# A community detection algorithm based on structural similarity

**Xuchao Guo, Xia Hao, Yaqiong Liu, Li Zhang, Lu Wang***

College of Information Science and Engineering(Shandong Agricultural University), Taian, 271018,Taian,China
json_guo@sdau.edu.cn,haohh@sdau.edu.cn,969903442@qq.com,2808637709@qq.com, wanglusdau@126.com*

**Abstract.** In order to further improve the efficiency and accuracy of community detection algorithm, a new algorithm named SSTCA (the community detection algorithm based on structural similarity with threshold) is proposed. In this algorithm, the structural similarities are taken as the weights of edges, and the threshold k is considered to remove multiple edges whose weights are less than the threshold, and improve the computational efficiency. Tests were done on the Zachary's network, Dolphins' social network and Football dataset by the proposed algorithm, and compared with GN and SSNCA algorithm. The results show that the new algorithm is superior to other algorithms in accuracy for the dense networks and the operating efficiency is improved obviously.

## 1 Introduction

In reality, many systems can be expressed by the network. Examples include the Internet, the network of scientists' cooperation, the transportation network, a variety of protein interaction networks and many others. It is called "complex network" because of the complex internal structure of these networks. The vertex in the network represents an entity as well as the edge is described as a relationship, such as an interpersonal social network, people are represented as vertices that can be connected by edges. There are many basic statistical characteristics of complex networks: the small-world property [16], power-low degree distribution [2] and the community structure [4]. Many networks have in common is the community structure, also called community clustering, which is a characteristic that network is composed of a plurality of sets with the same or similar functions [17]. In short, the connections between internal vertices are dense but external vertices are sparse. Analysing the community structure and understanding the functions of complex networks have very important application prospects and practical significance to discover the potential rules in complex networks and predict the behaviour of complex networks. For example, it is helpful to find the source of infection, cut off the route of transmission so as to achieve the purpose of curbing the spread of disease, through analysing the structural characteristics of the disease transmission network and the spread of the disease.

At present, many methods have been proposed in terms of the community detection from different perspectives. The first method called GN algorithm used for identifying the community structure is proposed [4], which deletes the edge of highest betweenness every time. This algorithm has a high precision, although recalculating the edge betweenness is time consuming. Liu Dayou presents a community detection algorithm based on loop compactness (LTA), which can reduce the time complexity effectively [7]. The problem of local solution is solved from the perspective of discrete particle swarm optimization [1]. Ma et al. proposes a LED algorithm for detecting overlapping communities [14], which reduces the threshold sensitivity in a way. In addition, there are many other algorithms such as Kernighan-Lin [6] and Spectrum Average method [3]. In this article, an improved algorithm, called the community detection algorithm based on Structural Similarity with threshold *k* (SSTCA) is proposed, which improves the operational efficiency obviously.

## 2 Network description and structural similarity construction

### *2.1 Network description*

Networks are usually described as graph G (*V, E*), where *V* contains vertices and *E* is a set of edges, with symmetric adjacency matrix. It can not only express the relationship between nodes and nodes in G, but also represent the relationship between nodes and edges. In this section, a method which converts the adjacency matrix to structural similarity will be described. The definition of symmetric adjacency matrix is given as follows:

**Definition 1**(*Adjacency matrix*). Let $v_i, v_j \in V$ , the adjacency matrix is defined as $A = (a_{ij})_{n\times n}$ ,

$$a_{ij} = a_{ji} = \begin{cases} 1 \ , & if \ \ (v_i, v_j) \in E, \\ 0 \ , & if \ \ (v_i, v_j) \notin E, \end{cases} \qquad (1)$$

As is shown in formula above，the degree $m_i$ of vertex *i* is defined as $m_i = \sum_{j=1}^{n} a_{ij}$ , the number of edges in whole graph is $m = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}$ . Obviously, it is easy for us to convert the *E* into adjacency matrix.

