Overlapping community detection algorithm based on similarity of node relationship

Hongtao Liu
Chongqing University of Posts and Telecommunications
Ning Wang (s180231003@stu.cqupt.edu.cn)
Chongqing University of Posts and Telecommunications https://orcid.org/0000-0003-1597-6471

Research Article

Keywords: Community detection, Local expansion, Relationship similarity, Overlapping Community

Posted Date: August 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-486226/v1

Version of Record: A version of this preprint was published at Soft Computing on May 6th, 2023. See the published version at https://doi.org/10.1007/s00500-023-08067-2.

Abstract

Community discovery is a vital link in the research of social networks. Aiming at the shortcomings of the current local extension-based community discovery algorithm in local community discovery and extension, we propose a based on relationship similarity and local extension Overlapping community detection algorithm(RSLO). First, use the node's relationship similarity strategy to find close seed communities. Then, according to the discovered seed community, the similarity between the neighbor nodes of the community and the community is calculated, and the nodes whose similarity meets the threshold are selected. After that, an adaptive optimization function is used to expand the community. Finally, the free nodes that have not been divided into the community are divided into communities, thereby achieving a more comprehensive community discovery. We conduct experiments on classic datasets and artificially generated networks. The results show that the RSLO algorithm can find accurate and objective community structures.

1. Introduction

Since humans have become the main body in social networks, their social relationships have also been projected onto the network. This kind of network structure containing social relationships is called a social network. Due to the complex relationships between people, it can be called a complex network. [1] We can summarize some people with high similarity in social networks into a community. The so-called community is composed of a group of closely connected nodes in the network. These nodes are relatively closely connected to each other, but are rarely connected to other closely connected nodes. Most intuitively, the process of community discovery is a clustering process. Community discovery aims to discover the most reasonable and true community structure in social networks, and it has become one of the important tasks to explore the network structure [2]. In order to discover excellent community structure, many scholars and researchers have proposed many community discovery algorithms and many algorithms have also been developed to a certain extent. At present, the research results discovered by the community can be applied to many fields, such as: personalized recommendation, network public opinion analysis, disease infection network, etc. [3].

Because the membership of some nodes in the community structure of social networks is not single, the classic discovery algorithm for non-overlapping communities is no longer suitable for overlapping community structures. Therefore, the research on the discovery of overlapping community structures has gradually become a hot spot in this field. After years of development, a large number of relevant research results have appeared in this field. These methods can be roughly divided into: faction filtering algorithms, local expansion algorithms, edge division discovery Algorithms and label propagation algorithms and other 4 categories [4]. The following is a brief introduction to the classic algorithms in these categories and the research in recent years:

The faction filtering algorithm CPM was first proposed by Palla et al. [5]. They believe that a community is composed of some connected complete subgraphs. The operation of the algorithm is to form a fully

connected subgraph composed of k nodes (that is, k Factions), search for neighboring factions consisting of k-1 identical nodes from the network to find overlapping community structures. Later, Palla et al. [6] proposed a CPMd algorithm that can handle directed graphs. The algorithm uses directed k factions instead of k factions in the CPM algorithm, completing the overlapping community discovery on the directed graph.. Lu et al. [7] proposed an overlapping community discovery algorithm based on the expansion of greedy factions. First, search for the largest faction in the network, calculate the link strength according to the degree of association between the factions, and convert the original network graph into the largest faction graph. Under the condition of maximizing the fitness function, greedily expand the seed factions in the largest faction graph for community discovery. Zhang et al. [8] merged the searched factions in the network according to the coupling strength, so as to hierarchically divide the obtained tree diagram to obtain the overlapping community structure. This type of method uses factions as the medium to explore the structure of overlapping communities, but its results are not ideal when dealing with relatively sparse network structures, and the time complexity of the algorithm is relatively high.

The basic idea of the local expansion algorithm is that in the network, the seed nodes are found according to the relevant strategies formulated, and then the community expansion is carried out according to the found seed nodes through the local optimization function to obtain the optimal community division [9]. For example, by Lancchinetti The proposed LFM algorithm [10] found the community based on the fitness function of the node, and then selected nodes outside the community as seed nodes for community expansion. Su et al. [11] use random walk strategy for community expansion. For such methods, the most important thing is the selection of seed nodes. For this reason, Wang et al. [12] proposed the concept of a structural central node, and used it as a seed node for local community expansion. Sun et al. [13] proposed the vertex cover growth rate to select the seed node, and combined the random walk strategy to expand the community to discover overlapping communities. Li et al. [14] proposed an overlapping community discovery algorithm based on the greedy expansion of seed nodes, which uses a greedy strategy based on the fitness function to expand the seed set according to the seed nodes. The algorithm can find high-quality overlapping community structures.

Based on the method of label propagation, each network node is assigned a label containing overlapping membership relationships, and through the propagation of these labels between neighboring nodes, the node's membership relationship with each community finally reaches a stable state, thereby obtaining community discovery results. Overlapping community discovery algorithm LPA [15] research. Typical representatives of this type of method in the field of overlapping community discovery are the multi-label COPRA [16] algorithm and the Speaker-listener model-based SLPA algorithm [17]. Lu et al. LPANNI [18] Algorithm

It is not difficult to find that the above algorithm is mainly for the study of nodes and their attributes in the network, in order to find overlapping community structures. In contrast, the edge division discovery algorithm starts from the perspective of edges and discovers the community structure. Wang [19] et al. proposed a label propagation algorithm based on edge propagation probability for the traditional

overlapping label propagation algorithm COPRA; GUO [20] proposed an overlapping community discovery algorithm based on edge density clustering. First, take edges as The research object uses density clustering to detect closely connected core edge communities. Then, according to the boundary edge attribution strategy, the boundary edge is divided into the core edge community closest to it. For isolated edges, a community attribution based on edge degree and edge is proposed Isolated edge processing strategy, Wang [21] proposed an adaptive overlapping community discovery algorithm based on the mixed parameters of edge trust, which defines the set of neighbors on the network side and the trust function between its neighbors, through information transmission Obtain the total amount of information of the edge to realize adaptive discovery of overlapping communities. At present, edge partition discovery algorithms have become an important class of overlapping community discovery algorithms.

However, many existing community discovery algorithms are based on the topology of the network, that is, the local information between nodes, while ignoring the influence of the connection relationship between nodes. Real social networks are based on the relationship between people living in reality, and independent people lead to individual differences. A real community should fully consider the importance of connections between nodes. Only in this way can we study the behavior patterns of the entire community through individuals. Many algorithms only perform cluster analysis for the purpose of simply discovering communities, thus ignoring the influence of node connection relationships. In order to overcome this problem, we will propose the relationship similarity of nodes to evaluate the value of nodes to the community. For all nodes in the community, there is a certain degree of similarity between nodes in the same community. We express this value through the local clustering coefficient of the node; and for a node in the community, each node Nodes directly connected to themselves have a higher similarity, and we express this value through the similarity of the relationship between the nodes. On this basis, the two are combined, and an overlapping community discovery algorithm based on relationship similarity and local expansion is proposed. The algorithm sets the node with the highest number as the core node, and then judges whether the node in the community can belong to multiple communities at the same time according to the tendency of the nodes in the community. If it is, mark it as an overlapping node and divide it into the community. The node is removed from the network. Iteratively obtain the division result of overlapping communities. Experimental results show that the algorithm has better performance than several classic overlapping community partitioning algorithms.

