Overlapping and Hierarchical Community Discovery Based on Spectral Method

Kun Guo[a], Nanzhou You[b], WenzhongGuo[c], Yuzhong Chen[d,*]

College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China

[a]gukn123@163.com, [b]532380512@qq.com, [c]fzugwz@163.com, [d]yzchen@fzu.edu.cn

* Corresponding author: Yuzhong Chen

Keywords:spectral method, overlapping community, hierarchical community.

Abstract. With the development of community detection algorithm, traditional clustering methods are hardly capable enough to meet our needs. This paper proposes a new algorithm based on spectral method that can identify the overlapping and hierarchical communities in complex networks. Experiments on artificial and real datasets are conducted to verify the algorithm by comparing it with other algorithms. The experimental results show that our algorithm can efficiently identify overlapping communities in complex networks and find hierarchies of the communities.

1. Introduction

Many systems in the real world exist in the form of networks, such as the network of personal relationships in the social system, collaborative network of scientists, telephone networks and the Internet, et [1]. The clustering network architecture is one of the most popular and the most important attributes of complex networks, which has the features that the same clustering nodes interconnect with dense connectivity and different clustering nodes interconnect with sparse connectivity[2]. The clustering methods of complex networks aim at revealing the cluster-shape network architecture that does exist in a complex network. The study of clustering methods of a complex network is of great significance to analyzing the complex network topology, finding the hidden laws within complex networks and forecasting the behavior of complex networks, which has been widely used in pattern recognition, image processing field, information retrieval and so on [3].

At present, scholars have proposed a lot of clustering algorithms, based on the different implementations of clustering, which can be divided into hierarchical clustering, partition clustering and others[2]. However, with the growing complexity of application environments and new technologies' emerging, the traditional clustering methods face many new challenges. Traditional clustering requires different clusters cannot have a common overlapping part, but today, we often need to find the overlapping communities. In 2005, Pallaetc first proposed the concept of overlapping communities[4], which leads the study of community discovery to a new direction, but the mentioned algorithm CPM is very low in efficiency, and cannot find the hierarchical characteristics of the communities. Lancichinettietc proposed a hierarchical community detective algorithm which can also find the overlapping communities.[5] They try to find a locally optimal solution of functions to detect the overlapping structure of communities, and center on a seed node to detect the hierarchical structure of overlapping communities. But the seed node is randomly assigned, so it cannot be guaranteed that all the overlapping communities have hierarchical structures. Then Shen etc proposed EAGLE algorithm[8]. The algorithm uses a maximal clique as the initial community, takes the condensation method to consolidate communities according to their similarities, so that it can find the overlapping and hierarchical community structure. However, EAGLE algorithm needs to compute the similarities among communities repeatedly, which is with high complexity.

This paper aims to provide a new overlapping and hierarchical community discovery algorithm by combining spectral clustering and hierarchical clustering. This new algorithm is based on spectral method and it can find overlapping and hierarchical communities from complex networks more effectively.

2. Hierarchical and Overlapping Spectral Clustering (HOSC)

2.1 General Idea

By learning and summarizing existing algorithms, we combine the advantages of spectral clustering and hierarchical clustering to propose a new clustering algorithm which can identify the hierarchical and overlapping communities. Its general process is: first, initialize all of the objects to a large cluster, use the spectral clustering method to partition the objects, divide the cluster into two parts and process the nodes of the two parts at the same time, and choose the overlapping nodes out, create copies of them, then place these copies into those two parts to form two new non-overlapping clusters; then calculate the internal connection density of each cluster to select a less close cluster for the next division as a loop that continually divide the developed clusters. The ending condition depends on the community modularity Q[7] according to the results of each division. When the value of Q reaches its maximum, the algorithm terminates, and outputs the result of the division which ends the loop.