### *2.2 Structural similarity construction*

The structural similarity is derived from the acquaintance model in sociology, which is used to measure the similarity strength between two people, that is, the more common neighbours of the two people, the greater the possibility of belonging to the same community[9]. It is widely applied to the community detection, because of the fast calculation speed for it does not need to consider the global nodes when the structural similarity is calculated. The method of structural similarity construction based on adjacency matrix is shown as follows:

**Definition 2**(Vertex neighbours). Let $i \in V$ , the vertex neighbours are defined as $\Gamma(i)$ by the vertex and its neighbours.

$$\Gamma(i) = \{ j \in V \mid (i, j) \in E\} \cup \{i\} \qquad (2)$$

**Definition 3**(Structural similarity) Let $i, j \in V$ , the structural similarity is donated by $\sigma(i, j)$ .

$$\sigma(i, j) = \frac{|\Gamma(i)| \cap |\Gamma(j)|}{\sqrt{|\Gamma(i)||| \Gamma(j)}} \qquad (3)$$

In this formula, the neighbours' number of vertex *i* is $|\Gamma(i)| = m_i + 1$ and the size of common neighbours of two neighbours is defined as follows:

$$|\Gamma(i)| \cap |\Gamma(j)| = \sum_{z=1}^{n} (a_{iz} \times a_{jz}) + 2 \qquad (4)$$

## 3 Community detection algorithm based on similarity

After calculating the structural similarities of network, it is necessary to take some strategies to delete the edges. Our algorithm is inspired by the idea of SSNCA algorithm, which is classified into four steps. A) The structural similarities in graph G are calculated as the weights of the edges. B) The weightless edges are deleted. C) Going to step A and B until there is no edge deleted. D) The community structure is detected [5]. In this algorithm, the structural similarity is used instead of the edge betweenness in the GN algorithm, which improves the computational speed. But only one edge is deleted at a time, the computational efficiency can be improved. Thus, we consider the structural similarity of two vertices as the weight if there is edge between the vertices. Then deleting one more edges once, which weights are less than the threshold.

### *3.1 The SSTCA algorithm*

In this section, we consider how to delete multiple edges, so as to reduce the iterations' times and time complexity. Therefore, this paper introduces threshold *k*, which is used to delete the edges of which weights are less than *k* , and gives the optimal community selection strategy to avoid reducing the accuracy of community partition and selecting *k* value blindly :The step size $\Delta k$ is defined and the modularity *Q* corresponding to the discrete threshold in the threshold interval is calculated to find out the maximum modularity, and then the best community partition is detected. So the idea of the algorithm is as follows:

A) Define the threshold interval $[k_1, k_2]$ , and $0 < k_1 < k_2 < 1$ , so the threshold $k_i = k_1 + (i-1)\Delta k$ , when the process is looping for *i* times, which *i* belongs to $0 < i < \frac{k_2 - k_1}{\Delta k}$ .

B) The structural similarity of two vertices is calculated as the weight of edge.

C) For the selected threshold $k_i$ , removing the edges of the network, which weights are less than the threshold $k_i$ , and repeating the step B, until there is no edge can be deleted. Thus the communities are detected via the breadth-first search, then current modularity $Q_i$ is calculated.

D) Let $k_{i+1} = k_i + \Delta k$ , go to step C, stopped until $k_{i+1} > k_2$ .

E) Find out the community partition which has the maximum modularity as the result of the optimal community partition.

### *3.2 Evaluation of community partition quality*

In order to describe the community quantitatively in the network, this algorithm uses the modular function *Q* proposed by Newman [10] as the standard of community partition quality evaluation. The formula is shown as follows:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \| e^2 \| \qquad (5)$$

Where *e* is symmetric matrix with $k \times k$ rows and columns, $e_{ij}$ is the proportion of the edges' number connected community *i* with community *j* in total edges' number of network. $e_{ii}$ is the proportion of all edges whose vertices in the community *i* . $a_i = \sum_j e_{ij}$ is the proportion of the total edges connected to the vertices of the community *i* [12].The larger the *Q*, the better the quality of community partition. In reality, the *Q* is 0.3~0.7, the upper limit of *Q* is 1, and *Q* is close to 1, indicating the community structure is more obvious. The negative value of *Q* indicates that the community structure is very poor [13].

