Weighted network community division based on co-neighbor nodes and similarity

Xiaoxia LUO1*, Xin SUN1, Xiangyu LUO1 and Yunzhong LUO2

1College of Computer Science and Technology, Xi’an University of Science and Technology, Xi’an Shanxi 710054, China
2College of Atmospheric Sciences, Nanjing University, Nanjing Jiangsu 210046, China
*Corresponding author’s e-mail: luoxx@xust.edu.cn

Abstract: In order to realize the accuracy of weighted network community division, based on the analysis of existing weighted network community dividing algorithm, a community dividing algorithm based on co-neighbor nodes and similarity is proposed. Firstly, the similarity between nodes is defined, and the weight of the edge is added to the local modularity function. Secondly, the Burt structural hole principle is used to select the important nodes. Then, the local clustering method is used to continuously expand the community. Finally, the judgment is made. The change of the Local modularity determines whether to continue the clustering, so as to obtain a more accurate result. Experiments on the simulation dataset and Zachary’s Karate club dataset show that compared with the CRMA algorithm, the algorithm is effectively improved in weighted network division time and accuracy.

1. Introduction
With the popularity of the Internet and the rapid development of computer technology, human social interaction is more complicated. Community discovery in social networks is easy to understand the network formed by the real world. How to quickly and effectively mine valuable information in such large-scale network data has become one of the current interdisciplinary research areas [1]. In such a large social network, the closeness of the relationship between individuals is different. If the weight is used to indicate the strength of the connection between individuals, the weighted community network is the social network with the weight between the nodes and the nodes [2]. By analyzing the structural characteristics of the weighted network, it has become a hot issue to find community relations from the middle. In 2004, Newman proposed that the WGN algorithm [2] applied the GN algorithm suitable for unweighted networks to the weighted network, and replaced the edge weights with the edge media to make the community discovery algorithm transfer from the unprivileged network to the weighted network. In 2008, Blondel et al. [3] proposed a BGLL algorithm based on modular gain but only for sparse networks. The algorithm is affected by the order of nodes, there are resolution limitations [4] and the lack of overlapping structure detection methods. In 2010, Han et al. [5] defined the incremental modularity of weights to achieve community partitioning, which improved the CNM algorithm, but the complexity of the algorithm was high. In 2011, Kumar [6] et al. proposed a density-based community discovery method, but this method only found a small number of communities in the Twitter data set. In the same year, Xie et al. [7] proposed an improved label propagation algorithm (SLPA), but the algorithm results were unstable. In 2014, Wang Kun et al. [8] proposed a similarity-based central clustering algorithm SCC to achieve community discovery of
weighted networks. In 2016, Guo et al. [9] improved the problem of the AGMA algorithm ignoring the weight of the edge between the current node and its undivided neighboring nodes, and proposed a weighted network community algorithm CRMA. The above algorithm ignores the relationship between two nodes and co-neighbor nodes. In view of the existing problems, this paper considers the direct relationship between nodes and nodes and the relationship of co-neighbor nodes, defines the similarity model between the weighted network nodes, and uses the local modularity as the threshold to accurately and effectively divide the weighted network.

2. Related definitions

2.1. Undirected weighted network related definition

In order to better describe and implement the weighted network community divisioning algorithm, the following definitions and descriptions are given:

Definition 1 An undirected weighted network \( G = (V, E, W) \), \( V \) representing a set of vertices, \( V = \{v_1, v_2, \ldots, v_i\} \), \( v_i \in V \), \( E \) representing a set of edges, \( E = \{e_1, e_2, \ldots, e_j\} \), \( e_j \in E \), \( W = (w_1, w_2, \ldots, w_q) \) representing a set of weights, \( W(v_i, v_j) \) representing the weight between the node \( v_i \) and the node \( v_j \).

Definition 2 The degree of the node \( i \) indicates the number of nodes directly connected to the node \( i \), which is recorded \( D(i) \).

Definition 3 For node \( i \) and node \( j \), if node \( i \) and node \( j \) have edges, the node \( i \) and node \( j \) are mutually referred to each other’s neighbors. If there is a node \( x \) such that there is an edge between the node \( i \) and the node \( j \), then the node \( x \) is the co-neighbor node of the nodes.

In the analysis of the node similarity function in the existing weighted network community partitioning algorithm, it is found that the similarity between nodes plays a decisive role in community division. In the literature [10], the proximity function is for the division of unweighted networks. After considering the weights of the weighted network edges, this paper redefines the similarity between nodes and nodes in the weighted network, and proposes and defines the proximity. The concept of making an unprivileged network model suitable for weighted networks.