2.2 Algorithm Implementation

Spectral Clustering. The first step is to divide the original cluster into two parts by spectral clustering. First, process the inputting data to obtain the similar matrix W, each element W_{ij} of W matrix represents the distance between the sample point S_i and S_j. For undirected graphs, it's a symmetric matrix that all diagonal elements are zero; then calculate the degree matrix D, each element D_{ii} of D matrix represents the degree of sample point S_i, in addition to diagonal elements, the other elements are all zero; calculate Laplacian matrix via matrix W and D, the non-normalized Laplacian matrix is: $L=D-W$, the normalized Laplacian matrix has two forms, namely,

$L_1=D^{-1/2}WD^{-1/2}$. (1)

$L_2=D^{-1}L=I-D^{-1}W$(2)

Where I means unit matrix.

The next step is to divide graphs based on the results of Laplacian matrix. The initial criterion of dividing a spectrogram is Min-cut. It divides the original graph into two sub-graphs (A, B), such that the sum of linked weights between sub-graphs is minimum, which is called Min-cut. Since the Min-cut criterion considers only the external connections of clusters and doesn't limit the scale of the clusters, that is, it does not consider the density within each cluster, and thus it is likely to be oblique division (with a bias to smaller areas). In order to overcome it, we often use N-cut:

$$Ncut(A,B) = \frac{cut(A,B)}{\mathrm{vol(A)}} + \frac{cut(A,B)}{vol(B)} \quad (3)$$

Where *vol(A)* is the sum of weights from A to all nodes.

The division corresponding to the value of N-cut division is optimal for graphs. We can calculate L_1 to get its eigenvalues, and take out the second largest eigenvalue, and then find the eigenvectors corresponding to eigenvalues, which will be put into community A if it's positive and others will be in community B. Therefore, we can get the first division result, called A, B community.

Search for the Overlapping Nodes

Due to the characteristic that overlapping nodes are more closely connected with multiple communities, we will traverse all the nodes in community A and B according to the last division, and proceed as follows: Suppose node S belonged to the community A, its degree is D, the number of nodes connected with community B is D_1, if D_1 / D exceeds the threshold value Y, then we regard it as the overlapping nodes between community A and B. When finding all the overlapping nodes, we will replicate these nodes to obtain a copy of them (such as: S and $S`$), then place the original nodes and copies into community A and B severally, so that we can obtain two non-overlapping communities A_1 and B_1. At this point, we will be able to find overlapping nodes between A and B.

Hierarchical Division

As for the two new communities A and B, we calculate the Laplacian matrix of each community, and figure out the second largest eigenvalue of each matrix, which will be the degreeof separation within the corresponding community. The two new eigenvalues will be added to separation sequence in descending order. At every division, we give priority to the community that has the highest degree of separation for the next division, and record the state of each division for the final output.

Search for the Ending Node

In order to select an appropriate time when the current division result is more reasonable, we introduce the concept of modularity Q[7]. Newman and Grivan have put forward the concept of modularity with the purpose of measuring the quality of community division. Its formula is:

$$Q(C) = \sum_{i=1}^{t} \left[\frac{l_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right] \quad (4)$$

Where l_i means the total sides of community i, d_i means the sum of all node degrees of community i, m means the total sides of networks.

However, the discovery of overlapping communities cannot be evaluated very well by the modularity function Q. So we use an extension of modularity EQ[8] instead of Q. EQ is proposed by Shen etc and can evaluate overlapping communities more easily than Q, its formula is:

$$EQ = \frac{1}{2m} \sum_{ij} \frac{1}{O_i O_j} \left(W_{ij} - \frac{d_i d_j}{2m} \right) \delta(C_i, C_j) \quad (5)$$

Where $O_i$means the number of communitie to which vertex v belongs, W_{ij} is the similar matrix, if $C_i$is the same to C_j, the value of function $\delta(C_i, C_j)$ is 1, otherwise the value is 0.

In the process of division, we write down every modularity of each community, and when the community is been divided, we calculate the modularity respectively of new divided communities. If their sum is not less than the modularity of the original community, then this division is considered reasonable, and proceed to the next division; otherwise this division did not get better results than the original, and because we have taken the community with highest degree of separation out for division every time, it can be considered that this division has reached the optimum result, and record the result of the division to end the algorithm.