## 4 Experimental results and analysis

In this section, tests were done on the Zachary's network, dolphins' social networks and football dataset by the proposed algorithm, and compared with GN and SSNCA algorithm. The experimental results show that the SSTCA algorithm can effectively discover the community structure, and the computation speed is significantly improved. The environment parameters of experiment were as follows:

Table1. The environment parameters of experiment

| Parameters | Description |
|---|---|
| CPU | Intel® Core™ i5-2400 CPU @3.10GHz 3.10 GHz |
| RAM | 6GB |
| Hard Disk | 500GB |
| Operating System | Windows 7 Ultimate |
| Software | MATLAB 2014b |

### *4.1 Zachary's Network*

Zachary's network [15] is a common data set in complex network community detection, which contains 34 vertices and 78 edges and reflects the social relations among the members of the American karate club. The club is divided into two major groups headed by the director and the principal because of the charges, as shown in Fig.1, the vertex 1 on behalf of the director and the president is vertex 33. The vertices of different colors represent the members of each group. The community result detected by SSTCA algorithm is as shown in Fig.2, of which solid vertices (yellow in the colourful figure) and hollow vertices (white in the colourful figure) represent two different communities, and the GN algorithm result slightly (Fig.3) but completely consistent with SSNCA algorithm classification results (see in Fig.4).

