J Inf Process Syst, Vol.15, No.5, pp.1055~1067, October 2019
https://doi.org/10.3745/JIPS.04.0133

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Community Discovery in Weighted Networks Based on the Similarity of Common Neighbors

Miaomiao Liu*, Jingfeng Guo**, and Jing Chen**

Abstract

In view of the deficiencies of existing weighted similarity indexes, a hierarchical clustering method initialize-expand-merge (IEM) is proposed based on the similarity of common neighbors for community discovery in weighted networks. Firstly, the similarity of the node pair is defined based on the attributes of their common neighbors. Secondly, the most closely related nodes are fast clustered according to their similarity to form initial communities and expand the communities. Finally, communities are merged through maximizing the modularity so as to optimize division results. Experiments are carried out on many weighted networks, which have verified the effectiveness of the proposed algorithm. And results show that IEM is superior to weighted common neighbor (CN), weighted Adamic-Adar (AA) and weighted resources allocation (RA) when using the weighted modularity as evaluation index. Moreover, the proposed algorithm can achieve more reasonable community division for weighted networks compared with cluster-recluster-merge-algorithm (CRMA) algorithm.

Keywords

Common Neighbors, Community Discovery, Similarity, Weighted Networks

1. Introduction

Community discovery in social networks has theoretical significance and practical value for understanding the topology and behavior patterns of the network. However, edges in real networks always have weights. For example, the closeness of relationships between individuals in social networks is different. If we use the weighted network to describe such a system, it can better express these relationships. Weighted networks are networks in which edges have weight attributes. The weight can not only express whether there is a relationship between two nodes, but can also express the closeness of this relationship. For example, the weight in the air transport network represents the number of flights between two airports and the weight in the communication network represents the talking time between two users. The weight can better express real systems and help to understand its nature. It also has practical significance for community discovery.

At present, there have been some researches on community discovery in weighted networks. Newman replaced the edge betweenness with the weighted edge betweenness and proposed weighted Girvan-

Manuscript received March 3, 2017; first revision April 18, 2017; accepted May 12,2017.
Corresponding Author: Miaomiao Liu (liumiaomiao82@163.com)
* Northeast Petroleum University, Daqing, China (liumiaomiao82@163.com)
**College of Information Science and Engineering, Yanshan University, Qinhuangdao, China (jfguo@ysu.edu.cn, 58558607@qq.com)

Newman (WGN) algorithm [1]. Subramani et al. [2] proposed a community mining method based on the variable density to cluster nodes. However, experimental results were not good and only a small number of communities were detected. A study of Liu et al. [3] put forward the attractiveness-based community detection (ABCD) algorithm for the clustering of large weighted networks based on the attractiveness between communities. Sharma [4] proposed automatic graph mining algorithm (AGMA) which can divide the weighted signed graph into several communities according to the link type and weights. Lu et al. [5] proposed intra-centrality and inter-centrality (I^2C) algorithm based on conductance which joined the edge that had the greatest degree of membership into the community and used the community conductivity to determine whether a new community would form. Wang et al. [6] proposed a central cluster algorithm based on similarity which selected the node with the largest center degree as the center of the community and achieved community discovery in weighted networks based on the degree of ownership of the node. Lin et al. [7] proposed a hierarchical community discovery method based on parallel decomposition of weighted graphs. Wang [8] proposed a splitting algorithm based on greedy selection strategy. However, experiments on weighted networks of karate club and dolphins showed that there were some deviations between the division results and the real datasets. Zhan [9] proposed an algorithm to find local communities in weighted networks, which used the node with the maximum local weight as starting node and found the local community by gradually adding nodes into it. Zhao and An [10] realized division of the service community in weighted networks by calculating the optimal path tree, similarity index and dispersion index of the community between mobile nodes. Yao [11] proposed a community discovery method in weighted short message network. Guo et al. [12] improved AGMA algorithm and proposed CRMA algorithm for community discovery in weighted networks.