2. Related Work

For an undirected and unweighted network $G = (V, E)$, V is a set composed of all nodes in the network, and E is a set composed of all edges in the network. Specifically: $V(G) = \{v_1, v_2, \cdots, v_n\}$, $E(G) = \{e_{ij}\}\, e_{ij} = (v_i, v_j), v_i, v_j \in V$. The number of nodes is $|V| = N$, the number of edges is $|E| = m$, $\Gamma(v_i)$ is a set of neighbor nodes of node v_i, and k_i represents the degree of node v_i.

2.1 Local clustering coefficient

The clustering coefficient [22] is a parameter used to measure the degree of clustering between nodes in the network. In a real social network, the clustering parameter represents the close relationship between friends. Specifically, it measures how close a node is to its neighbors. The local clustering coefficient of an undirected graph can be defined as:

$$LCC_i = \frac{2e_i}{k_i (k_i - 1)}$$

Among them, k_i is the degree of node v_i, and e_i is the number of nodes connected between neighbors of node v_i. LCC_i is the local clustering coefficient of node v_i, and its value is between [0,1]. Under certain circumstances, $LCC_i = 0$ means that the neighbor nodes of node v_i have no relationship with each other, and $LCC_i = 1$ represents all neighbor nodes of v_i are connected to each other.

2.2 Node relationship similarity

In social networks, the similarity of nodes reflects the similarity between nodes. The intimacy between nodes is reflected by the similarity, which can better reflect the relationship between nodes. Lü [23] and others summarized the currently popular similarity indexes, which are shown in Table 1, which include the Salton index, Jaccard index, Sorensen index, Resource Allocation (RA) index, etc.

Table 1
Similarity Index

Similarity Name	Formula
Salton Index	$\frac{\lvert\Gamma(i)\cap\Gamma(j)\rvert}{\sqrt{k_i \bullet k_j}}$
Jaccard Index	$\frac{\lvert\Gamma(i)\cap\Gamma(j)\rvert}{\lvert\Gamma(i)\cup\Gamma(j)\rvert}$
Sorensen Index	$\frac{2\lvert\Gamma(i)\cap\Gamma(j)\rvert}{k_i+k_j}$
Resource Allocation Index	$\sum_{t\in\Gamma(i)\cap\Gamma(j)} \frac{1}{k_t}$

In most real networks, nodes tend to be relatively closely connected groups, which are characterized by high local connection density. The higher the clustering coefficient of a node, the stronger the cohesion of its neighbor nodes. Professionals have done a lot of research on the exploration and optimization of cluster analysis. In many studies on clustering coefficients, the focus is on the closeness between two nodes and their adjacent nodes, while ignoring the connection between the two nodes themselves, which will reduce the accuracy of similarity. For example: In a real social relationship, two people who become friends will have a friend relationship, which will increase the similarity between them. Moreover, the two

displays are friends but do not have a mutual friend. Existing, obviously according to the individual clustering coefficient, its similarity is 0. Therefore, in the process of studying the similarity, we not only consider the node and its neighbor nodes, but also consider the connection relationship between the nodes themselves. Through the research of these works, in the external algorithm, we use the similarity of the node relationship based on the local clustering coefficient. The relationship similarity of nodes is as follows:

$$\mathrm{RSim}_{i,j} = \sum_{t \in \Gamma(i) \cap \Gamma(j)} \mathrm{LCC}_t$$

Among them, LCC_t is the local clustering coefficient of ν_t, and t is the set of common neighbor nodes of nodes ν_i and ν_j. Then the relationship importance of nodes can be expressed as:

$$\mathrm{RI}(i) = \sum_{t \in \Gamma(i) \cap \Gamma(j)} \mathrm{RSim}_{i,j} \times \sum_{z} \frac{1}{k(z)}$$

Among them, There are two situations for the parameter z: when the nodes ν_i and ν_j are directly connected, $z \in t \cup \{v_i, v_j\}$; when the nodes ν_i and ν_j are not directly connected, $z = t$. The similarity not only measures the aggregation coefficient of the common neighbors of two nodes, but also considers the connection relationship between the two nodes themselves. The closer the relationship between the common neighbor nodes, the closer the relationship between the two nodes. The aggregation coefficient of common neighbors is an important indicator for calculating two nodes. The larger the index, the higher the similarity of nodes. The relationship between nodes also affects the similarity of two nodes to a certain extent. Therefore, the similarity between the two and the node is positively correlated.

2.3 Local expansion related concepts

In the community discovery algorithm based on local expansion, in addition to precise similarity, several basic concept parameters are still required, which are defined as follows:

2.3.1 Community neighbor set

The community neighbor set $N_v(C)$ represents the set of nodes connected to the community C:

$$N_v(C) = \forall_{v \in c} \Gamma(v) - C$$

Among them, C represents a community, and $\Gamma(v)$ represents the neighbor nodes directly connected to node v.

2.3.2 Similarity between node and community

The similarity $S_{vc}(v_i, C)$ between node v_i and community C is defined as:

$$S_{vc}(v_i, C) = \frac{|N_v(C) \cap \Gamma(v_i)|}{|N_v(C) \cup \Gamma(v_i)|}$$

At this time, the node v_i not belongs to the community C. The larger the value of $S_{vc}(v_i, C)$, the greater the probability that the node v_i belongs to the community C.

2.3.3 Community similarity

The similarity $S_{cc}(C_i, C_j)$ between community C_i and community C_j is:

$$S_{cc}(C_i, C_j) = \frac{|C_i \cap C_j|}{\min(|C_i|, |C_j|)}$$

Among them,$|C_i \cap C_j|$ represents the number of the same nodes in the community C_i and the community C_j. $\min(|C_i|, |C_j|)$ represents the minimum of the number of nodes in the community C_i and the number of nodes in the community C_j. The larger $S_{cc}(C_i, C_j)$, the more similar the structure of the community C_i and the community C_j. Usually, when $S_{cc}(C_i, C_j) > 0.5$, the two communities can be merged.

2.3.4 Adaptive function

The adaptive fitness function is used to measure the tightness of a group of nodes, and its formula is defined as follows:

$$Q_f(C) = \frac{Com_{in}^g}{(Com_{in}^g + Com_{out}^g)^\alpha}$$

Among them, Com_{in}^g and Com_{out}^g are respectively the sum of the internal degree and the external degree of community C. In addition, $Q_f^+(C, v) = Q_f(C \cup v)$means, The value of Q_f after adding node v to the current community; $Q_f^-(C, v) = Q_f(C - v)$ means the value of Q_f after node v is removed in the current community. The parameter α is a positive real number. To control the size of the discovery community.

3. Rslo Algorithm

In this part, we will briefly sort out the algorithm flow and the pseudo code of RSLO, see Algorithm 1.