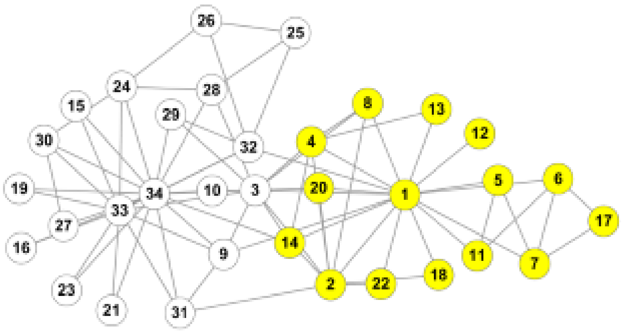

Fig.1 The original communities of Zachary's network

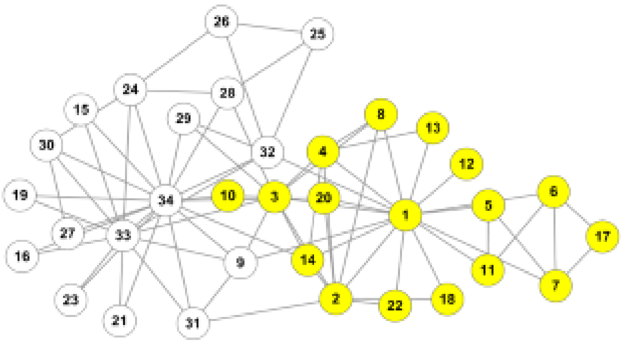

Fig.2 The result of Zachary's network by SSTCA algorithm

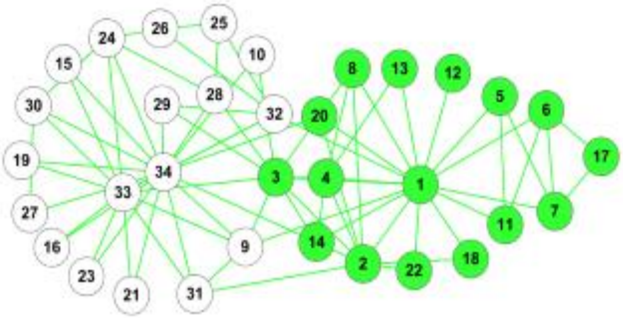

Fig.3 The result of Zachary's network by GN algorithm

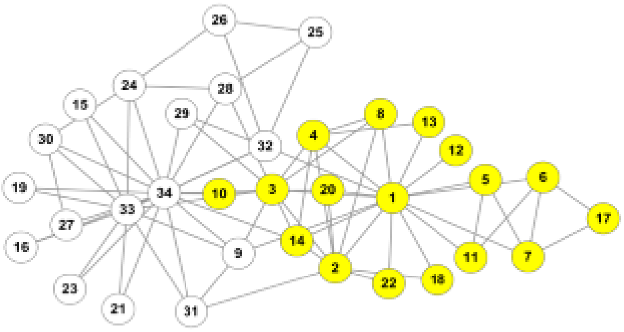

Fig.4 The result of Zachary's network by SSNCA algorithm

*4.2 Football club*

Football club dataset contains 115 vertices and 613 edges, in which each of vertices represents a team and each of edges represents a regular season between two teams. The teams are randomly divided into 12 groups, each group containing 8-12 teams. The number of teams in the group matches more than the number of matches between groups. Therefore, the community structure of network is obvious [11]. The result of SSTCA algorithm on Football club is as shown in Fig.5. The algorithm divides the network into 12 communities. Fig.6 shows the statistics of the size for every community, which is consistent with the actual situation in the 8~12 branch. Therefore, the algorithm can detect the community structure effectively.

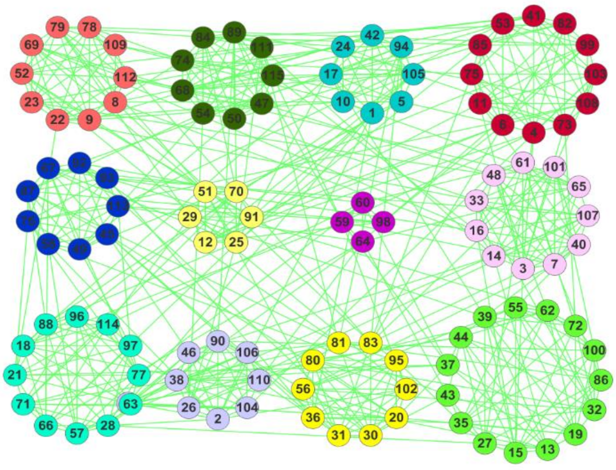

Fig.5 The result of SSTCA algorithm for football network

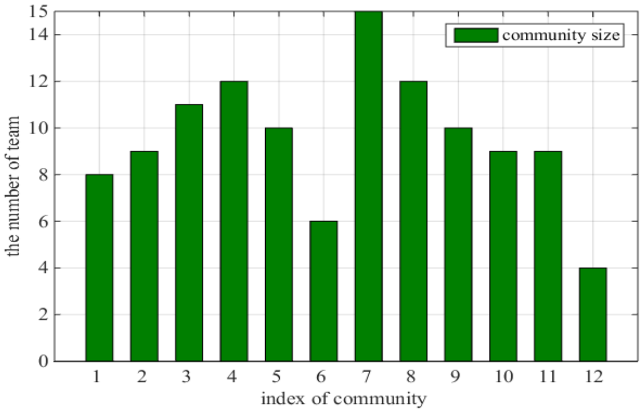

Fig.6 The statistics of size for every community

*4.3 Dolphins' social Network*

Lusseau constructs a dolphins' social network with 62 vertices and 159 edges via observing the habits of the 62 bottlenose dolphins. Each vertex represents a bottlenose dolphin, and the edge represents two dolphins' frequent activities. Lusseau find that these dolphins communicate with a specific pattern, that is, a certain community structure [8]. The experiment result of SSTCA algorithm for Dolphins' social network is shown in Fig.7. The vertices in the network represent a successful community in circular sets, different colors represent the different community. Obviously, the algorithm divides the network into 3 communities. The other contains outliers 37, 40, and 56 (as shown in the square shape). The experimental result is consistent with the actual network community.

