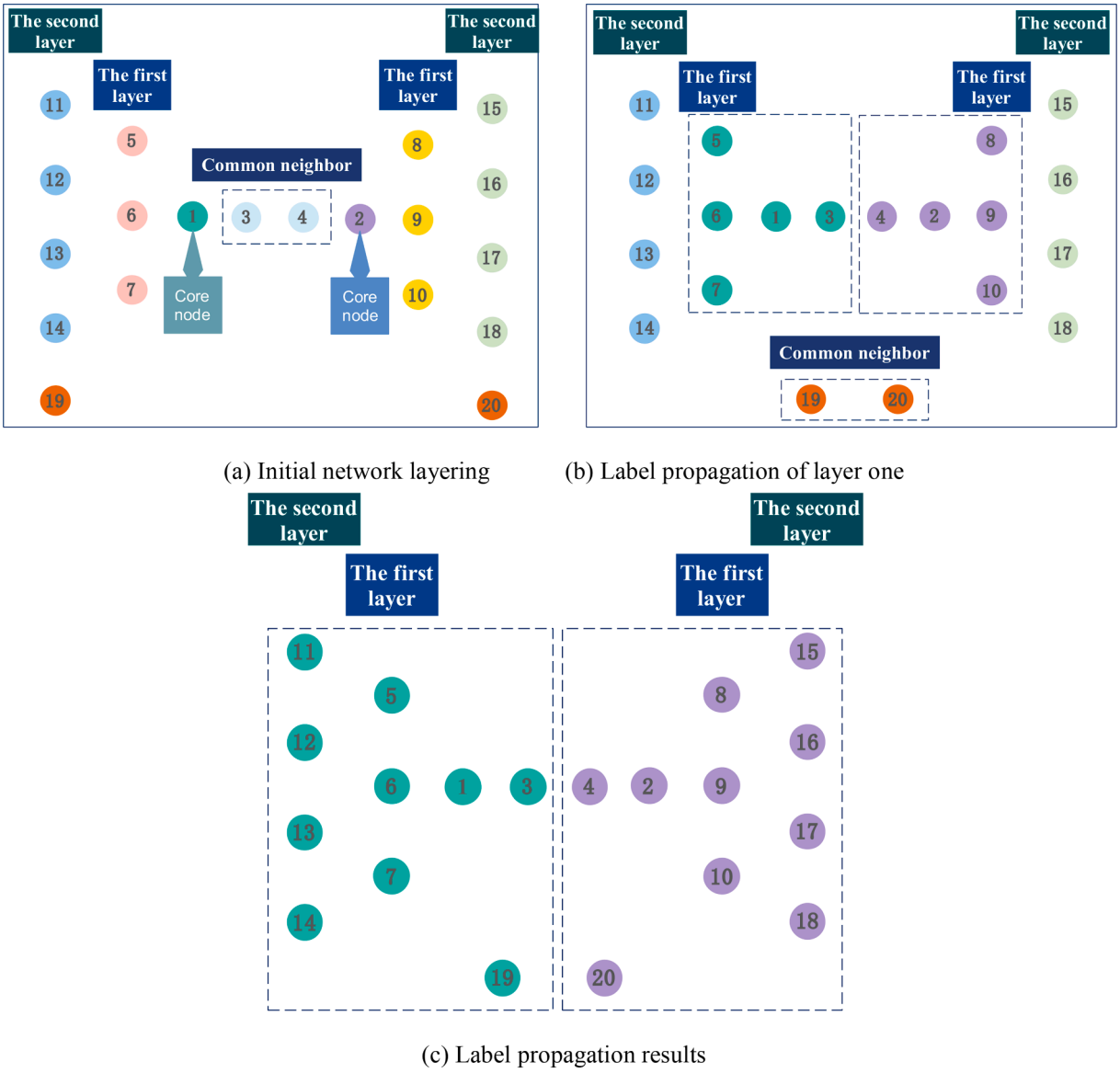

Fig. 4. Example of layer-by-layer label propagation.

carried out. At first, the common neighbors of core nodes 1 and 2 are nodes 3 and 4. The similarity between nodes 3 and 1, nodes 4 and 2 is calculated according to Eq. (1). Assuming that node 3 is more similar to node 1 and node 4 is more similar to node 2, according to the label propagation rules in this paper, node 3 is added to the community where node 1 belongs to. Node 4 is added to the community where node 2 belongs to. At this time, if the similarity is equal, the label will be randomly selected for propagation. The labels of nodes 5, 6 and 7 directly connected to node 1 are updated to the labels of node 1, and the labels of nodes 8, 9 and 10 directly connected to node 2 are updated to the labels of node 2. The division of the network community structure at this time is shown in Fig. 4 (b).

Then, the second layer of label propagation is carried out. At this time, the community structure of the core node is regarded as a new set of core nodes, and its common neighbors are nodes 19 and 20. Similarly, the similarity between nodes 19 and 20 and the core node set is calculated according to Eq. (1). In this example, assuming that node 19 is more similar to core node 1 and node 20 is more similar to core node 2. Node 19 is added to the community where node 1 belongs, and node 20 is added to the community where node 2 belongs according to the label propagation rules in this paper. Then, the labels of nodes 11, 12, 13 and 14 directly connected to core node 1 are updated to the labels of node 1, and the labels of nodes 15, 16, 17 and 18 directly connected to core node 2 are updated to the labels of node 2. At this time, the layer-by-layer label propagation ends, and the final community structure division of the network is shown in Fig. 4 (c).

3.3. Node label calibration based on community-node attraction

After the above label propagation is terminated, all node labels in the graph have been assigned. At this time, effective community structure information has been obtained. To obtain more accurate community structure information, a node label calibration method based on the degree of community attraction to nodes is introduced. The wrong label updates in the process of label propagation are calibrated and corrected. The specific calibration rules are based on the following Equation.

$$Ca(i,c) = \frac{l_{ic}}{|c|} \tag{2}$$

where, $Ca(i, c)$ represents the degree of attraction of community c to node i, $|c|$ represents the number of nodes in community c. Since

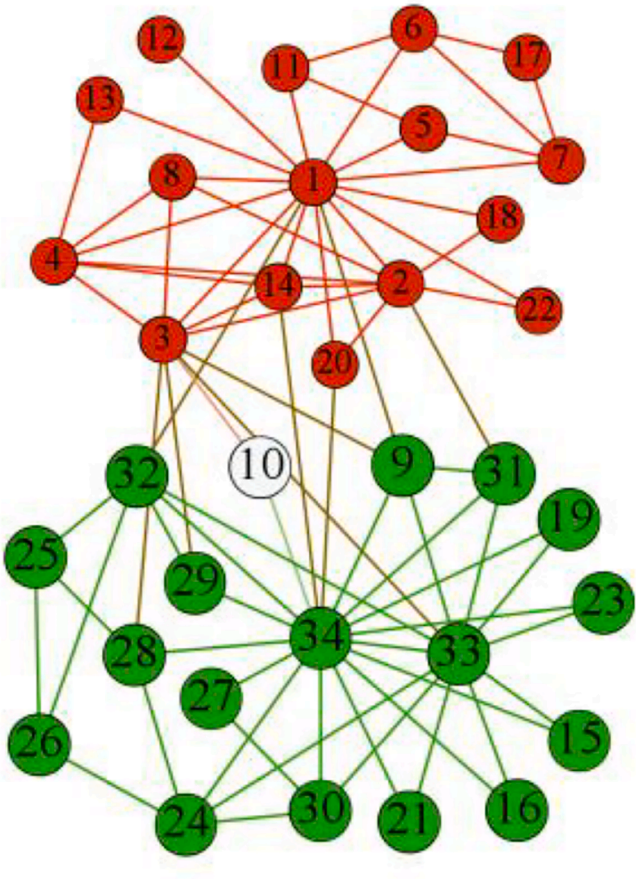

Fig. 5. Detection of overlapping community nodes on the Karate dataset.

many nodes are connected to multiple communities at the same time, the calibration formula is used to determine which community is more suitable for the node. To prevent communities with low information transmission ability from affecting the results in the calibration process, the node label calibration in this section starts from the nodes with a larger degree.