Definition 4 The proximity between a node \( i \) and a node \( j \) is defined as:

\[
S_{ij} = \frac{W(i, j)}{D(i)}
\]  

among them, \( W(i, j) \) represent the weight value of the edge when there is an edge between the node \( i \) and the node \( j \). When there is no edge between the node \( i \) and the node \( j \), \( W(i, j) = 0 \).

Definition 5 The similarity between node \( i \) and node \( j \) is defined as:

\[
\delta_{ij} = S_{ij} + \sum_{s \neq i, j} S_{is} \cdot S_{js}
\]  

For example, in Figure 2.1, there are 10 nodes and 20 edges, and the weights of the edges are \((0, 1)\). For a node \( v_6 \), its neighbor node has four nodes, its \( \{v_2, v_4, v_8, v_9\} \). According to the definition 4, you can know, \( \delta_{62} = 0.131 \), \( \delta_{64} = 0.0625 \), \( \delta_{68} = 0.03625 \), \( \delta_{69} = 0.1515 \). By using the result, you can get the node \( v_9 \) that is most similar to the current node \( v_6 \).
In order to calculate the similarity between nodes and each community, a definition is proposed.

Definition 6 The similarity between node \( i \) and communities \( C \) is defined as:

\[
\rho(i, C) = \sum_{s \in C} \rho_{js}
\]

(3)

2.2. Local modularity definition

The modularity is a function proposed by Newman-Girvan [2] to judge the quality of community division. The larger the module value, the better the corresponding community division result. However, due to the inherent resolution limitation of the modularity definition [11], it tends to find community structures of similar size. That is to say, when the sum of the edges of the unknown network and the local information of the network are known, the modularity proposed by the GN algorithm cannot be used. It is necessary to extend the community outward from a node, so Clauset [12] proposed the concept of partial modularity. The expression for the local modularity is:

\[
Q = \frac{\sum_{i,j} C_{ij}\delta(i, j)}{\sum_{i,j} C_{ij}}
\]

(4)

The formula (4) represents the total number of relationships in the community, and the denominator represents all the edges in the community, where \( \delta(i, j) = 1 \), this shows that node \( i \) and node \( j \) are in the community \( C \), otherwise \( \delta(i, j) = 0 \).

The traditional partial modularity definition of equation (4) has no weight. In this paper, equation (5) is redefined so that it can express the weight between two adjacent nodes. The defined representation is: where \( \delta(i, j) = W(i, j) \), this means that there is an edge between the node \( i \) and the node \( j \) and at least one node is in \( C \), otherwise \( \delta(i, j) = 0 \).

3. Algorithm design

3.1. Algorithm thought

The basic idea of weighted network community dividing algorithm based on co-neighbor nodes and similarity is: firstly, Burt structural hole theory is used to select important nodes from the community network that needs to be divided [13]; then, from the important nodes, combined with proximity utilization The method of local clustering divides the node with the largest similarity into the community continuously; finally judges the change of the local modularity. When the local module degree decreases, then the clustering is stopped, a community division is completed, and the next community division is started. When the local modularity increases, the next node with the greatest similarity to the divided community is searched until the local module value becomes smaller and the community dividing algorithm ends.

![Figure 1. Schematic diagram of undirected weighted network](image-url)
3.2. Algorithm Description
Step 1: For a given undirected weighted network $G$, use the Burt structural hole principle \cite{13} to calculate the influence of each node, select the node $v_i$ with the most influence as the important node, add it to the community $C_m$, and delete the node $v_i$ from the network $G$.

Step 2: Calculate the node $v_j$ with the highest similarity to the important node $v_i$ according to formula (2), and add the found node $v_j$ to the initial community $C_m$, and form a new community $C_m$.

Step 3: Calculate the current community modularity $Q$ by using equation (4), and compare it with the current maximum modularity $Q_{\text{max}}$. In the case of $Q > Q_{\text{max}}$, judge $G$ whether it is empty. If it is empty, the algorithm ends, otherwise return to step 2; if $Q \leq Q_{\text{max}}$, then Go back to step 1.

4. Experiment and result analysis

4.1. Experimental environment
The experiment in this paper is that the processor is Intel Core i5-4200M 2.5GHz dual core, 8GB of memory, the operating system is Microsoft Windows 7, and the programming language is Python3.6.

4.2. Data source
In order to verify the algorithm, the simulation data set in \cite{14} and the real weighted network data downloaded from http://konect.cc/networks/ are used. The data set information is as follows:

(1) Simulation data set: There are 36 nodes and 70 edges, and the weight range of the edges is (0, 1).

(2) Zachary’s Karate club: The data was from 1970 to 1972 when Wayne Zachary observed the interaction between members of a US karate club and established a network of relationships between members. The data set contains a total of 34 nodes and 78 edges. The researcher uses nodes to represent the members of the club, and indicates that there is communication between the members and the members. The weight of the side indicates the closeness of the interaction between the members and the members.