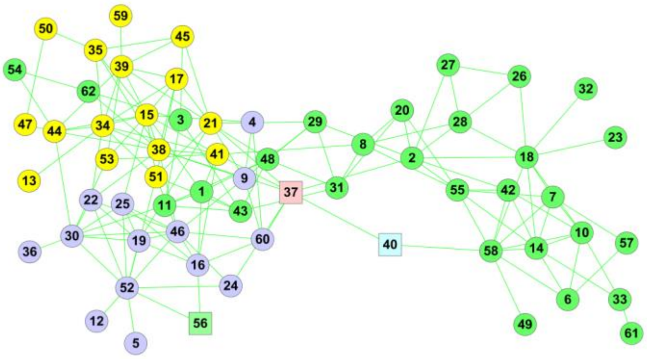

Fig.7 The result of SSTCA for dolphins' social network

*4.4 Comprehensive comparison of experimental results*

Fig.8 shows different $Q$ values of three datasets with the change of threshold $k$ by SSTCA algorithm ( $\Delta k = 0.02$ ), where the X axis represents a different threshold $k$, and the Y axis represents the $Q$ value obtained in the case of the selected threshold $k$. As can be seen from Fig.8, the best communities of Zachary are detected when the $k \in [0.32, 0.34]$ , and $Q$ gets maximum value 0.37, the community structure of football is detected when $k \in [0.38, 0.40]$, and dolphins' social network gets communities at 0.48. $Q$ remains unchanged in the threshold intervals [0.32, 0.34] and [0, 0.30] on Zacchary's network indicates that deleting a part of the network will not affect the overall community structure.

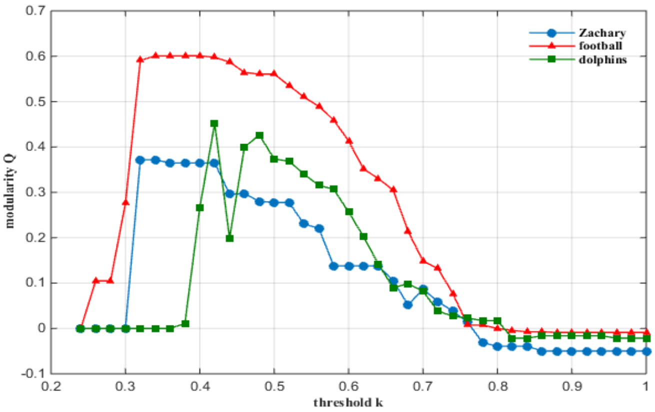

Fig.8 Modularity with different threshold ( $\Delta k = 0.02$ )

The detailed comparison of the parameters for the 3 datasets are shown in Table 2, where the average clustering coefficient indicates the density of the network, the larger the average clustering coefficient, the denser the network. The statistics of time consuming on SSTCA, GN and SSNCA algorithm for Zachary's dataset, dolphins' social network and football club are shown in Table 3. The parameters in the Table 3 are the average of 100 times. The experimental results of different data sets are shown by three algorithms: for the same dataset, the larger the network size, the longer the computing time; for the same dataset in different algorithms, the SSTCA algorithm is the fastest, followed by SSNCA, and the GN is the slowest. Because the average clustering coefficient of dolphins' social network is small, that is, the network is sparse, the partition effect is poor, what's more, the accuracy is low, and the Zachary's club and the football club network are dense, the SSTCA partition effect is better. In conclusion, the experimental results show that our algorithm can find the community structure effectively, and improves the computational efficiency of SSNCA algorithm.

Table 2. The parameters of datasets

| datasets | vertices | edges | average degree | average clustering coefficient |
|---|---|---|---|---|
| Zachary | 34 | 78 | 4.588 | 0.588 |
| dolphins | 62 | 159 | 5.129 | 0.303 |
| football | 115 | 613 | 10.661 | 0.403 |

Table 3. Run times of algorithms for different datasets

| algorithms | GN | SSNCA | SSTCA ($\Delta k = 0.01$) | SSTCA ($\Delta k = 0.02$) |
|---|---|---|---|---|
| Zachary | 0.279s | 0.206s | 0.257s | 0.129s |
| dolphins | 1.200s | 1.057s | 0.766s | 0.385s |
| football | 14.842s | 11.046s | 2.386s | 1.209s |