2.3 The HOSC Algorithm

Our algorithm can be summarized as follows:

Algorithm. HOSC

Input: network $G = (V,E)$, where V is the vertex set and E is the edge set
Output: clusters $C = \{C_1, C_2 \ldots C_i\}$

(1) construct W of input Dataset;
(2) compute D and $L=D^{-1/2}WD^{-1/2}$;
(3) compute eigenvalues$\{\lambda_i\}$ of L, take out the second largest λ, and its corresponding eigenvalues V;
(4) while true:
(5) divide V to 2 clusters $S_1, S_2$by its sign;
(6) for each node from S_1, S_2 do
if (the connection degree of node s>Y) then
clone s as s`;
put s` into the other cluster S;
end if
end for
(7) construct W,D,L of new groups,
(8) compute $\{\lambda_i\}$ of L, take out the second largest λ, and its corresponding eigenvalues V,
(9) record the new clusters with their details as C_i, put them to sequence C,
delete the old cluster from sequence C;
(10) put λ_1 and λ_2 to an ascending sequence l;
(11) if EQ = 1.0 then
break;
else
choose the first of sequence l for next devision;
end if
end while
(12) return clusters $C = \{C_1, C_2 \ldots C_i\}$.

2.4 Complexity Analysis

Let n be the number of nodes, m be the number of edges. In step (5), it takes $O(n)$ operations to visit all nodes, and in step (6), it takes $O(m)$ operations to visit all edges. Matrix M, D and L will be constructed in step (7) which costs $O(n^2)$ operations. Suppose $M(n)$ is the cost of a matrix-vector computation of $n \times n$, in step (8), it costs $O(M(n))$ operations to calculate the eigenvalues and eigenvectors of L. Suppose the dataset was divided into several clusters finally (we use letter t to represent the amount of clusters), the total time complexity is $O(n^2t+tM(n))$ operations in total and the space complexity is $O(n^2)$. So the algorithm's main computational cost lies in the matrix inversion. How to reduce the time needed for matrix operations under the condition of certain level of accuracy, is the key to reduce the time complexity of the algorithm.

3. Experiment

In order to verify the performance of the algorithm, we apply this algorithm to artificial datasets and real datasets, and compare the results with other clustering methods. Experimental results show that the proposed algorithm can effectively find the overlapping community in the network.

3.1 Experiments on Artificial Datasets

We generated a large amount of network of different scales ($100<N<2000$, N is the number of nodes of the network) and mixing parameter ($0.1<\mu<0.5$) by Lancichinetti's famous tools[6]. The details of the parameters are shown in Table 1. The proposed algorithm is compared with the COPRA algorithm for clustering accuracy on both artificial and real networks. We use EQ[8] as the index to evaluate the accuracy of community structure found by the algorithm. Experimental results show that our algorithm has higher accuracy than the COPRA algorithm.

Table 1 Detailed parameters for generating the artificial datasets

average degree	maximum degree	mixing parameter	minimum for the community size	maximum for the community size	number of overlapping nodes	number of memberships of the overlapping nodes
10	50	0.3	10	50	5%*N	4

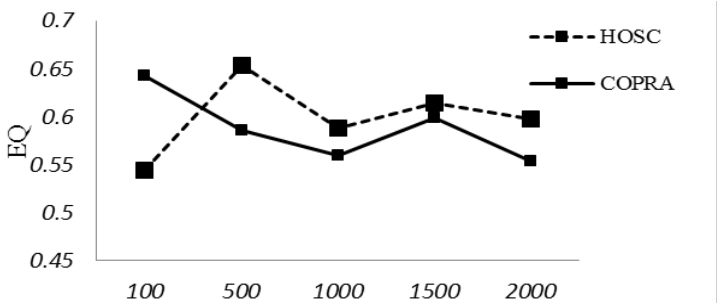

Fig.2 EQ with varying network size (u =1000)

From Fig. 2 it can be seen that the modularity obtained by our algorithm is slightly lower than that of copra algorithm when mixing parameter in a small data set is 0.3. But when the data set size becomes larger, the accuracy of our algorithm is improved and stabilized at a higher value, and that of copra are decreased. It can be seen that our algorithm has a better adaptability on a large data set.