3.4. Overlapping community detection method based on node community membership

So far, complete non-overlapping community test results have been obtained. In most real networks, the same node may have many different attributes, so it may belong to multiple different subgraphs at the same time. For example, authors in cooperative networks can collaborate with multiple research groups, and users in social networks can join multiple interest groups [36]. This section introduces a method to detect the ownership of nodes belonging to communities base on non-overlapping communities. The calculation equation for node i belonging to community c is as follows.

$$As(i,c) = \frac{l_{ic}}{d_i} \tag{3}$$

where, $As(i, c)$ represents the ownership degree of node i to community c, d_i represents the node degree of node i. When the maximum value of $As(i, c)$ is not unique, it indicates that node i may belong to multiple community structures at the same time. At this time, node i is recorded as the overlapping community node. Fig. 5 shows the detection of overlapping community nodes on the Karate dataset.

After completing the non-overlapping community detection, the Karate network is divided into two community structures. By calculating the node community membership, it can be found that, as shown in Fig. 5, node 10 belongs to two different communities to the same extent. It is detected as an overlapping node in the overlapping community structure.

3.5. Time complexity analysis

In this paper, non-overlapping community detection mainly includes three steps: core node search, layer-by-layer label propagation, and node label calibration. Overlapping community detection mainly includes four steps. An overlapping community detection process based on node-community membership is added to the three steps of non-overlapping community detection. Suppose there are n nodes and m edges in the network. In the core node search part, first, calculate the degree of each node in the network. The time complexity is O(n). Then, each node and its neighbor nodes are traversed for core node filtering. The complexity is about O(m). The complexity of layer-by-layer label propagation is about O(lm), where l is the average number of layers of label propagation. The complexity of node label calibration is O(n). Therefore, the time complexity of non-overlapping community detection is approximately O(m). The complexity of overlapping community detection processes based on node-community membership is about O(n). Therefore, the time complexity of overlapping community detection is also approximately O(m).

4. Experimental results and analysis

4.1. Comparison algorithm and datasets

This paper selects four non-overlapping community detection methods and two overlapping community detection methods for comparison. Non-overlapping community detection methods: LPA [25], IsoFdp[37], ECD [15], RMOEA[38], and CDEP [34]. Overlapping community detection methods: DNMF [39] and EMOFM [16]. The parameters of each comparison algorithm refer to the original references. In the ECD algorithm, the maximum number of iterations is set to 100, the population size is set to 100, the mutation probability is set to 0.2, and the migration probability is set to 0.5. In the RMOEA algorithm, the probability of crossover and mutation is set to 0.9 and 0.1 respectively, the neighborhood size is set to 40. In the EMOFM algorithm, the crossover probability is set to 1, the mutation probability is set to $1/n$, where n is the number of nodes, and the distribution index of crossover and mutation is set to 20. In addition, all the algorithms are implemented using MATLAB. The processor used in the experiment is Intel(R) Core(TM) i5-4590 CPU@3.30 GHz and memory is 8.00 GB.

To verify the effectiveness of this algorithm, non-overlapping community detection will be performed on 3 synthetic networks and 12 real network datasets, and extended overlapping community detection experiments will be performed on the real network datasets. Among them, the LFR network is used as the synthetic network, which contains many parameters and can adjust the network size, the degree distribution of nodes (average degree of nodes $k_{average}$, maximum node degree k_{max}, degree distribution index τ_1 and τ_2), community size ($[C_{min}, C_{max}]$), etc. In addition, parameter μ controls the fuzzy degree of the LFR network community boundary. With the increase of parameter μ, the community structure in the network becomes fuzzier, and the detection difficulty increases. In this section, three LFR networks of different scales with the number of nodes of 1000 (LFR1), 5000 (LFR2) and 10,000 (LFR3) are selected.

Table 1
Specific information on synthetic networks.

Datasets	nodes	kaverage	kmax	τ_1	τ_2	$[C_{min}, C_{max}]$	Ref
LFR1	1000	20	50	2	1	[10,50]	[40]
LFR2	5000	20	50	2	1	[20,100]	[40]
LFR3	10,000	20	50	2	1	[20,100]	[40]

The node scale of datasets ranges from tens to millions. See Tables 1 and 2 for details.

4.2. Evaluation indicators

In this paper, the evaluation indicators are divided into non-overlapping community evaluation indicators and overlapping community evaluation indicators.

Non-overlapping community evaluation indicators:

(1) Normalized mutual information, *NMI* [49].

Normalized mutual information detects the effectiveness of network division results based on real labels, and its definition is shown in Eq. (4).

$$NMI\ (A,B) = \frac{-2\sum_{u=1}^{M_A}\sum_{v=1}^{M_B} M_{uv}\cdot\left(\frac{M_{uv}\cdot n}{M_u\cdot M_v}\right)}{\sum_{u=1}^{M_A} M_u \log\left(\frac{M_u}{n}\right) + \sum_{v=1}^{M_B} M_v \log\left(\frac{M_v}{n}\right)} \tag{4}$$

where $A(B)$ represent two kinds of different divisions, M represents the division information matrix, the element M_{uv} in the matrix indicates that the node belonging to the u community in division A belongs to the v community in division B, M_A indicates the total number of communities in division A, M_B indicates the total number of communities in division B, M_u indicates the sum of the u-th row elements in the division information matrix M, and M_v indicates the sum of the v-th row elements in the division information matrix M.

(2) Modularity, Q.

The modularity function is one of the most commonly used indicators to evaluate the advantages and disadvantages of network community structure division. The larger value of modularity presents the closer the internal connection of the community structure in the network. For undirected and unweighted networks, the modularity function is calculated as in Eq. (5).

$$Q = \frac{1}{2m}\sum_{i,j=1}^{n}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\sigma(c_u, c_v) \tag{5}$$

where i, j represent nodes in the network, n represents the number of communities, m represents the total number of edges in the whole network, d_i represents the node degree of node i, c represents the community number, A_{ij} represents the adjacency matrix of the network structure, when $A_{ij} = 1$, there is an edge connection between nodes i and j, and when $A_{ij} = 0$, there is no edge between them, σ represents the kronecker's delta.

The modularity function can also be expressed as Eq. (6).

$$Q = \frac{1}{2m}\sum_{c=1}^{k}\left[2l_c - \frac{(d_c)^2}{2m}\right] \tag{6}$$

where l_c represents the total number of edges and d_c represents the sum of node degrees in community c.