Algorithm 1. RSLO

Input: Network $G = (V, E)$, α, γ, δ

Output: Overlapping Community Result $O = \{C_1, \cdots C_k\}$

Begin:

1. $O = \Phi, Seed_c, List_{RI}, List_{core}$

2. For $node$ in $V(G)$: $List_{RI}(node) = RI(node)$ END For

3. For $node_i$ in $V(G)$:

4. $n_{cc} = 0, n_{sum} = |\Gamma(node_i)|$

5. For $node_j$ in $\Gamma(node_i)$:

6. If $RI(node_i) > RI(node_j)$: n_{cc}+=1 End If;End For

7. If $(n_{cc}/n_{sum}) > \delta$: $List_{core} = List_{core} \cup node_i$ End If; End For

8. For $node_i$ in $List_{core}$:

9. If $node$ not in $Seed_c$: $C_s = \Phi$, $C_s = C_s \cup node_i$

10. For $node_j$ in $\Gamma(node_i)$: $sim = S_{vc}(node_i, C_s)$

11. If $sim > \gamma$: $C_s = C_s \cup node_j$ End If; End For

12. $Seed_c = Seed_c \cup C_s$;End If; End For

13. $Seed_m = \Phi$

14. For $node$ in $Seed_c$:

Algorithm 1. RSLO

15. If $\mathrm{Seed_m} = \Phi$:$\mathrm{Seed_m} = \mathrm{Seed_m} \cup \mathrm{node}$ End If

16. $\mathrm{merge} = \mathrm{True}$

17. For $\mathrm{node_m}$ in $\mathrm{Seed_m}$:$\mathrm{sim} = \mathrm{S_{vc}}\,(\mathrm{node}, \mathrm{node_m})$

18. IF $\mathrm{sim} > \gamma$:$\mathrm{s_{merge}} = \mathrm{node} \cup \mathrm{node_m}$

19. $\mathrm{Seed_m} = \mathrm{Seed_m} \cup \mathrm{s_{merge}}$,$\mathrm{Seed_m} = \mathrm{Seed_m} - \mathrm{node_m}$

20. $\mathrm{merge} = \mathrm{False}$; BREAK

21. End If; End For

22. If $\mathrm{merge} = \mathrm{True}$:$\mathrm{Seed_m} = \mathrm{Seed_m} \cup \mathrm{node}$ End If; End For

23. For nodes in $\mathrm{Seed_m}$: $\mathrm{C_I} = \mathrm{nodes}$;

24. While $\mathrm{N_v}\,(\mathrm{C_I})! = \Phi$:$\mathrm{List_{can}} = \Phi$

25. For node in $\mathrm{N_v}\,(\mathrm{C_I})$:

26. $\mathrm{sim} = \mathrm{S_{vc}}\,(\mathrm{node}, \mathrm{C_I})$

27. If $\mathrm{sim} > \gamma$: $\mathrm{List_{can}} = \mathrm{List_{can}} \cup \mathrm{node}$ End If; End For

28. For $\mathrm{node_L}$ in $\mathrm{List_{can}}$: If $\mathrm{Q_f^{+}}\,(\mathrm{C_I}, \mathrm{node_L}) > \mathrm{Q_f}\,(\mathrm{C_I})$:$\mathrm{C_I} = \mathrm{C_I} \cup \mathrm{node_L}$

29. For $\mathrm{node_c}$ in $\mathrm{C_I}$: If $\mathrm{Q_f^{-}}\,(\mathrm{C_I}, \mathrm{node_c}) > \mathrm{Q_f}\,(\mathrm{C_I})$:$\mathrm{C_I} = \mathrm{C_I} - \mathrm{node_c}$;End If

30. End For; End If; End For

31. $\mathrm{recalculate}\mathrm{N_v}\,(\mathrm{C_I})$

Algorithm 1. RSLO
32. End While;$O = O \cup C_I$; End For

The overlapping community discovery algorithm based on relationship similarity and local expansion consists of 4 parts: 1) seed community discovery; 2) seed community merging; 3) local community expansion; 4) community final optimization. The discovery part of the seed community is mainly by calculating the relationship similarity of the nodes in the network, and then selecting the core node of the seed community according to the local information of the neighboring nodes, and forming the seed community with the neighboring nodes of the tight structure; in the local community In the expansion part, using the relevant information of local nodes and their neighbor nodes, select nodes that have high relationship similarity with the community and can optimize the adaptive fitness function to join the community, and realize the community division of all nodes based on this. Since each seed community expands the community based on the set of community neighbors, overlapping structures in the network can be found.

3.1 Seed community discovery stage

In the community discovery algorithm based on local expansion, the selection of core nodes in the seed community will directly affect the accuracy of the seed community discovery. This algorithm designs a new core node selection process. First calculate the local clustering coefficient of the nodes in the network, and obtain the relationship similarity of each node v based on this. Then, according to the relationship importance of each node v, count the number n_{cc} whose value is greater than the relationship importance of its neighbor nodes, if the ratio of n_{cc} to the number of neighbors of node v, $|\Gamma(v)|$is greater than the set threshold γ, then node v is divided into the core nodes of the seed community; then the similarity S_{vc} between the neighbor nodes of node v and the seed community is calculated, Find the node whose value is greater than the threshold δ, and add it to the seed community to get the final seed community PC. The specific steps are shown in lines 1–12 of Algorithm 1.

3.2 Seed community merger stage

In the process of discovering seed communities, there may be situations where two seed communities have a high degree of similarity. To avoid repetitive calculations in the third part, we can first merge seed communities with higher community similarity. Calculate the similarity between the seed community PC_i and the seed community PC_j according to S_{cc}. If $S_{cc}(PC_i, PC_j)$ is greater than the threshold δ, then the two seed communities will be merged to obtain a tighter set of seed communities PC_t. See the 13 ~ 22 lines of Algorithm 1 for specific steps

3.3 Community expansion phase

After obtaining a close seed community in the second stage, the community can be expanded. The steps of this part are as follows: first obtain the neighbor set N_v of the seed community, calculate the similarity

S_{vc} with the community according to each node in N_v, select the nodes with similarity greater than the threshold δ as candidate nodes, and then calculate the nodes in the candidate node set After joining the fitness function value of the local community, the candidate node that can increase the fitness function value is added to the community, otherwise it is set as a free node in the network, and the node that causes the fitness function value in the community to become smaller is deleted. Finally update N_v and repeat the above steps until N_v is empty. See lines 23 ~ 32 of Algorithm 1 for specific steps.

3.4 Community optimization stage

In the third part, because some nodes are set as free nodes, they may not belong to any community, and after the community is expanded, there may be communities with higher similarity. Therefore, the expanded community needs to be optimized. When optimizing, first calculate the similarity S_{vc} between the free node and each community. When S_{vc} is greater than the threshold δ, add it to the community, otherwise let it become a separate community; then, calculate the similarity between the community and the community again The degree S_{cc} then merges the communities whose value is greater than the threshold δ. Finally, the final community division result is obtained.

4. Experiment And Result Analysis

The experimental environment of RSLO is: an all-in-one computer configured with Intel(R) Core(TM) i5-8400 CPU @ 2.80 GHz and 8G memory. The operating system is Windows 10 Home Edition 64-bit. The algorithm code is implemented based on python3. The python modules used include: networkx2.4, python-igraph0.8.2, matplotlib3.3, and the java-based network visualization tool Gephi0.9.2

4.1 Datasets

In order to test the performance of the RSLO algorithm, we selected several real network data sets that are widely used in research experiments, and data sets based on artificial synthesis. Real network data sets, especially Karate club Karate, Dolphin network Dolphins, American college football game network Football, American political books network Pol books, etc. have become the most classic data sets in the field of community discovery. Almost all algorithms are in these Experiments on data sets, it can be said that these data sets have become benchmark data sets for measuring community discovery algorithms. The artificially synthesized real data set usually refers to the synthetic data set generated based on the LFR-benchmark [24] program. It has good node representation and community distribution differences. It is also the artificially synthesized data used in many studies in recent years. Set standards. A brief overview of these data sets will be given below, see in Table 2.