Overall, most of the existing algorithms are suitable for community discovery in traditional social networks in which the weight of the edge is always 1, and there are relatively few researches on community discovery in weighted social networks. Additionally, community discovery in weighted networks should not only consider whether the nodes are connected, but also consider the closeness of these relationships. So the weight should be taken as an important factor in the clustering process. However, existing weighted similarity indexes such as weighted common neighbor (CN) and weighted Adamic-Adar (AA) only consider the influence of the weight information of the CNs on the similarity, which ignore the effect of the degree and strength of CNs on the similarity. As a result, these algorithms have poor performance in community division in certain networks in which most node pairs have less common neighbors such as US Airports network. Besides, the hierarchical clustering method single based on the modularity is complex. Moreover, a good algorithm should meet two requirements at the same time, namely, the higher accuracy and the lower complexity. However, it is difficult for most existing algorithms to achieve the both.

For all these reasons, the algorithm initialize-expand-merge (IEM) is proposed in order to achieve a higher quality of community division in weighed networks and ensure the feasibility and effectiveness of the time meanwhile. Firstly, the similarity between nodes based on their CNs is defined so as to complete the clustering fast. Then it merges communities based on the target of maximizing the modularity. Lastly, the effectiveness and correctness of the algorithm are verified through experiments. The paper analyzes related current research firstly. Then the main idea, the definition and description of IEM algorithm are given. The last two sections are experiments and the conclusion.

2. Ideas and Preliminaries

2.1 Main Ideas

Community division for weighted networks should ensure that nodes in the same community are densely and closely connected. Additionally, according to the hierarchical clustering algorithms, the more similarity the two nodes have, the greater possibility of their belonging to the same community is. So the key of the algorithm is to effectively capture topological properties which affect the similarity and reasonably define the similarity index to complete the clustering and community discovery. In order to reduce the complexity, only the effects of the degree, the strength and the weight on the similarity of the two nodes are taken into account in our algorithm. We think that if the two nodes are not directly connected, their similarity is 0. Otherwise, if they are directly connected, their similarity depends on the contribution of their common neighbors.

Firstly, when measuring contributions of CNs to the similarity of the two nodes, we think the more CNs the two nodes have, the higher similarity they would have. Moreover, in weighted networks the strengths of the two nodes with the same degree are not necessarily the same and vice versa. Therefore, it is believed that the CN with the lower degree and the higher strength would contribute more to the similarity than those of CNs with higher degree and lower strength. Based on this, unit weight of the node is defined to be used to measure the similarity contribution of the node to its neighbors. It should be proportional to the node's strength and inversely proportional to its degree. In other words, the greater the unit weight of the CN is, the more its contribution to the similarity of the two nodes is.

Secondly, when measuring the contribution of weights of the two edges that connect the two nodes to their CN, we think that the greater ratio of the sum of weights of these two edges to the sum of weights of these two nodes is, the higher similarity these two nodes have. Based on this, the effect coefficient of the CN is defined to be used to measure the contribution extent of this CN compared with all neighbors of these two nodes. Moreover, on the basis of the above two definitions, the concept of the joint strength of the neighbor node is proposed, which equals the product of the unit weight of the CN and its effect coefficient. The higher the value is, the more contribution of the CN is.

Finally, in terms of two nodes, we take the sum of the joint strength of all their common neighbors as their total similarity. Here, there is a special situation that the two nodes have no CNs. Then the concept of the edge weight strength of the node pair is introduced as the similarity measurement. It is defined as the ratio of the weight of the edge that connects these two nodes to the sum of weights of all edges that connect with these two nodes. The larger edge weight strength means the two nodes are more closely connected and they have higher similarity.

Based on above definitions, we can fast cluster nodes and their neighbors by calculating the similarity so as to get initial communities. In the expanding phase, as to one of the two nodes in the node pair, if the node having the maximal similarity with the current node is just another one, these two nodes would be clustered together to form a community. If there are many such node pairs in the network, it would form many small communities leading to a lower modularity. So we further optimize the division results by gradually merging communities on condition that the merger can increase the modularity.