(3) Modularity density, D.

For the resolution of Modularity, another commonly used evaluation standard in community detection is the modularity density function. The modularity density function is shown in Eq. (7).

$$D = \sum_{u=1}^{N}\frac{L(c_u, c_u) - L(c_u, \overline{c_u})}{|c_u|} \tag{7}$$

where c_u represents a community in the network partition result, $L(c_u, c_u)$ represents the number of connections between nodes in the community c_u, $L(c_u, \overline{c_u})$ represents the number of connections between nodes in the community c_u and nodes outside the community, $|c_u|$ represents the number of nodes in the community c_u.

Table 2
Specific information on real networks.

Datasets	nodes	edges	average node degree	Ref
Karate	34	78	4.5882	[41]
Dolphin	62	159	5.1290	[42]
Football	115	613	10.6609	[41]
Polbooks	105	441	8.4000	[43]
Email	1133	5451	9.6222	[44]
PGP	10,681	24,316	4.5536	[45]
CA_AstroPh	18,772	396,160	21.1038	[46]
CA_CondMat	23,133	186,936	8.0784	[46]
Email_Enron	36,692	183,831	10.0202	[46]
soc_Epinions	75,879	508,837	10.6944	[47]
Email_EuAll	265,214	420,045	2.7486	[46]
com_Youtube	1,134,890	2,987,624	5.1607	[48]

(4) Overlapping community evaluation indicators, Q_{ov}.
The expansion of modularity Q_{ov} is shown in Eq. (8).

$$Q_{ov} = \frac{1}{2m}\sum_{c=1}^{k}\sum_{i\in c, j\in c}\frac{1}{O_i O_j}\left(A_{ij} - \frac{d_i d_j}{2m}\right) \tag{8}$$

where $O_{i(j)}$ represents the number of communities to which node $i(j)$ belongs.

4.3. Non-overlapping community detection results and analysis

4.3.1. Synthetic network

The corresponding curve of community detection results of this algorithm and five non-overlapping community detection methods running 10 times on the LFR1 dataset is shown in Fig. 6.

As can be seen from Fig. 6, on the LFR1 network, when $\mu = 0 \sim 0.2$, the LPA algorithm, the ECD algorithm and RMOEA algorithm can obtain better community structure detection results. But the accuracy of the ECD algorithm decreases significantly when μ increases. The RMOEA algorithm has maintained good detection results until $\mu = 0.55$. When μ greater than 0.55, it can be seen that the CNLLP algorithm can still maintain a high *NMI* value. This shows that the proposed algorithm CNLLP can still detect the effective community structures in the networks when the network community structures blur.

The corresponding curve of community detection results of this algorithm and five non-overlapping community detection methods running 10 times on the LFR2 dataset is shown in Fig. 7.

As can be seen from Fig. 7, when the LFR network size is 5000, there is no obvious difference in the accuracy of community detection between the CNLLP algorithm and the RMOEA algorithm when the community structure is clear. But when μ greater than 0.5, the RMOEA algorithm failed to maintain a high *NMI* value. The CNLLP algorithm can maintain effective community structure detection.

The corresponding curve of community detection results of this algorithm and five non-overlapping community detection methods running 10 times on the LFR3 dataset is shown in Fig. 8.

As can be seen from Fig. 8, In the same way, when the LFR network size is 10000, there is no obvious difference in the accuracy of community detection between the CNLLP algorithm and the RMOEA algorithm when the community structure is clear. But when μ greater than 0.5, *NMI* values obtained by other algorithms except the CNLLP algorithm decreased. When μ continues to increase, that is, the community structure is gradually blurred, the CNLLP algorithm has always maintained more effective community structure detection results than other algorithms.

4.3.2. Real networks

The community detection results of the proposed algorithm and four non-overlapping community detection methods on 12 datasets are shown in Table 3. The maximum and average values of the corresponding modularity Q are given. "NA" indicates exceeding memory on computers with the same configuration or runing for more than 6 days.

It is found from Table 3 that the LPA algorithm obtains the second highest modularity value on the Karate network, but the average modularity value on this dataset is the lowest. In large datasets soc_Epinions and com_Youtube, the modularity value of the community structure obtained by the LPA algorithm is less than 0.1, which indicates that there is little difference in the density of internal and external connections in the community structure, which is inconsistent with the definition of community structure. In addition, observing the detection results of the LPA algorithm on all datasets, it can be seen that the average value and maximum value obtained by the LPA algorithm on most datasets are quite different. Thus, the LPA algorithm has strong randomness. The ECD algorithm achieves good results on small-scale datasets, in which the maximum and average values on three datasets are the highest. However,

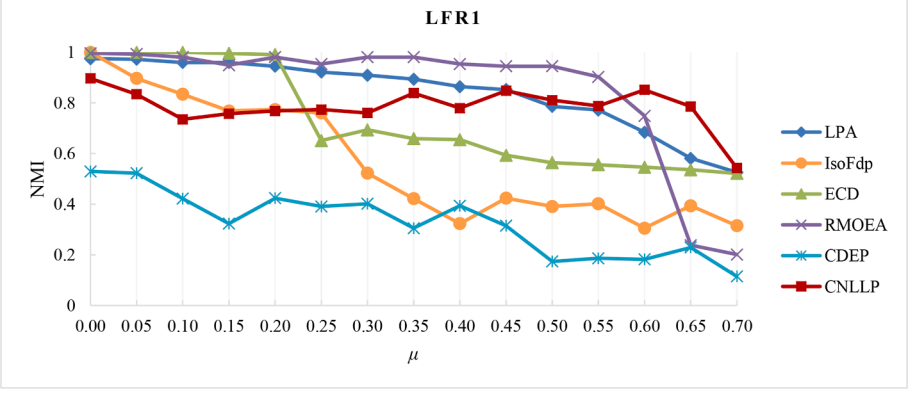

Fig. 6. Non-overlapping community detection results on the LFR1 network dataset.

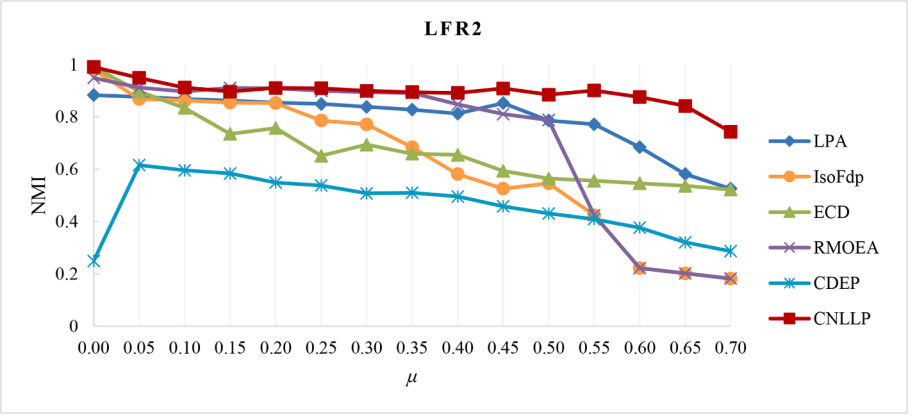

Fig. 7. Non-overlapping community detection results on the LFR2 network dataset.