Table 2
Real-network datasets

No.	Name	Nodes number	Edges number	Description
1	Karate	34	78	Karate Club Network
2	Dolphins	62	159	Dolphin social network
3	Football	115	613	American Football Match Network
4	Pol Books	105	441	American Political Books Network

Karate [25]: A well-known real network data set based on long-term observation of 34 members of the American college student karate club. The nodes represent the members of the karate club, and the edges represent the relationships between the members. There is an obvious community structure, one around the coach and the other around the club owner.

Dolphins [26]: A real network data set based on long-term observation of the contact behavior between 62 dolphins. The nodes represent dolphins, and the edges represent frequent contacts between dolphins.

Football: According to the real data set obtained by American college students in the 2000 regular football game. The nodes represent the players, and the edges represent the friendship between the players.

Pol books [27] According to statistics, the United States is a network of political books sold by the online bookstore Amazon during the 2004 presidential election. The nodes represent books, and the edges represent the purchase relationship of connected books by the same buyer.

LFR-benchmark: It is an artificial synthetic network used to generate LFR benchmarks. Unlike real data sets, the LFR synthetic network has a clearer community structure and plays an important role in testing the performance of the community discovery algorithm. See the specific parameter settings in Table 3.

Table 3
LFR network parameter table

Parameter	Description
N	Total number of nodes
k	Node average degree
$maxk$	Maximum degree of node
$minc$	Number of nodes in the smallest community
$maxc$	Number of nodes in the largest community
μ	Mixed parameters
O_n	Number of overlapping nodes
O_m	Number of communities that overlapping nodes can belong to

The artificial data set parameters used in this article are as follows in Table 4:

Table 4
LFR datasets parameter setting

Name	N	k	$maxk$	$minc$	$maxc$	μ	O_n	O_m
LFR1	2000	20	50	20	100	0.1 ~ 0.7	100	3
LFR2	1000 ~ 10000	20	50	20	100	0.3 0.5	100	3
LFR3	2000	20	50	20	100	0.3	50 100	2 ~ 10

4.2 Evaluation index

In the field of overlapping community discovery, the two most commonly used evaluation indicators are modularity EQ [28] and Normalized Mutual Information(NMI) [29]. The two are briefly summarized below:

Modularity EQ is used to evaluate the quality of overlapping community structure, which is mainly used as an evaluation index for the division of real network data sets. The closer the EQ value is to 1, the better

the quality of the communities discovered by the algorithm. Usually, if its value is between 0.3 and 0.7, we think that the community discovery result of an algorithm is reasonable. The definition of modularity EQ is as follows:

$$EQ = \frac{1}{2m}\sum_{k=1}^{c}\sum_{i,j\in c_i}\frac{1}{O_i O_j}[A_{ij} - \frac{k_i k_j}{2m}]$$

Among them, m is the sum of the number of edges in the network; c is the number of communities found after the algorithm runs; O_i is the number of communities to which node i belongs; k_i is the degree of node i; A_{ij} is used to determine whether there is a node i and node j Connection, if the connection exists $A_{ij} = 1$, otherwise its value is 0.

Standardized mutual information NMI, on the premise of knowing the real community division result, can use NMI to measure the fit between the division result and the actual division. It measures the similarity between two vectors from the perspective of information theory. By combining with the real community results, the accuracy of the algorithm's community discovery results is objectively evaluated. The value range is between 0 and 1. For two different division results of A and B, the formula is defined as follows:

$$NMI = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} C_{ij} \times \log\left(\frac{C_{ij}\times N}{C_i \times C_j}\right)}{\sum_{i=1}^{C_A} C_i \times \log\left(\frac{C_{i.}}{N}\right) + \sum_{j=1}^{C_B} C_j \times \log\left(\frac{C_{.j}}{N}\right)}$$

Among them, C is a confusion matrix, and the element C_{ij} in the matrix represents the number of nodes that belong to the community i in the A division and the community j in the B division at the same time. C_A is the number of communities in the A division. $C_{i.}$ represents the sum of the elements in the i-th row of the confusion matrix C, and $C_{.j}$ represents the sum of the elements in the j-th column of the confusion matrix C. The larger the value of NMI, it means that the community discovery result of the algorithm is closer to the real community structure. Especially when the value of NMI is 1, the situation of the two communities is same.

4.3 Algorithm comparison

In order to verify the performance of the RSLO algorithm, the algorithm is compared with several classic overlapping community discovery algorithms. The contrasted algorithms include LFM algorithm, CPM algorithm, SLPA algorithm, DEMON algorithm [30]. The real network data set and the artificially synthesized data set are compared to verify the effect of the RSLO algorithm.

4.3.1 Comparison on real data sets

Figure 1 shows the results based on the EQ value of the RSLO algorithm and the classic algorithms found in 4 other communities on 4 real sets. It can be seen from the figure that the RSLO algorithm has good performance in general, except that the performance on some data sets is not as good as other

algorithms. Because the algorithm can find high-quality seed communities, and then through community expansion, community optimization and other steps to effectively discover the community structure in the real network. Figure 2 and Figure 3 show the results of Karate and Dolphins. The heterochromatic node is the overlapping node of the two communities.

4.3.2 Comparison on synthetic data sets

According to the network data set generated by the LFR benchmark, the parameter μ represents the complexity of the community structure. The closer the value of μ is to 1, the more complex the community structure in the synthesized network; on the contrary, the simpler the synthesized community structure. The following experiments are carried out on the effect of different algorithms according to different μ values. Figure 4 is the running results of different algorithms on the LFR1 artificial synthesis network data set. It can be seen from the figure that although μ takes different values, the value of NMI obtained according to the RSLO algorithm is higher than other algorithms, and as the LFR benchmark parameter μ increases, the downward trend of the RSLO algorithm is also greater than that of other algorithms. The slowness proves that the algorithm has better performance in more complex networks for other algorithms, which is mainly due to the accuracy of the selection of the core nodes of the seed community.

In the previous section, We ran an experiment on LFR2.we presented the experimental results of the community complexity of different algorithms in different artificial synthetic data sets. Next, we mainly explain the performance of each algorithm for different community sizes. It can be seen from the Fig. 5 and Fig. 6 that for different parameters μ, the total number of nodes in the network gradually increases, and the NMI value of the RSLO algorithm is basically stable and higher than some algorithms. This is mainly because in addition to the precise selection of the core nodes of the seed community mentioned above, the algorithm also processes the free nodes in the network to ensure the rationality of the community structure to a certain extent. Therefore, the RSLO algorithm has good performance in different community structures and artificial synthetic networks of different scales.