2.2 Relevant Definitions

Let $G=(V,E,W)$ represent the undirected and weighted network, where V is the node set, E is the edge

set and W is the set of weights. $\forall x,y \in V$, $\Gamma(x)$ represents the neighbor set of x, e_{xy} represents the edge that connects x and y, and w_{xy} represents the weight of e_{xy}. Let $s(x)$ represent the strength of x, namely, $s(x) = \sum_{z \in \Gamma(x)} w_{xz}$.

Definition 2.1 (Unit weight of the node) $\forall x \in V$, the unit weight of x is defined as the average of weights of all edges connected to the node x, which is denoted by $u(x)$.

$$u(x) = \frac{\sum_{z \in \Gamma(x)} w_{xz}}{|\Gamma(x)|} \tag{1}$$

Definition 2.2 (Effect coefficient of the node) $x,y \in V$, $\forall z \in V \bigcap \Gamma(x) \bigcap \Gamma(y)$, the effect coefficient of z to the node pair <x,y> is defined as the ratio of the sum of w_{xz} and w_{zy} to the sum of weights of all edges connected with x and y, which is denoted by $\varepsilon_z^{CN}(x,y)$.

$$\varepsilon_z^{CN}(x,y) = \frac{w_{xz} + w_{zy}}{s(x) + s(y) - w_{xy}} \tag{2}$$

Definition 2.3 (Joint strength of the common neighbor) $x,y \in V$, $\forall z \in V \bigcap \Gamma(x) \bigcap \Gamma(y)$, the joint strength of z to the node pair <x,y> is defined as the product of the unit weight of z and its effect coefficient to <x,y>, which is denoted by $Sim_z^{CN}(x,y)$.

$$Sim_z^{CN}(x,y) = u(z)\varepsilon_z^{CN}(x,y) \tag{3}$$

Definition 2.4 (Edge weight strength of the node pair) $\forall x,y \in V$, the edge weight strength of the node pair <x,y> is defined as the ratio of w_{xy} to the sum of weights of all edges that connected to x and y. We denote it by $sw(x,y)$.

$$sw(x,y) = \frac{w_{xy}}{s(x) + s(y) - w_{xy}} \tag{4}$$

Definition 2.5 (Weighted similarity based on common neighbors) $\forall x,y \in V$, the weighted similarity of the node pair <x,y> based on their common neighbors is defined as the sum of joint strength of all common neighbors of the two nodes, and we denote it by Sim_{xy}^{IEM}.

$$Sim_{xy}^{IEM} = \begin{cases} 0 & , e_{xy} \notin E \\ sw(x,y) & , e_{xy} \in E \wedge \Gamma(x) \bigcap \Gamma(y) = \Phi \\ \sum_{z \in \Gamma(x) \bigcap \Gamma(y)} Sim_z^{CN}(x,y) & , e_{xy} \in E \wedge \Gamma(x) \bigcap \Gamma(y) \neq \Phi \end{cases} \tag{5}$$

3. IEM Algorithm

3.1 Description of IEM algorithm

Based on the above definitions, the community division algorithm IEM is proposed which mainly

consists of three parts, namely, forming the initial community, expanding the community and merging communities. The description of IEM is as follows.

1) Forming the Initial Community: Calculate the similarity of any node pairs in the network and store them in the matrix. With each node as a community, a node is selected randomly from the network as the starting node and set as the current node. Find the node v_j that has the largest similarity with the current node and merge the community containing v_j with the community containing the current node to form a community. Then take the merged community as the current community.