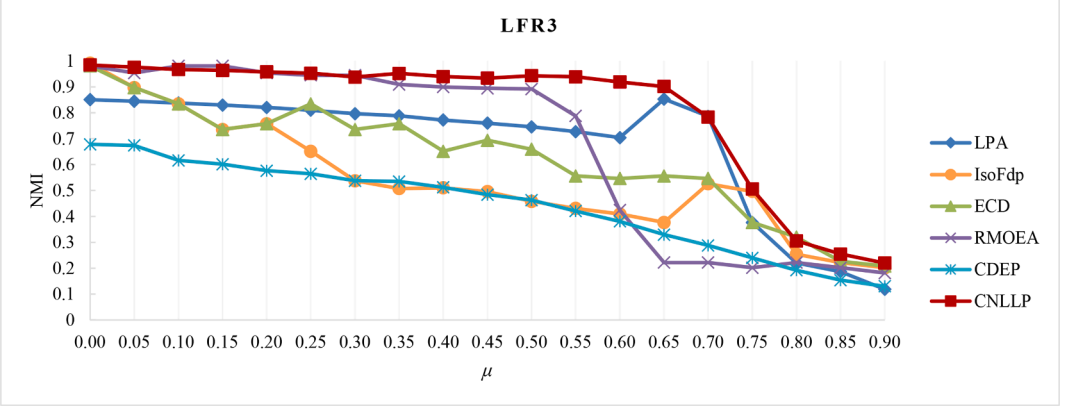

Fig. 8. Non-overlapping community detection results on the LFR3 network dataset.

Table 3
Modulatiry values of non-overlapping community detection results of real network datasets (Q).

Datasets		LPA	IsoFdp	ECD	RMOEA	CDEP	CNLLP
Karate	best	0.3991	0.3715	**0.4020**	0.3715	0.3715	0.3718
	ave	0.3366	0.3715	**0.3943**	0.3715	0.3715	0.3718
Dolphin	best	0.5034	0.4786	0.5202	**0.5246**	0.3797	**0.5246**
	ave	0.4926	0.4786	0.5070	**0.5246**	0.3797	**0.5246**
Football	best	0.5837	0.5695	0.5932	**0.6044**	0.3477	0.5811
	ave	0.5560	0.5695	0.5925	**0.6044**	0.3477	0.5667
Polbooks	best	0.4424	0.4825	0.5122	**0.5262**	0.3359	0.4569
	ave	0.4396	0.4825	0.5080	**0.5262**	0.3358	0.4569
Email	best	0.1122	0.4931	0.3400	0.4658	0.3837	**0.4938**
	ave	0.1074	0.4931	0.3350	0.4593	0.3837	**0.4938**
PGP	best	0.5902	0.7369	NA	0.7386	0.5211	**0.7400**
	ave	0.5818	0.7267	NA	0.7268	0.5211	**0.7374**
CA_AstroPh	best	0.3120	NA	NA	**0.5355**	0.5326	0.3442
	ave	0.2993	NA	NA	**0.5294**	0.5217	0.3440
CA_CondMat	best	0.5929	NA	NA	0.6190	0.4714	**0.6596**
	ave	0.5910	NA	NA	0.6102	0.4714	**0.6581**
Email_Enron	best	0.3199	NA	NA	0.4563	0.4537	**0.5374**
	ave	0.3085	NA	NA	0.4512	0.4537	**0.5365**
soc_Epinions	best	0.0151	NA	NA	NA	0.2205	**0.2943**
	ave	0.0150	NA	NA	NA	0.2205	**0.2943**
Email_EuAll	best	0.6068	NA	NA	NA	0.5463	**0.6972**
	ave	0.5760	NA	NA	NA	0.5463	**0.6940**
com_Youtube	best	0.0744	NA	NA	NA	0.2753	**0.3139**
	ave	0.0736	NA	NA	NA	0.2581	**0.3075**

due to its high algorithm complexity, when the size of network nodes increases to more than 10000, the ECD algorithm shows that the memory is exceeded with the same configuration, and it is unable to obtain effective community detection results. In the Dolphin, Football, Polbook and CA_AstroPh networks, the RMOEA algorithm achieved the highest modularity value, but it did not perform well on other large-scale datasets. When the network size increases to more than 200 thousand, the running time of the RMOEA algorithm increases significantly, and it is unable to obtain effective community detection results in a relatively short time. The CDEP algorithm has low complexity and can obtain effective community structure information on all datasets. In CA_AstroPh dataset, the CDEP algorithm obtains the highest modularity value. However, on other datasets, CDEP algorithm does not achieve the best results. In CA_CondMat and Email_EuAll datasets, the corresponding modularity values of community structure obtained are lower than that of the LPA algorithm. The CNLLP algorithm obtains the highest modularity value of non-overlapping community detection on 8 kinds of network datasets, and the improvement is obvious. On small-scale datasets, the modularity values obtained by the CNLLP algorithm are not far from other algorithms, which verifies the effectiveness of the CNLLP algorithm.

The community detection results of the proposed algorithm and four non-overlapping community detection methods on 4 datasets with true labels are shown in Table 4.

It can be seen from Table 4 that the RMOEA algorithm has the highest *NMI* value on three networks on four datasets with real community structures. The CNLLP algorithm obtains the highest *NMI* value on the Polbooks network. By comparison with Table 3, it can be found that when the *NMI* value is high, the corresponding modularity value is often low, and when the modularity value is high, the corresponding *NMI* value is low. For example, for Karate network, when the community detection result is the same as the real network partition, the *NMI* value is 1, but the modularity value at this time is 0.3715. As the number of detected communities increases, the modularity value may also increase. ECD algorithm obtains the highest Q value on the Karate network, but its *NMI* value is the lowest. This is because the *NMI* value is calculated according to the real division of the network, while the Q value is calculated according to the connection of the community structure inside and outside. The results of real community division in real networks are often of practical significance, and the connections between individuals in the network are often more complex than the connections of simple edges. However, only a few of the complex networks in real world can get the real community result information.

The maximum and average values of module density evaluation index D corresponding to the community detection results of the algorithm and five non-overlapping community detection methods running 10 times on 12 real network datasets are shown in Table 5. "NA" indicates exceeding memory on computers with the same configuration or runing for more than 6 days.

As shown in Table 5, the RMOEA algorithm has achieved the highest modularity density value on six networks. The LPA algorithm obtained the highest modularity density value on three datasets. The CNLLP algorithm only obtains the highest modularity density value on the Email dataset. From the introduction in section 4.2, we can see that both the modularity density function and the modularity function correspond to larger values when the community is more closely connected. The difference is that high modularity values often correspond to large-scale community structures, while high modularity density values often correspond to small-scale community structures. During the experiment, it was found that in PGP, CA_AstroPh, CA_CondMat, Email_Enron, soc_Epinions, Email_EuAll and com_Youtube networks, the number of communities obtained by LPA algorithm is 937, 1091, 2321, 1562, 695, 367 and 160, respectively. In PGP, CA_AstroPh, CA_CondMat and Email_Enron networks, the number of community structures detected by the RMOEA algorithm is 1045, 1746, 1985 and 2541, respectively. While the number of communities detected by the proposed algorithm CNLLP on these seven networks is 133, 3, 71, 30, 7, 156 and 87, respectively. Therefore, it can be seen that when the number of detected community structures is large, it is often easy to correspond to a large modularity density value. Therefore, the algorithm CNLLP in this paper has no sufficient advantage in the modularity density index. However, since the LPA algorithm, CDEP algorithm and CNLLP algorithm are based on the label propagation algorithm, it can be seen that the LPA algorithm ends the propagation of node labels after very few iterations, so the number of communities is very large. Therefore, its modularity density value is larger. Compared with the CDEP algorithm, the CNLLP algorithm in this paper obtains a higher modularity density value after effective label propagation, which verifies the effectiveness of the proposed algorithm CNLLP in this paper.