In order to detect the performance of overlapping community structures, we have compared several algorithms for different network overlap degrees $\mathrm{O_m}$. We conducted experiments based on the data set LFR3. It can be seen from the Fig. 7 that despite the different values of μ, the NMI value of the RSLO algorithm tends to be stable and leads other algorithms to a certain extent. With the increase of $\mathrm{O_m}$, the discovery of the network structure becomes more difficult, and the performance of all algorithms will deteriorate. This is due to the fact that after the initial seed community is issued, the discovered seed communities are preferentially merged to ensure the quality of the discovered communities when the community is expanded, and then the adaptive function fitness and relationship similarity are used to compare different seed communities. Expand the community to discover nodes in an overlapping structure in the network. Therefore, the RSLO algorithm can realize the discovery of overlapping structures in the network structure.

5. Conclusion

This paper proposes an overlapping community discovery algorithm (RSLO) based on relationship similarity and local expansion, which can discover overlapping community structures in the network. First, calculate the local clustering coefficient of each node, and then calculate the relationship similarity of the nodes according to the local clustering coefficient and the connection relationship between the nodes, and find the core node of the seed community according to the local clustering coefficient, and compare it with Nodes with close relationships together constitute a seed community. Then, the discovered seed communities are merged according to the similarity of the communities to reduce the amount of calculation in the community expansion stage. After that, the similarity between the neighbor nodes of the seed community and the community is calculated, and the adaptive fitness function is used to expand the community. Finally, optimize the result of community division, add free nodes in the network to the community with the greatest similarity, and merge the communities with too high similarity again to ensure the quality of the community structure discovered by the algorithm. Experimental results show that the algorithm performs well in some real network data sets and artificially synthesized data sets.

Declarations

Funding (information that explains whether and by whom the research was supported): This work was supported in part by the National Social Science Fund of China 18BGL266 and National Natural Science Foundation of China 41571401.

Conflicts of interest/Competing interests (include appropriate disclosures)

Availability of data and material (data transparency): The data set comes from the classic open data set in the community discovery domain

Code availability (software application or custom code): Follow the algorithm to complete the code

Authors' contributions

Hongtao Liu received his Master's degree in Computer Application Technology from Chongqing China Normal University in 2001, and his Doctor`s degree in artificial intelligence from Chongqing China Southwest University in 2004. His research interests include natural language processing, social networking and swarm intelligence. He is now an associate professor in the School of Computer science, Chongqing University of Posts and Telecommunications

And

Ning Wang received a Bachelor of Science degree in 2017. She is currently a graduate student at Chongqing University of Posts and Telecommunications. Her research interests include machine learning and social networks

References

1. Girvan M, Newman MEJ. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821–7826
2. Radicchi F, Castellano C, Cecconi F et al. Defining and identifying communities in networks[J]. Proceedings of the national academy of sciences, 2004, 101(9): 2658–2663
3. Yang S (2013) Networks: An Introduction by MEJ Newman, 720 pp. $85.00[J]. Oxford University Press, Oxford
4. Javed MA, Younis MS, Latif S et al (2018) Community detection in networks: A multidisciplinary review[J]. Journal of Network Computer Applications 108:87–111
5. Palla G, Derényi I, Farkas I et al (2005) Uncovering the overlapping community structure of complex networks in nature and society[J]. nature 435(7043):814–818
6. Palla G, Farkas IJ, Pollner P et al (2007) Directed network modules[J]. New journal of physics 9(6):186
7. Lu Zhigang Wu, Lu. Discovery of overlapping communities based on greedy Faction Expansion in ESN [J]. Computer engineering,2019 (7): 6
8. Zhang Z, Wang Z (2015) Mining overlapping and hierarchical communities in complex networks[J]. Physica A 421:25–33
9. Chen Junyu Z, Gang N, Yu et al (2016) A semi-supervised locally extended overlapping community discovery method [J]. Computer research development 53(6):1376
10. Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks[J]. New journal of physics 11(3):033015
11. Su Y, Wang B, Zhang X (2017) A seed-expanding method based on random walks for community detection in networks with ambiguous community structures[J]. Sci Rep 7:41830
12. Wang X, Liu G, Li J (2017) Overlapping community detection based on structural centrality in complex networks[J]. IEEE Access 5:25258–25269
13. Sun L, Liu J, Zheng X et al (2018) An efficient and adaptive method for overlapping community detection in real-world networks[J]. Chin J Electron 27(6):1126–1132
14. Li Y, Jing H, Youxi Wu (2019) Discovery method of seed node greedy Expansion in overlapping communities [J]. Minicomputer system, (5): 39
15. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E 76(3):036106
16. Gregory S (2010) Finding overlapping communities in networks by label propagation[J]. New journal of Physics 12(10):103018
17. Xie J, Szymanski BK, Liu X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]//2011 ieee 11th international conference on data mining workshops. IEEE, 2011: 344–349

18. Lu M, Zhang Z, Qu Z et al (2018) LPANNI: Overlapping community detection using label propagation in large-scale complex networks[J]. IEEE Trans Knowl Data Eng 31(9):1736–1749
19. Wang Gang (2018) Overlapping Community Discovery Algorithm based on Edge Propagation Probability [J]. Computer Knowledge and Technology, (21): 23
20. Guo Kun C, Erbao GW. Overlapping Community Discovery Algorithm based on Edge density Clustering [J]. Pattern recognition and artificial intelligence, 2018 (2018) å¹´ 08): 693–703
21. Wang Qing GU, Chun-mei ZHAO, Jian-jun CUI, Xin HONG, Wen-xing (2019) XU Wen-jing. Hybrid Parameter Adaptive Overlapping Community Discovery Algorithm based on Edge Trust [J]. Journal of Tianjin University (Natural Science Engineering Technology Edition 52(06):618–624
22. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks[J]. nature 393(6684):440–442
23. Lü L, Zhou T (2011) Link prediction in complex networks: A survey[J]. Physica A: statistical mechanics its applications 390(6):1150–1170
24. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms[J]. Physical review E 78(4):046110
25. Zachary WW (1977) An information flow model for conflict and fission in small groups[J]. Journal of anthropological research 33(4):452–473
26. Lusseau D, Schneider K, Boisseau OJ et al (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behav Ecol Sociobiol 54(4):396–405
27. V Krebs. Books About US Politics.[Online].Available:http://www.orgnet.com/,2004
28. Shen H, Cheng X, Cai K et al (2009) Detect overlapping and hierarchical community structure in networks[J]. Physica A 388(8):1706–1712
29. Danon L, Diaz-Guilera A, Duch J et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(09): P09008
30. Coscia M, Rossetti G, Giannotti F et al (2012) Demon: a local-first discovery method for overlapping communities[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 615–623