2) Expanding the Community: Find the node v_k that has the largest similarity with v_j and take v_k as the next node to be clustered. If v_k does not belong to the current community, it means the current initial community has formed. In such a situation, we set v_k as current node and continue to look for the node having the largest similarity with v_k so as to form the next new community. Or else, it means v_k has been clustered into the current community. In such a situation, a node that has not been visited was randomly selected from the network and taken as the current node then another node was also selected to form the next community. Execute above steps repeatedly until all nodes have been visited, which means all initial communities have been expanded.

3) Merging communities: Calculate the modularity of the network. On the basis of the current community structure, calculate the modularity of the corresponding network that would form after merging any two communities. Then initialize the modularity matrix Q. That is to say, the Q_{ij} in Q equals the corresponding modularity of the network that would form after merging the community i and the community j. If max(Q_{ij}) is higher than the modularity of the current network, merge the community i and the community j, and update the community structure of the network. Execute above operations repeatedly until no two communities can be merged, that is, the modularity would not increase no matter which two communities are merged. Then it means the final community structure forms.

3.2 Implementation of IEM Algorithm

The implementation of IEM algorithm is as follows.

Input: G=(V,E,W)
Output: Division results of G, where G={$C_1,C_2,\ldots,C_t$} and $\forall i\neq j$, $C_i\cap C_j=\varnothing$
// **Initializing**
Step 1: ReadFile from Dataset.Txt and Return adjMatrix A;
Step 2: for each node v_i in V do get $\Gamma(v_i)$ endfor;
Step 3: New a SimMatrix S; for i, j=1 to n do S(i,j)= S_{ij}^{IEM} endfor; Return S;
Step 4: New a List SList, for each v_i in V do new ArrayList simlist(v_i);
Step 5: for each v_j in $\Gamma(v_i)$ do get max(S(i,j)); simlist(v_i).add(v_j) endfor;
Step 6: SList.add(v_i,simlist(v_i)) endfor; p=0;
// **Forming the Initial Community**
Step 7: select a node v_i in V randomly where v_i.visited=false; currentnode=v_i;
Step 8: p++;
Step 9: new a List community C_p; C_p.add(currentnode);
// **Expanding**

Step 10: get node_max from SList(currentnode, simlist(currentnode));
Step 11: v_j=node_max; C_p.add(node_max); currentnode=v_j;
Step 12: get node_max from SList(currentnode, simlist(currentnode)); v_k=node_max;
Step 13: if v_k not in C_p, {C_p.add(v_k); currentnode=v_k; goto step 12; }
Step 14: else goto step 7;
Step 15: until all v_i in V, v_i.visited=true; then get G_{current}={C_1,C_2, …,C_p}
// **Merging**
Step 16: Cal $Q_w(G_{\text{current}})$ and Initialize Matrix Q;
Step 17: for i, j=1 to p do Q_{ij}=Q_w(G.Merge(C_i,C_j)) endfor;
Step 18: Return max(Q_{ij});
Step 19: if max(Q_{ij}) > $Q_w(G_{\text{current}})$ {C_i= $C_i \cup C_j$; update G; goto step 16;}
Step 20: else Return G={C_1, C_2,…,C_t}($t \leqslant p$).

4. Experiments and Analysis

Experiments were done on several weighted networks, which show that IEM algorithm is superior to other algorithms for community division in weighted networks with the higher accuracy and relatively lower complexity.

4.1 Datasets

Five real weighted networks were got from the network (http://konect.uni-koblenz.de/networks/) and descriptions of these datasets are as follows.

(1) Zachary's Karate club: It is a relationship network between members of a karate club in a university of America. There are 34 nodes and 78 edges in the network where a node represents a member, an edge represents the close relationship between the two members and the weight represents the close degree of the two members.

(2) Les Misérables: It is a character relationship network originated from the novel of Les Misérables. There are 77 nodes and 254 edges in the network where a node represents a character, an edge represents the appearance of the two characters in the same scene and the weight represents the times they appeared simultaneously.