4.4. Results of overlapping community detection

The community detection results of the proposed algorithm and two overlapping community detection methods on 12 datasets are shown in Table 6. The maximum and average values of the Q_{ov} are given. "NA" indicates exceeding memory on computers with the same configuration or runing for more than 6 days.

Table 4
Normalized mutual information values of non-overlapping community detection of real network datasets (*NMI*).

Datasets		LPA	IsoFdp	ECD	RMOEA	CDEP	CNLLP
Karate	best	0.6995	**1**	0.5618	**1**	**1**	**1**
	ave	0.6857	**1**	0.5532	**1**	**1**	0.8372
Dolphin	best	0.0181	0.0246	0.0181	**1**	0.5996	0.6011
	ave	0.0137	0.0229	0.0154	**1**	0.5917	0.6091
Football	best	0.2131	0.2384	0.2169	**0.9366**	0.2393	0.2856
	ave	0.2086	0.2267	0.2146	**0.9366**	0.2351	0.2763
Polbooks	best	0.4256	0.3743	0.4318	0.4412	0.3460	**0.4829**
	ave	0.4039	0.3704	0.4251	0.4412	0.3447	**0.4803**

Table 5
Modulatiry density values of non-overlapping community detection results of real network datasets (D).

Datasets		LPA	IsoFdp	ECD	RMOEA	CDEP	CNLLP
Karate	best	**7.8450**	6.8333	7.6619	6.8333	6.8333	6.8235
	ave	**7.8421**	6.8333	7.6582	6.8333	6.8333	6.8235
Dolphin	best	11.380	10.092	**11.381**	9.0952	8.4017	10.997
	ave	11.273	10.002	**11.327**	9.0952	8.4015	10.586
Football	best	38.650	32.878	41.642	**44.488**	–33.537	34.292
	ave	37.275	32.878	41.278	**43.914**	–33.538	34.198
Polbooks	best	18.594	15.337	18.455	**20.626**	–7.0224	15.352
	ave	18.017	15.327	18.207	**19.003**	–7.0238	15.287
Email	best	9.6222	18.044	16.378	11.871	2.5365	**18.449**
	ave	9.3789	18.036	16.337	11.563	2.5364	**18.038**
PGP	best	575.89	30.964	NA	**1611.02**	13.775	362.89
	ave	570.74	30.278	NA	**1598.53**	13.774	361.17
CA_AstroPh	best	268.05	NA	NA	**1735.87**	10.358	156.25
	ave	265.78	NA	NA	**1763.71**	10.256	155.27
CA_CondMat	best	257.41	NA	NA	**2333.13**	19.870	223.72
	ave	255.31	NA	NA	**2329.72**	19.762	222.73
Email_Enron	best	144.33	NA	NA	**2802.64**	11.888	80.422
	ave	140.85	NA	NA	**2767.15**	11.786	79.938
soc_Epinions	best	**78.195**	NA	NA	NA	10.616	22.945
	ave	**73.766**	NA	NA	NA	10.593	22.276
Email_EuAll	best	**260.89**	NA	NA	NA	9.4612	167.09
	ave	**244.87**	NA	NA	NA	9.4278	165.37
com_Youtube	best	**210.36**	NA	NA	NA	8.3695	136.32
	ave	**205.34**	NA	NA	NA	8.0249	134.37

It can be seen from Table 6 that the DNMF algorithm and EMOFM algorithm can obtain effective overlapping community detection results when the dataset size is small. Among them, the DNMF algorithm obtains the highest extended modularity value on the Karate network and Polbooks network, and obtains the highest average extended modularity value on the Football dataset. The advantage of non-overlapping community detection results of the proposed algorithm on small-scale datasets is not very obvious, but it still obtains the highest value of extended modularity on the Dolphin network and the maximum value of extended modularity on Football network. On the Email and PGP datasets, the proposed algorithm achieves the highest value of extended modularity. When the size of the dataset continues to increase and runs on computers with the same configuration, both the DNMF algorithm and EMOFM algorithm show that they are out of memory and cannot obtain effective overlapping community detection results. The proposed algorithm can obtain effective overlapping community detection results on all datasets. Compared with Table 3, it can be found that based on the detection results of non-overlapping communities, the proposed algorithm judges the overlapping nodes according to the degree of the

Table 6
Results of overlapping community detection (Q_{ov}).

Datasets		DNMF	EMOFM	CNLLP
Karate	best	**0.3754**	0.2347	0.3717
	ave	**0.3623**	0.2347	0.3717
Dolphin	best	0.5164	0.2745	**0.5235**
	ave	0.5042	0.2733	**0.5235**
Football	best	0.5753	0.3065	**0.5809**
	ave	**0.5669**	0.3063	0.5665
Polbooks	best	**0.5028**	0.2703	0.4569
	ave	**0.4874**	0.2703	0.4569
Email	best	0.4839	0.2783	**0.4924**
	ave	0.4742	0.2748	**0.4923**
PGP	best	0.3204	0.4218	**0.7402**
	ave	0.3038	0.4200	**0.7374**
CA_AstroPh	best	NA	NA	**0.3437**
	ave	NA	NA	**0.3435**
CA_CondMat	best	NA	NA	**0.6597**
	ave	NA	NA	**0.6581**
Email_Enron	best	NA	NA	**0.5349**
	ave	NA	NA	**0.5340**
soc_Epinions	best	NA	NA	**0.2886**
	ave	NA	NA	**0.2837**
Email_EuAll	best	NA	NA	**0.6852**
	ave	NA	NA	**0.6731**
com_Youtube	best	NA	NA	**0.3952**
	ave	NA	NA	**0.3941**

community to which the nodes belong, so the detection results of overlapping communities obtained are guaranteed by the preliminary work. In the subsequent experiments, the effective running time of all algorithms to detect non-overlapping communities and overlapping communities will be given respectively to further verify the effectiveness of each algorithm.

4.5. Statistical experiment

To better prove the stability of the proposed algorithm, the results obtained by comparing the CNLLP algorithm with five non-overlapping community detection algorithms on 12 real network datasets are statistically analyzed. Because of the significant difference between the evaluation indicators *NMI* and *D*, this section only analyzes the *Q* value. The specific method is Wilcoxon Signed Rank Test. The specific results are shown in Table 7.

Here p represents the probability that the median values of the two samples are equal. When p is close to 0, the zero hypotheses should be questioned. h is the test result, and $h = 0$ indicates that the median difference between the two samples is not significant. While $h = 1$ indicates that the difference between the median of the two samples is significant. According to Tables 3 and 7, the difference between the CNLLP algorithm and RMOEA algorithm in Dolphin and PGP networks is not obvious. The difference between the CNLLP algorithm and LPA algorithm in Karate and Football networks is not obvious. For other comparison algorithms and datasets, the proposed algorithm CNLLP has obvious advantages when the *Q* value is better. The overall results of the modularity value are significantly improved.