Figures

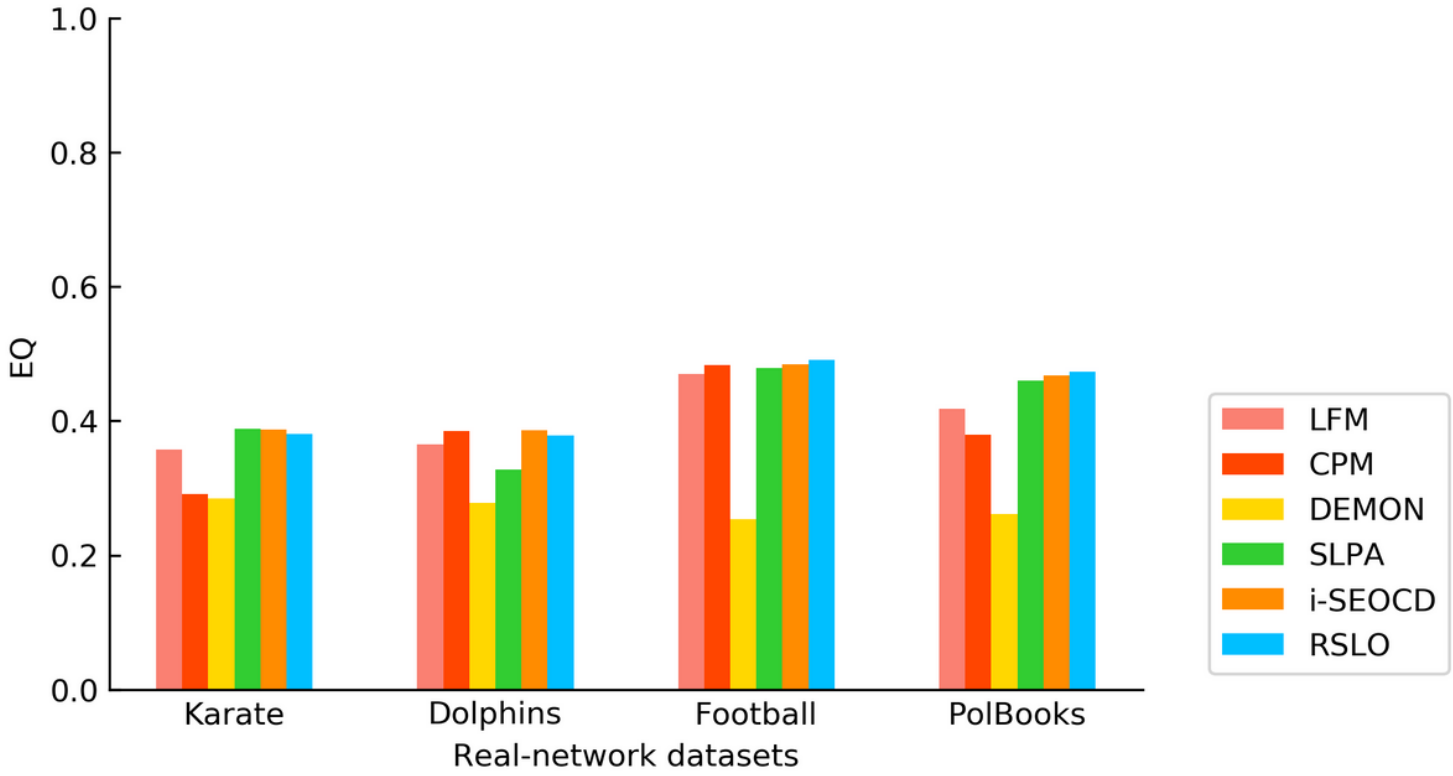

Figure 1

EQ comparison

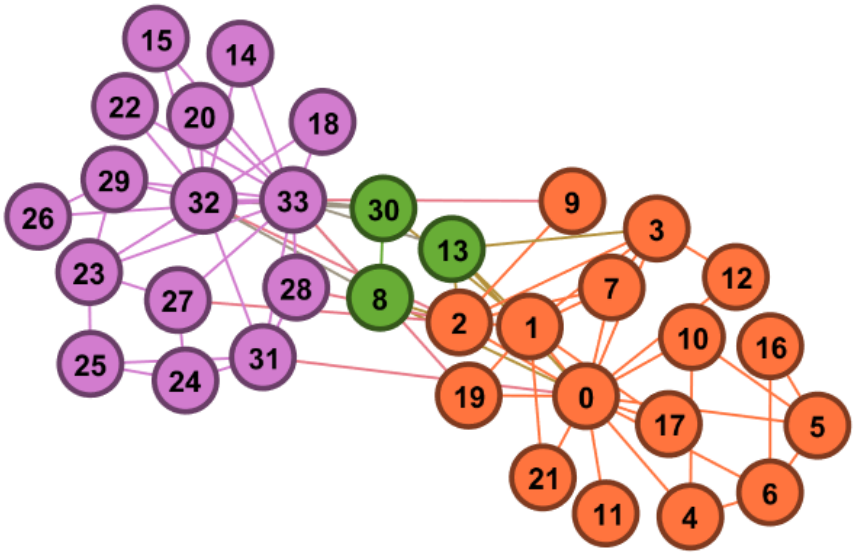

Figure 2

Karate overlapping community discovery results by RSLO

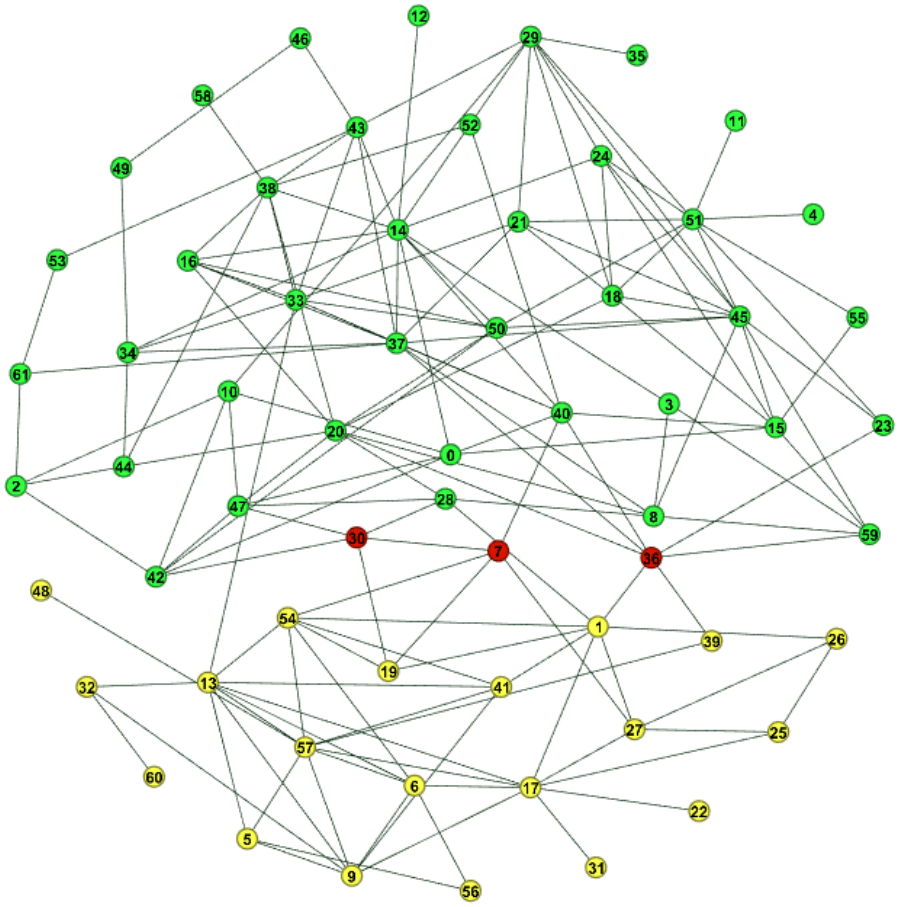