4.3. Analysis of experimental results
This algorithm and the algorithm CRMA are implemented by using two data sets in 4.2. The results of the two algorithms are shown in Figure 2, Figure 3, and Figure 4. The results in the number of communities, the spent times and modularity are shown in Table 1.
Figure 2. Simulation data set partition results

Figure 3. CRMA algorithm for Zachary’s Karate club dataset partition results

Figure 4. The number of Zachary’s Karate club algorithms in this paper

Table 1. Experimental results of two algorithms on different data sets.

| Algorithm          | Number of communities | The time spent (s) | Modularity | Number of communities | The time spent (s) | Modularity |
|--------------------|-----------------------|-------------------|------------|-----------------------|-------------------|------------|
| CRMA               | 7                     | 0.1153            | 0.8039     | 7                     | 0.1120            | 0.8039     |
| Algorithm          | 2                     | 0.1229            | 0.4547     | 4                     | 0.1410            | 0.4752     |

It can be concluded from Table 1 that for the simulation data set in [14], the results of the two algorithm divisions are identical, as shown in Figure 2, and the weighted modularity after division is 0.8039, but the two algorithms run. The time taken to end is different, and the time used by the algorithm in this paper is 2.86% higher than the CRMA algorithm time. For the Zachary’s Karate club dataset, the CRMA algorithm divides it into two communities, as shown in Figure 3. The algorithm divides it into four communities, as shown in Figure 4. Although the time used is increased, the weighted modularity of the partitioned community is 4.51% higher than that of the CRMA algorithm. Generally, for the same data set, the network module size obtained after execution
of different algorithms and the number of communities obtained after partitioning will be different. Generally, the number of communities corresponding to the module is more accurate.

5. Conclusion
Based on the co-neighbor nodes of nodes, this paper proposes the concept of similarity between two nodes combined with local modularity for weighted network community dividing. On the premise of the structural integrity and accuracy of the community, the two sets of data sets are verified. The results show that the proposed algorithm can divide the weighted network more accurately and effectively, and the segmentation result is improved in rate and accuracy.

The algorithm has achieved certain effects on the weighted network community dividing problem, but some problems need to be considered in this research field. For example, the community finds that in the fields of sentiment analysis, data mining and knowledge mapping, the node attributes are often complicated [15].

Acknowledgments
This work was supported by National Natural Science Foundation of China (No.61702408). Technicians at the Xi'an University of Science and Technology who provided tools for the tests and who helped to conduct the experiments are also acknowledged.

References
[1] Barabasi A L, Albert,R,Jeong H, et al. Power law distribution of the World Wide Web[J]. Science, 2000, 287(5461).
[2] Newman M E J. Analysis of weighted networks[J]. Phys Rev E, 2004, 70: 56133.
[3] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.
[4] Yang Chunming, Wang Yujin. Analysis of Weighted Complex Network Community Discovery Algorithm Based on Module Degree Optimization[J]. Journal of Southwest University of Science and Technology, 2016, 31(04): 84-89.
[5] Han H, Wang H, A new algorithm to detect community structures on weighted network[C]// Proceedings of 2010 International Conference on Computational Intelligence and Software Engineering. Wuhan: IEEE, 2010: 1.
[6] Subramani K, Velkov A, Ntoutsiti I, et al. Density based community detection in social networks[C]// Proceedings of 2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application. Bangalore, India: IEEE IMSAA, 2011.
[7] Xie J R, Symanski B K. Community detection using a neighborhood strength driven label propagation algorithm. 2011 IEEE Neworks Science Workshop. West Point, NY, USA, 2011: 188-195.
[8] Wang Kun, Lu Guanghong, Liang Zhaowe, Ye Maoying. A Method of Weighted Complex Network Community Discovery Based on Similarity[J]. Journal of Sichuan University (Natural Science Edition), 2014, 51(06): 1170-1176.
[9] Guo J F, Liu M M, Liu L L, et al. An improved community discovery algorithm in weighted social networks[J]. ICIC Express Lett, 2016, 10: 35.
[10] StanDan. Division of dynamic social network community structure [D]. Xi'an University of Science and Technology, 2018.
[11] Fortunato S, Barthélémy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104 (1): 36-41.
[12] Clauset Aaron, Newman M E J, Moore Cristopher. Finding community structure in very large networks.[J]. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 2005, 70(6 Pt 2).
[13] Burt R S. Structural holes[M]. America: Harvard University Press, 1992.
[14] Yang B, Liu J M. An autonomy oriented computing approach to distributed network community mining[C]//Proceedings of the First International Conference on Self-Adaptive and Self-Organizing Systems. Boston, USA: IEEE SASO, 2007.
[15] Zhang Xin. Research on community discovery methods in complex networks [D]. Harbin Institute of Technology, 2017.