4.6. Algorithm analysis

After obtaining the community detection results based on the core node and the layer-by-layer label propagation algorithm, the proposed algorithm calibrates the current community detection results according to the attraction of the community to the nodes. The experimental results on 12 real networks are shown in Fig. 9.

It can be seen from Fig. 9 that the calibration method introduced in this paper has significantly improved the corresponding modularity value of the community structure on 12 real networks, except for the two datasets of the Karate network and PGP network. Where the modularity value corresponding to the CA_AstroPh dataset increased by 23.9%, the CA_CondMat dataset increased by 15.2%, the Email_Enron dataset increased by 42.9%, and Email_EuAll increased by 75.5% on the full dataset. Especially on the soc_Epinions dataset, the improvement effect is very considerable.

4.7. Running time

The proposed algorithm CNLLP and 5 non-overlapping community detection methods are run 10 times on 12 datasets. The average time of algorithm operation is analyzed. To more intuitively reflect the results, the running time of 5 small-scale datasets is shown in Table 8.

The running time of 7 large-scale datasets is shown in Fig. 10. For the algorithms that cannot obtain effective community structure detection results in a relatively limited time, the longest time (6 days) is used in Fig. 10 to represent their running time.

It can be seen from Table 8 and Fig. 10, the proposed algorithm can detect the effective community structure information fastest on the 3 datasets of Karate network, Dolphin network and Polbooks network. ECD algorithm is obviously the slowest, and its running time on small-scale datasets is thousands of times that of the other four non-overlapping community detection algorithms. When the dataset size increases to more than 1000 nodes, such as Email dataset, the running time of ECD algorithm increases sharply. It can be expected that when the dataset size reaches one million, ECD algorithm will not be able to obtain effective community detection results in a limited period of time. The running time of IsoFdp algorithm on small-scale datasets is not much different from that of LPA algorithm, CDEP algorithm and CNLLP algorithm. It can also quickly obtain effective community detection results on Email datasets. However, when the dataset size increases to more than 10,000 nodes, the running time of IsoFdp algorithm also increases sharply. It can also be expected that when the dataset size reaches one million, IsoFdp algorithm will not be able to obtain effective community detection

Table 7
The statistical results of the *Q* value on 12 real networks.

Datasets	LPA/CNLLP		IsoFdp/CNLLP		ECD/CNLLP		RMOEA/CNLLP		CDEP/CNLLP	
	p	*h*	*p*	*h*	*p*	*h*	*p*	*h*	*p*	*h*
Karate	1.000000	0	0.000016	1	0.000047	1	0.000016	1	0.000016	1
Dolphin	0.000047	1	0.000016	1	0.000047	1	0.076716	0	0.000016	1
Football	1.000000	0	0.013336	1	0.000085	1	0.000033	1	0.000033	1
Polbooks	0.000047	1	0.000016	1	0.000047	1	0.000016	1	0.000047	1
Email	0.000047	1	0.000016	1	0.000047	1	0.000047	1	0.000016	1
PGP	0.000114	1	0.000114	1	NA	NA	0.056085	0	0.000047	1
CA_AstroPh	0.000114	1	NA	NA	NA	NA	0.000114	1	0.000114	1
CA_CondMat	0.000114	1	NA	NA	NA	NA	0.000114	1	0.000047	1
Email_Enron	0.000114	1	NA	NA	NA	NA	0.000114	1	0.000047	1
soc_Epinions	0.000047	1	NA	NA	NA	NA	NA	NA	0.000016	1
Email_EuAll	0.000114	1	NA	NA	NA	NA	NA	NA	0.000047	1
com_Youtube	0.000114	1	NA	NA	NA	NA	NA	NA	0.000114	1

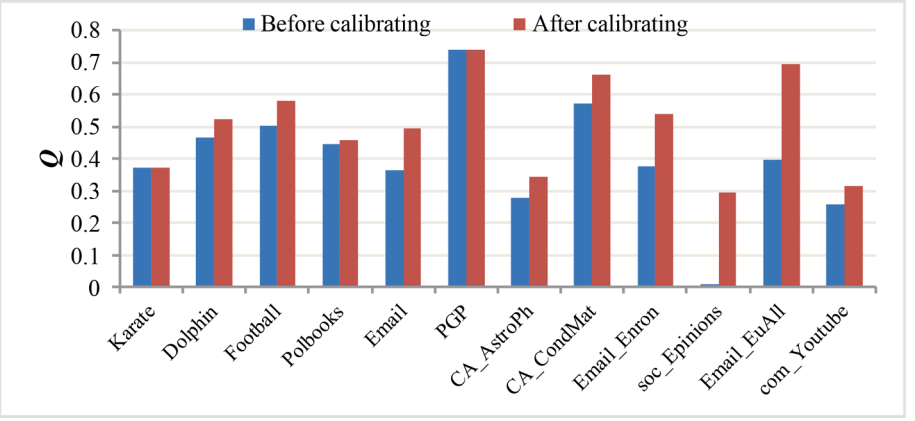

Fig. 9. Experimental results of node label calibration effectiveness.

Table 8
Running time of non-overlapping community detection on 5 small-scale datasets.

Time (s)	LPA	IsoFdp	ECD	RMOEA	CDEP	CNLLP
Karate	0.0328	0.0370	15.2883	26.686374	0.0293	**0.0168**
Dolphin	0.0762	0.0516	32.5041	30.41287	0.0404	**0.0301**
Football	**0.0631**	0.1177	67.6755	37.89031	0.0851	0.0776
Polbooks	0.0789	0.1143	25.6074	36.40158	0.0568	**0.0397**
Email	31.6803	6.1803	4385.8941	218.161	**0.2931**	0.7773

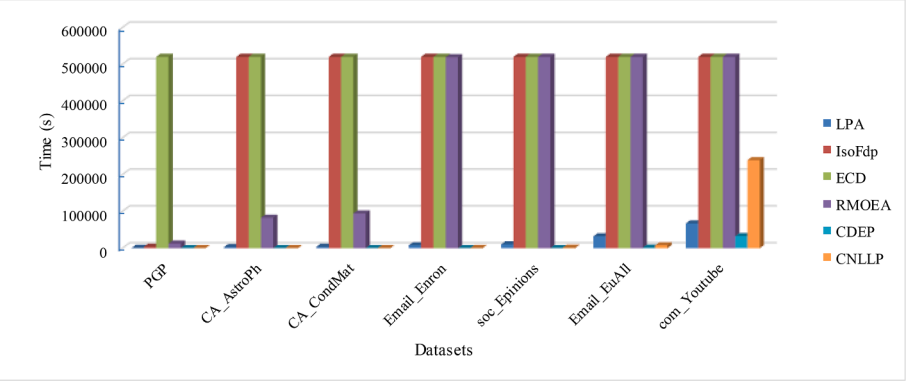

Fig. 10. Running time of 6 non-overlapping community detection methods on 7 datasets.

results in a limited period of time. LPA algorithm has relatively low complexity and relatively fast running speed. It can obtain community detection results in a limited time. CDEP algorithm runs faster than the CNLLP algorithm on multiple datasets, but compared with Table 3, although CNLLP algorithm sacrifices limited algorithm running time, it has achieved a significant improvement in the final community detection results.