In addition, we have tested and compared the datasets of different mixing parameter, the results are shown in Fig. 3 and Fig. 4, details of parameters are as shown in Table 2.

Table 2 Parameters of Fig. 3 and Fig. 4

average degree	maximum degree	minimum for the community size	maximum for the community size	number of overlapping nodes	number of memberships of the overlapping nodes
10	50	20	50	5%*N	4

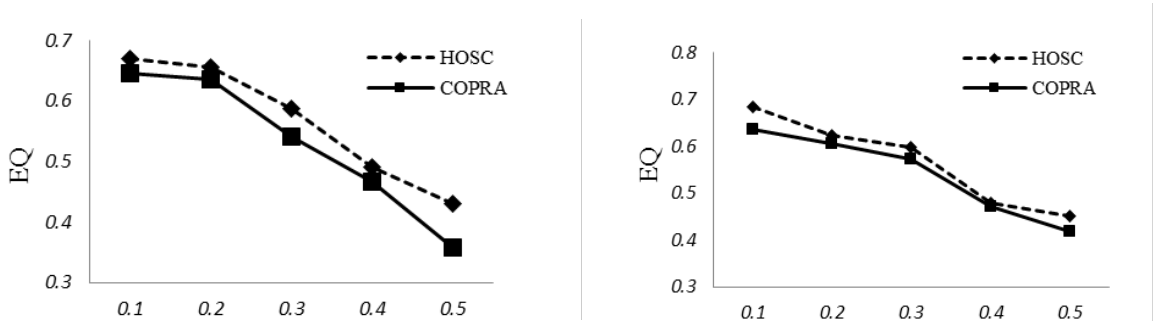

Fig.3 EQ with varying u (N=1000) Fig.4 EQ with varying u (N=2000)

From fig. 3 it can be seen that when the mixing parameter is increased, the modularity of the COPRA algorithm has dropped significantly, even for some datasets we cannot get a suitable partition. But our algorithm still can find communities with an acceptable modularity (about 0.45).

From fig. 4 it can be seen that when the mixing parameter is increased, although the modularity of our algorithm also dropped a little bit, it is still in a relatively higher level than COPRA algorithm, and the results are relatively stable. This shows that our algorithm has better adaptability and stability compared with the COPRA algorithm in the complex case of datasets.

3.2 Experimental Results on Real Datasets

In addition to the artificial datasets, we use two real datasets (Zachary 's Karate Club Network[9] and American college football Network[10]) to test the algorithm, and compare it with the WCF algorithm[11].

Zachary's Karate Club Network. The real dataset, Zachary's Karate Club Network, contains 34 nodes, 78 edges, the detailed network is shown in Fig. 5.

Fig. 5 Original Network of Zachary's Karate Club Network

Our algorithm divided the original network into four communities, and found overlapping nodes in the network including <3, 9, 10>. This result is consistent with the results of the comparison algorithm, whose results are shown in Fig. 6.