Figure 3

Dolphins overlapping community discovery results by RSLO

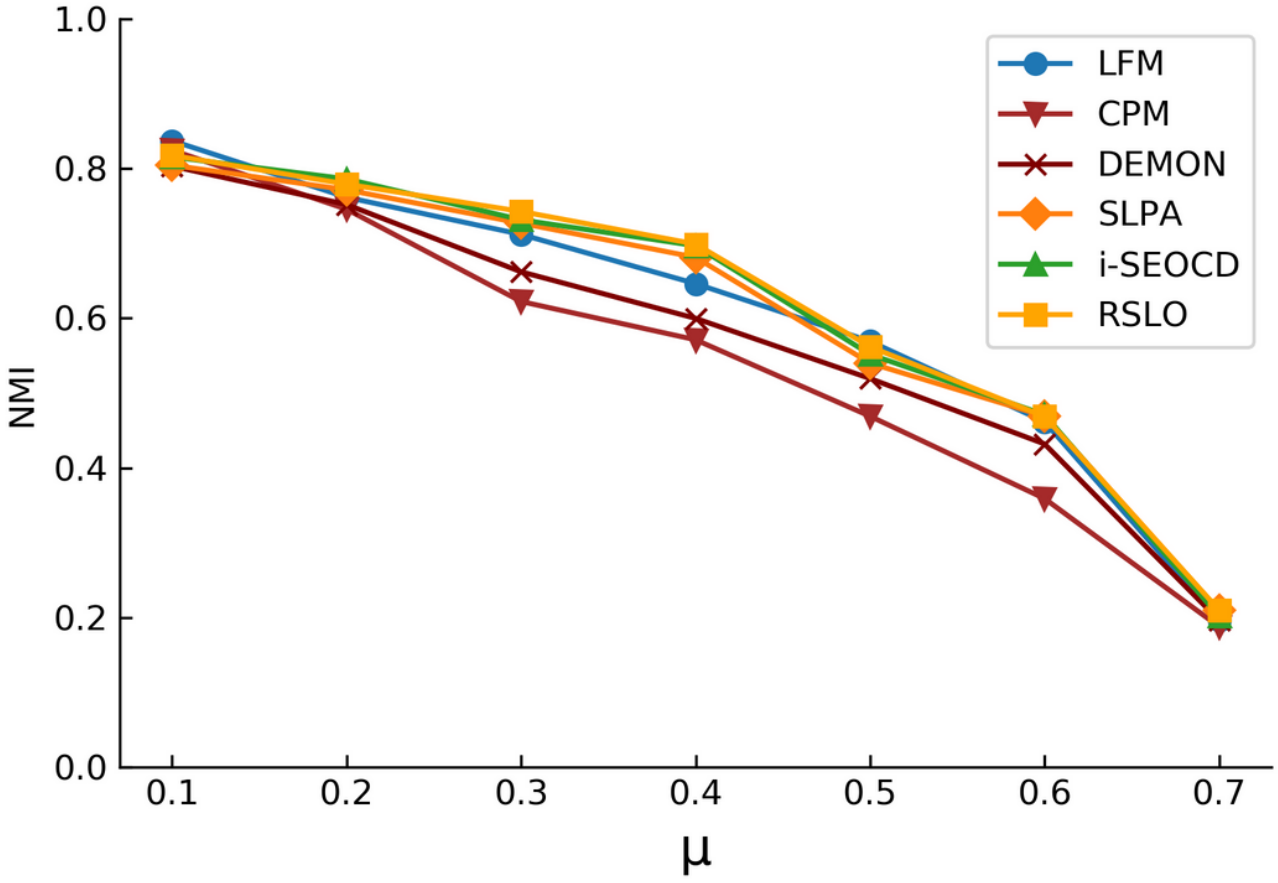

Figure 4

The result of complex network parameters μ

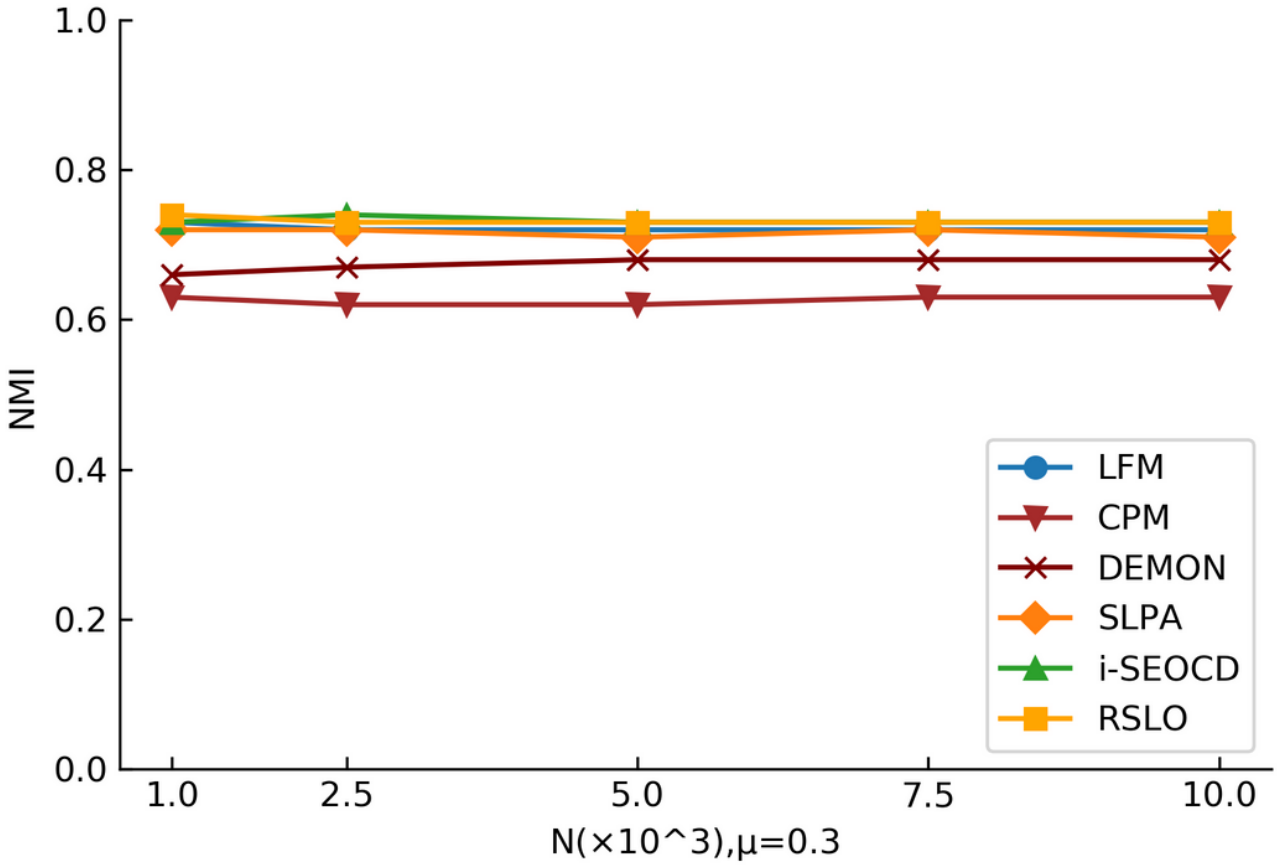

Figure 5