The CNLLP algorithm and 2 overlapping community detection methods are run 10 times on 12 datasets. The average time of algorithm operation is analyzed. In order to more intuitively reflect the results, the running time of 5 small-scale datasets is shown in Table 9. It can be seen from Table 9, the CNLLP algorithm in this paper runs faster than DNMF algorithm and EMOFM algorithm. The running time difference between DNMF algorithm and the CNLLP algorithm on small-scale datasets is not very obvious. When the dataset size reaches more than 1000 nodes, the running time of DNMF algorithm reaches hundreds of times that of the CNLLP algorithm.

The running time of 7 large-scale datasets is shown in Fig. 11. For the algorithms that cannot obtain effective community structure detection results in a relatively limited time, the longest time (6 days) is used in Fig. 11 to represent their running time.

It can be seen from Fig. 11, when the dataset size reaches more than 10,000 nodes, the operation time of DNMF algorithm is about 730 times that of the CNLLP algorithm. It can be expected that when the dataset size reaches one million, DNMF algorithm will not be able to obtain effective overlapping community detection results in a limited period of time. The running time of EMOFM algorithm on

Table 9
Running time of overlapping community detection on 5 small-scale datasets.

Time (s)	DNMF	EMOFM	CNLLP
Karate	0.0805	6.377	**0.0221**
Dolphin	0.5164	10.6310	**0.0407**
Football	0.4168	16.4462	**0.0984**
Polbooks	0.4369	12.1523	**0.0567**
Email	281.1487	1552.2730	**1.20471**

small-scale datasets has shown its high complexity. When the dataset size reaches more than 1000 nodes, its running time is five times that of DNMF algorithm. Therefore, it can be expected that when the dataset size reaches one million, EMOFM algorithm will also not be able to obtain effective overlapping community detection results in a limited period of time. To sum up, the CNLLP algorithm runs faster and can obtain effective overlapping community detection results in a limited time.

5. Conclusion

For large-scale complex network community detection tasks, a large-scale community detection algorithm based on core nodes and layer-by-layer label propagation is proposed in this paper. First, the core nodes whose node degree is greater than the average node degree in the graph are found, which can effectively use the feature of core nodes as potential community centers, and avoid the impact of nodes with low node degree on community structure detection. Then, starting from the core node, label propagation is carried out layer-by-layer according to the node degree and node connection, which effectively improves the accuracy of community detection. The node labels are calibrated after label propagation according to the current attraction of the community to the nodes, which effectively improves the misclassification in early community structure detection. Finally, based on the non-overlapping community structure, overlapping community detection is carried out to make the overlapping community detection results more accurate and interpretable. Under different evaluation indicators, the detection results of the non-overlapping and overlapping community structures of the synthetic network datasets and the real network datasets show that the proposed algorithm in this paper can improve the accuracy of the label propagation algorithm, while largely preserving its linear complexity. This not only provides ideas for large-scale community detection, but also helps to expand the application of community detection to large-scale complex scenes. However, there is still room for further improvement of the proposed algorithm. From the detection results and analysis of modularity and modularity density, it can be found that the detection results of different algorithms for datasets without real labels may have great differences. The proposed algorithm does not achieve the optimal modularity density value when it obtains a good modularity value. Therefore, relying on only evaluation indicators may not be sufficient to fully explain the advantages and disadvantages of the algorithms. Especially in most realistic scenarios, data with real labels will be even less common. The next research work is to continue to explore the application scenarios of the large-scale community detection algorithm in this paper, and give play to its advantages in solving large-scale complex scene problems.

CRediT authorship contribution statement

Weitong Zhang: Methodology, Investigation. **Ronghua Shang:** Supervision, Project administration. **Licheng Jiao:** Supervision, Project administration, Funding acquisition.

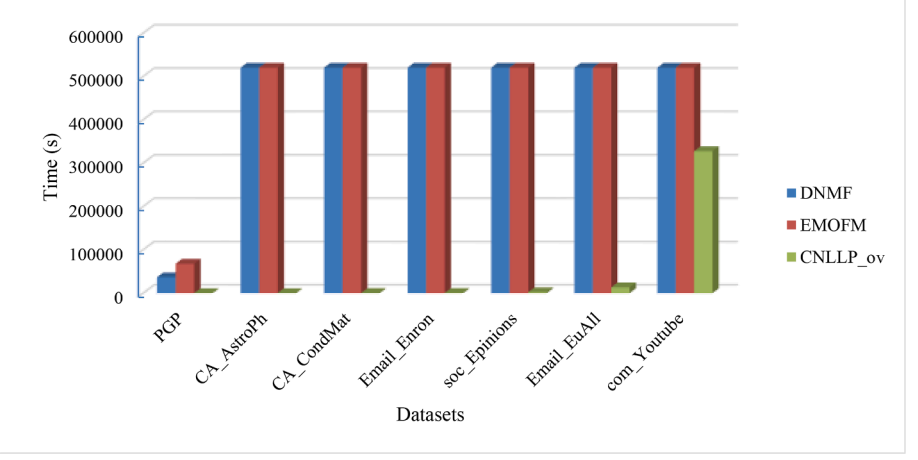

Fig. 11. Running time of 3 overlapping community detection methods on 7 datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was partially supported by the Natural Science Basic Research Program of Shaanxi under Grant 2022JQ-616 and Nos.2022JC-45, the Fundamental Research Funds for the Central Universities under Grant XJS221903, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2021A1515110686, the National Natural Science Foundation of China under Grants Nos. 62176200 and 62271374, the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB03 and the Research Project of SongShan Laboratory under Grant YYJC052022004.