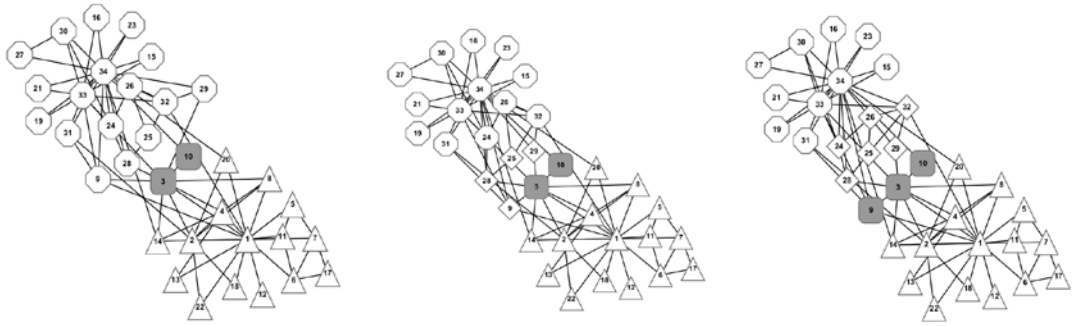

Fig. 6 the Result of Three Divisions

American College Football Network. The real dataset, American College Football Network, has 115 nodes, 616 edges, and the detailed network is shown in Fig. 7.

On the American College Football Network, our algorithm found 11 communities, which is consistent with the number of actual American college football network communities. At the same time, our algorithm also unearthed the overlapping nodes in the network, including <3,15,16,33,37,48,59,60,64,70,83,98>. The results are shown in Fig. 8.

It can be seen from the above results that, in the first division, the original network is divided into two balanced parts, and find the overlapping nodes between the two parts. In the subsequent divisions,

our algorithm chooses the community which has the highest degree of separation to next division, and as far as possible to avoid oblique division. In each division, the algorithm can detect the overlapping nodes of the network. After comparing with the results obtained from the other algorithms, our algorithm can find the overlapping nodes in the community with high accuracy.

Fig. 7 the Original Network of American College Football Network

Fig. 8 the Final Result of All Divisions

4. Conclusions

In this paper, we propose a new overlapping community detection algorithm, which combines the characteristics of spectral clustering and hierarchical clustering. The new algorithm can effectively identify the overlapping and hierarchical communities in complex networks. Then we use artificial datasets and real datasets to verify the effectiveness of the algorithm. The experimental results show that the algorithm can accurately identify the communities in network, and has a good adaptability to the change of the network size and mixing parameter. At the same time, the algorithm can find the overlapping nodes in the network and detect the overlapping communities. However, due to the complexity of the computation of matrix eigenvalues and eigenvectors, the computation speed can be improved. Our future work will focus on the reduction of the complexity of the algorithm, the improvement of the accuracy and the efficiency of the algorithm.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China under Grants No. 61103175 and No. 61300104, the Key Project of Chinese Ministry of Education under Grant No.212086, the Fujian Province High School Science Fund for Distinguished Young Scholars under Grand No.JA12016, the Program for New Century Excellent Talents in Fujian Province University under Grant No. JA13021, the Fujian Natural Science Funds for Distinguished Young Scholar under Grant No. 2014J06017, and the Natural Science Foundation of Fujian Province under Grant No. 2013J01230.

References

[1]. Yang B, Liu D.Y, Jin D, Ma H.B. Complex network clustering method[J].Journal of Software,2009,01:54-66

[2]. Sun J.G, Liu J, Zhao L.Y. Study on Clustering Algorithms[J]. Journal of Software, 2008, 19(1): 48-61.

[3]. Xu T.S. Research on Spectral Clustering[J]. Computer Knowledge and Technology,2012,16:3948-3950.

[4]. Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society. Nature, 2005,435(7043):814−818.

[5]. Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2008, 625(15):19-44.

[6]. Lancichinetti A and Fortunato S, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E, 2009,vol. 80, no. 1, pp. 1–8.

[7]. Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E. 2004, 69 (6): 066133.

[8]. Shen H, Cheng X.Q, Cai K, Hu M.B. Detect overlapping and hierarchical community structure in networks[J]. Physical A, 2009, 388: 1706-1712.

[9]. Zachary W W. An information flow model for conflict and fission in small groups[J].Journal of Anthropological Research, 1977, 33(4):452-473.

[10]. Girvan M,Newman M E J. Community structure in social and biological networks[C]//Proceedings of the National Academy of Sciences,2002,99(12):7821-7826.

[11]. Wan X.F, Chen D.B, Fu Y. Heuristic algorithm for detecting overlapping communities[J]. Computer Engineering and Applications,2010,03:36-38+41.