## 5 Conclusion

In this paper, an algorithm based on structural similarity of the community detection algorithm SSTCA was proposed by setting the threshold *k* and considering the deletion of multiple edges. Firstly, the definition of complex network and the method of constructing the structure similarity based on adjacency matrix were listed. Secondly, the main idea of the SSTCA algorithm and the modularity function Q which was used for evaluating the quality of community partition were described. Finally, the SSTCA algorithm was tested on Zachary's network, football club and dolphins' social network via comparing with the GN and SSNCA algorithm. The experimental results show that the algorithm has good performance for dense network and can improve the computation speed greatly. The next work is to optimize the threshold determination strategy and improve the accuracy of the algorithm for sparse networks.

## Acknowledgements

The author thanks Lu Wang and Xia Hao for their useful comments. This work was supported by the National Natural Science Foundation of China (General Program) (31571306).

## References

[1] Bai yun, Ren Guoxia. “Complex network community mining Based on Particle Swarm Optimization”. Computer Engineering, **41**, pp. 177-181, (2015).

[2] Barabasi A L, Albert R. “Emergence of scaling in random networks”. Science, 286, pp. 509-512,(1999).

[3] Fiedler M. Algebraic connectivity of graphs. Czech. Math. J. **23**, pp.298-305(1973).

[4] Girvan M, Newman M E J. “Community structure in social and biological networks”. Proceedings of the National Academy of Sciences of the United States of America, **99**, pp. 7821-7826, (2002).

[5] Jin Di, Liu Jie, Jia Zhengxue, et al. “K-nearest-neighbour network based data clustering algorithm”. PR&AI, **23**, pp. 546-551, (2010).

[6] Karypis G, Kumar V. “Parallel Multilevel k-way Partitioning Scheme for Irregular Graphs”. Journal of Parallel & Distributed Computing, **48**, pp. 278-300, (1996).

[7] Liu Dayou, Yang Jianning, Zhao Xuehua et al. “Community mining from complex networks based on loop tightness”. Journal of Jilin University (Engineering and Technology Edition), **43**, pp. 98-105,(2013).

[8] Lusseau, K. Schneider, O. J. Boisseau, P. Haase, et al., “The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations”, Behavioral Ecology and Sociobiology. **54**, pp. 396-405, (2003).

[9] M. Girvan, M. E. J. Newman, “Community structure in social and biological networks”, Proc. Natl. Acad. Sci. USA. **99**, pp.7821-7826, (2002).

[10] Newman M E J. “Community detection in networks: Modularity optimization and maximum likelihood are equivalent”. **16606,** pp.1-8.**(**2016).

[11] Newman M E, Girvan M. “Finding and evaluating community structure in networks”. Physical Review E Statistical Nonlinear & Soft Matter Physics, **69**, pp.1-16, (2003).

[12] Shi Wei, Fu Hegang, Zhang Cheng. “Overlapping communities detecting based on similarity of edge”. Application Research of Computers. **30**, pp.221-223, (2013).

[13] Sun Xijing, Si Shoukui. “Complex Network Algorithms and Applications”. Beijing: National Defense Industry Press, pp.224-225. (2015)

[14] Tinghuai Ma, Yao Wang, Meili Tang, et.al. “LED: A fast overlapping communities detection algorithm based on structural clustering”. Neurocomputing, **207**, pp. 488-500, (2016).

[15] W. W. Zachary, “An information flow model for conflict and fission in small groups”. Journal of Anthropological Research, **33**, pp. 452-473, (1976).

[16] Watts D J, Strogatz S H. “Collective dynamics of 'small-world' networks”. Nature, **393**, pp. 440-442,(1998).

[17] Žalik K R, Žalik B. “Multi-objective evolutionary algorithm using problem-specific genetic operators for community detection in networks”. Neural Computing & Applications, **10**, pp.1-14, (2017).