References

[1] S. Fortunato, Community detection in graphs, Physics Reports. 486 (3-5) (2010) 75–174.
[2] R. Shang, W. Zhang, L. Jiao, X. Zhang, R. Stolkin, Dynamic Immunization Node Model for Complex Networks Based on Community Structure and Threshold, IEEE Trans. Cybern. 52 (3) (2022) 1539–1552.
[3] Y. Chen, D. Mo, Community detection for multilayer weighted networks, Information Sciences. 595 (2022) 119–141.
[4] A. Nocaj, M. Ortmann, U. Brandes, Adaptive Disentanglement Based on Local Clustering in Small-World Network Visualization, IEEE Trans. Visual. Comput. Graphics. 22 (6) (2016) 1662–1671.
[5] H. Peng, S. Si, M.K. Awad, N. Zhang, H. Zhao, X.S. Shen, Toward Energy-Efficient and Robust Large-Scale WSNs: A Scale-Free Network Approach, IEEE J. Select. Areas Commun. 34 (12) (2016) 4035–4047.
[6] C. He, Y. Zheng, J. Cheng, Y. Tang, G. Chen, H. Liu, Semi-supervised overlapping community detection in attributed graph with graph convolutional autoencoder, Information Sciences. 608 (2022) 1464–1479.
[7] R. Shang, W. Zhang, J. Zhang, L. Jiao, Y. Li, R. Stolkin, Local Community Detection Algorithm Based on Alternating Strategy of Strong Fusion and Weak Fusion, IEEE Trans. Cybern. (2022) 1–14.
[8] R. Shang, K. Zhao, W. Zhang, J. Feng, Y. Li, L. Jiao, Evolutionary multiobjective overlapping community detection based on similarity matrix and node correction, Applied Soft Computing. 127 (2022), 109397.
[9] M. Chen, K. Kuzmin, B.K. Szymanski, Community Detection via Maximization of Modularity and Its Variants, IEEE Trans. Comput. Soc. Syst. 1 (1) (2014) 46–65.
[10] R. Guimerà, L.A. Nunes Amaral, Functional cartography of complex metabolic networks, Nature. 433 (7028) (2005) 895–900.
[11] Q. Cai, L. Ma, M. Gong, D. Tian, A survey on network community detection based on evolutionary computation, IJBIC. 8 (2016) 84.
[12] C. Pizzuti, Boosting the detection of modular community structure with genetic algorithms and local search, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, Trento Italy, (2012) 226-231.
[13] J. Ji, X. Song, C. Liu, X. Zhang, Ant colony clustering with fitness perception and pheromone diffusion for community detection in complex networks, Physica A: Statist. Mechan. Appl.. 392 (15) (2013) 3260–3272.
[14] M. Nasser Al-Andoli, S. Chiang Tan, W. Ping Cheah, Distributed parallel deep learning with a hybrid backpropagation-particle swarm optimization for community detection in large complex networks, Information Sciences. 600 (2022) 94–117.
[15] F. Liu, J. Wu, C. Zhou, J. Yang, Evolutionary Community Detection in Dynamic Social Networks, in: in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, Budapest, Hungary, 2019, pp. 1–7.
[16] Y. Tian, S. Yang, X. Zhang, An Evolutionary Multiobjective Optimization Based Fuzzy Method for Overlapping Community Detection, IEEE Trans. Fuzzy Syst. 28 (11) (2019) 2841–2855.
[17] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. U.S.A. 104 (1) (2007) 36–41.
[18] R. Shang, S. Luo, W. Zhang, R. Stolkin, L. Jiao, A multiobjective evolutionary algorithm to find community structures based on affinity propagation, Phys. A: Statist. Mechan. Appl.. 453 (2016) 203–227.
[19] R. Shang, W. Zhang, L. Jiao, R. Stolkin, Y. Xue, A community integration strategy based on an improved modularity density increment for large-scale networks, Phys. A: Statist. Mechan. Appl. 469 (2017) 471–485.
[20] F. Ye, S. Li, Z. Lin, C. Chen, Z. Zheng, Adaptive Affinity Learning for Accurate Community Detection, in: in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, Singapore, 2018, pp. 1374–1379.
[21] X. Kang, L. Zhu, A. Ming, Dynamic Random Walk for Superpixel Segmentation, IEEE Trans. Image Process. 29 (2020) 3871–3884.
[22] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York New York USA, (2014) 701-710.
[23] A. Grover, J. Leskovec, node2vec: Scalable Feature Learning for Networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, (2016) 855-864.
[24] W. Zhang, R. Zhang, R. Shang, L. Jiao, Weighted compactness function based label propagation algorithm for community detection, Phys. A: Statist. Mechan. Appl. 492 (2018) 767–780.
[25] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E. 76 (2007), 036106.
[26] H. Yu, R.u. Ma, J. Chao, F. Zhang, An Overlapping Community Detection Approach Based on Deepwalk and Improved Label Propagation, IEEE Trans. Comput. Soc. Syst. 10 (1) (2023) 311–321.
[27] M. Lu, Z. Zhang, Z. Qu, Y. Kang, LPANNI: Overlapping Community Detection Using Label Propagation in Large-Scale Complex Networks, IEEE Trans. Knowl. Data Eng. 31 (2019) 1736–1749.
[28] X. Wu, C. Zhang, Detecting Network Community by Propagating Labels Based on Contact-Specific Constraint, in: J.J. Park, H. Jin, Y.-.-S. Jeong, M.K. Khan (Eds.), Advanced Multimedia and Ubiquitous Engineering, Springer Singapore, Singapore, 2016, pp. 697–705.
[29] P. Schuetz, A. Caflisch, Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement, Phys. Rev. E. 77 (2008), 046112.
[30] X. Liu, T. Murata, Advanced modularity-specialized label propagation algorithm for detecting communities in networks, Phys. A: Statist. Mechan. Appl. 389 (7) (2010) 1493–1500.
[31] S. Yazdanparast, M. Jamalabdollahi, T.C. Havens, Linear Time Community Detection by a Novel Modularity Gain Acceleration in Label Propagation, IEEE Trans. Big Data. 7 (6) (2021) 961–966.
[32] Z. Lin, X. Zheng, N. Xin, D. Chen, CK-LPA: Efficient community detection algorithm based on label propagation with community kernel, Phys. A: Statist. Mechan. Appl. 416 (2014) 386–399.

[33] S. Zhang, S. Yu, X. E, R.u. Huo, Z. Sui, Label Propagation Algorithm Joint Multilayer Neighborhood Overlap and Historic Label Similarity for Community Detection, IEEE Syst. J. 16 (2) (2022) 2626–2634.
[34] X. Zhao, J. Liang, J. Wang, A community detection algorithm based on graph compression for large-scale social networks, Information Sciences. 551 (2021) 358–372.
[35] Y. Chen, A.Q. Hu, J. Hu, et al., A method for finding the most vital node in communication networks, High Technol. Lett. 1 (2) (2004) 573–575.
[36] H. Roghani, A. Bouyer, A Fast Local Balanced Label Diffusion Algorithm for Community Detection in Social Networks, IEEE Trans. Knowl. Data Eng. (2022) Early access, DOI: 10.1109/TKDE.2022.3162161.
[37] J. Chen, H. Zheng, X. Lin, Y. Wu, M. Su, A novel image segmentation method based on fast density clustering algorithm, Eng. Appl. Artific. Intellig. 73 (2018) 92–110.
[38] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, Y. Jin, A Network Reduction-Based Multiobjective Evolutionary Algorithm for Community Detection in Large-Scale Complex Networks, IEEE Trans. Cybern. 50 (2) (2020) 703–716.
[39] F. Ye, C. Chen, Z. Zheng, R.-H. Li, J.X. Yu, Discrete Overlapping Community Detection with Pseudo Supervision, in: in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE, Beijing, China, 2019, pp. 708–717.
[40] A. Lancichinetti, S. Fortunato, Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities, Phys. Rev. E 80 (1) (2009), 016118.
[41] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Nat. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[42] D. Lusseau, K. Schneider, O.J. Boisseau, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 392–405.
[43] V. Krebs, Books about us politics, http://www.orgnet.com/, 2004.
[44] R. Guimer, L. Danon, A. Daz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (6) (2003), 065103.
[45] B. Marin, P.S. Romualdo, D.G. Albert, A. Alex, Models of social networks based on social distance attachment, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 70 (2) (2004), 056122.
[46] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: densification and shrinking diameters, ACM Trans. Knowl. Discovery Data 1 (1) (2006) 2.
[47] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, Lect. Notes Comput. Sci. 284 (10) (2003) 351–368.
[48] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, in: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, (2012) 1–8.
[49] G.E. Hinton, N. Srivastava, A. Krizhevsky, et al., Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, Comput. Sci. 3 (4) (2012) 212–223.