The result of different scales when μ=0.3

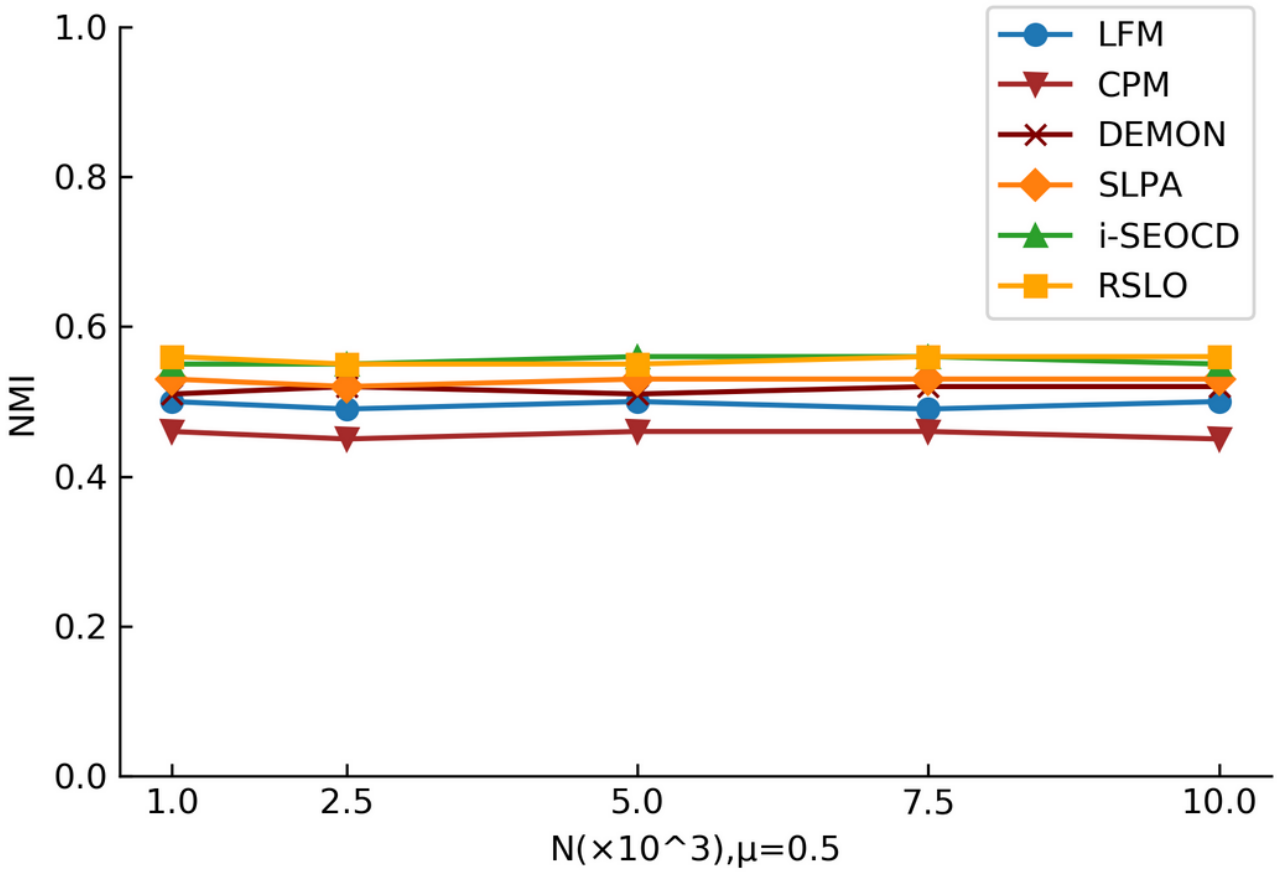

Figure 6

The result of different scales when μ=0.3

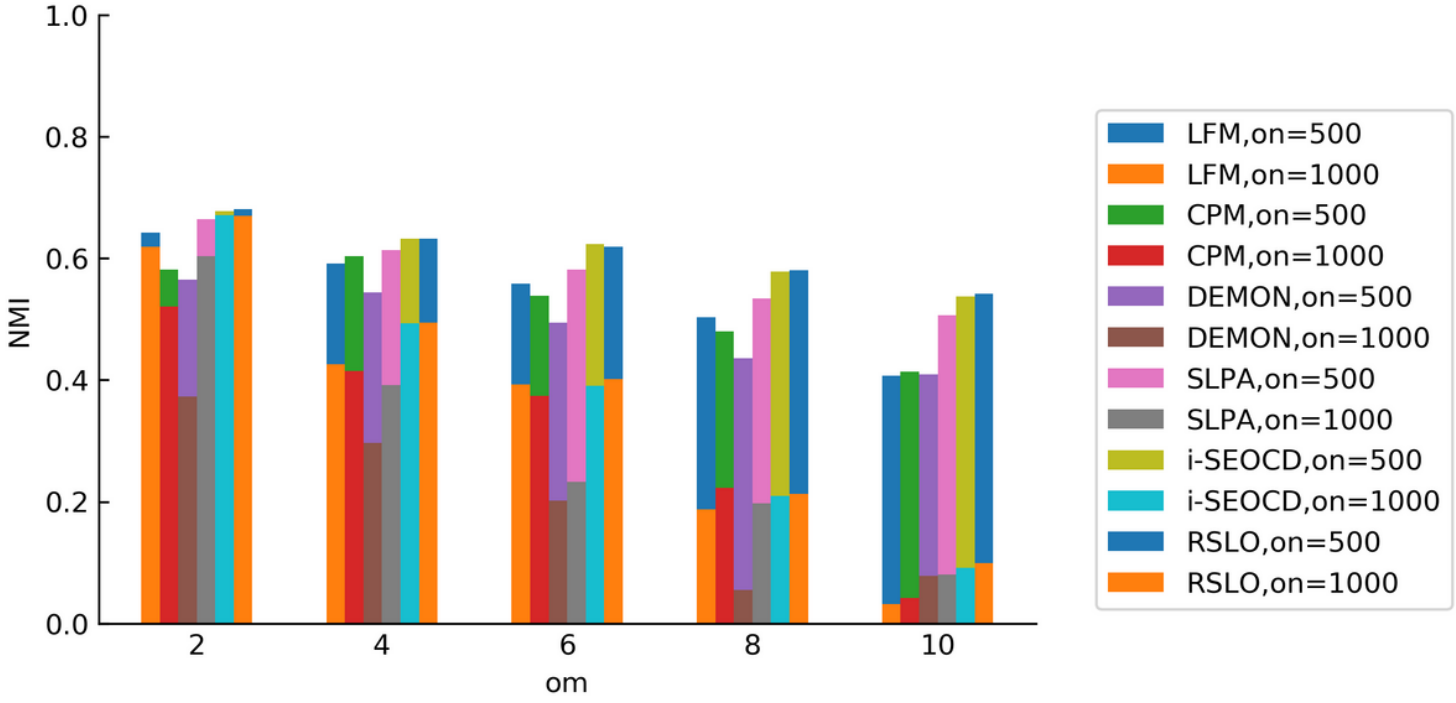

Figure 7

Results of different overlaps of networks