(3) Madrid Train Bombing: This is a terrorist network in the train bombings in Madrid, Spain in 2004. There are 64 nodes and 243 edges in the network where a node represents a terrorist, an edge represents the cooperation or communication between the two terrorists in train bombings, and the weight represents the frequency of their contact.

(4) US Airport: It is a US air transport network that has 332 nodes and 2,126 edges. In this network a node represents an airport, an edge represents there is a route between the two airports and the weight represents the number of flights between these two airports.

(5) Net Science: This is a network of scientists that published papers cooperatively. There are 379 nodes and 914 edges in the network where a node represents a scientist, an edge represents the two scientists have worked together and the weight represents the number of their cooperation.

4.2 Evaluation Index

Modularity Q is a commonly used standard to evaluate the community division quality of algorithms. For a certain division of the network, the larger modularity always means the more reasonable division of the network. Usually, the value of Q is between 0.3 and 0.7. So many algorithms try to optimize the community division results of the network by maximizing the modularity function.

In the paper, the weighted modularity Q_w [11] is used as an evaluation index. Its definition is as follows.

$$Q_w = \frac{1}{2W}\sum_{ij}(w_{ij} - \frac{w_i w_j}{2W})\delta(C_i, C_j) \tag{6}$$

where, $v_i, v_j \in V$ and w_{ij} represents the weight of e_{ij}.

$w_i = \sum_j w_{ij}$ represents the strength of v_i. $w_j = \sum_i w_{ij}$ represents the strength of v_j.

$W = \sum_{ij} w_{ij}$ represents the sum of the weights of all edges in the network.

$\delta(C_i, C_j)$ is a function. If the node v_i and v_j are in the same community, $\delta(C_i, C_j)$ is 1, or else it equals 0.

4.3 Weighted Similarity Index

In the measurement of the similarity between nodes, the similarity can be defined according to the local attributes of the nodes or the topological information of the network. In general, there are three algorithms based on the similarity. They are similarity algorithms based on the CNs, the node degree and the path of the network. The following is a brief introduction of three classical weighted similarity indexes used in our experimental comparison, namely, weighted CN (written $S_{xy}^{W\_CN}$), weighted AA (written $S_{xy}^{W\_AA}$) and weighted RA (written $S_{xy}^{W\_RA}$), in which S_{xy} represents the similarity between the node v_x and v_y, w_{xy} represents the weight of the edge connecting v_x and v_y, $\Gamma(x)$ represents the set of neighbors of the node v_x and $S(x)$ represents the strength of the node v_x as mentioned above.

$$S_{xy}^{W\_CN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{zy}}{2} \tag{7}$$

$$S_{xy}^{W\_AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{zy}}{\log(1 + s(z))} \tag{8}$$

$$S_{xy}^{W\_RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{s(z)} \tag{9}$$

4.4 Comparison of Experimental Results

In terms of the above datasets, we compared IEM algorithm with three classical similarity indexes as described in [11], namely, weighted CN, weighted AA and weighted RA. We have also done a comparison of IEM with CRMA algorithm. Experimental results were shown in Table 1 where the first column listed the five networks, the second column listed the number of nodes and edges of each network, five algorithms were listed respectively from the third column to the seventh column, the community findings were expressed by the number of the community and the modularity of the network was presented by p/Q_w.

Table 1. Community discovery results of five algorithms on five datasets

Dataset	$\|V\|$ / $\|E\|$	weighted CN	weighted AA	weighted RA	CRMA	IEM
Karate Club	34 / 78	2 / 0.4547	2 / 0.4547	2 / 0.4547	2 / 0.4547	4 / 0.4950
Les Misérables	77 / 254	1 / 0.0350	3 / 0.4185	3 / 0.4577	9 / 0.5222	5 / 0.5427
Train Bombing	64 / 243	1 / 0.0303	4 / 0.3626	4 / 0.3604	4 / 0.4420	5 / 0.4579
US Airport	332/ 2,126	2 / 0.0174	3 / 0.0987	3 / 0.1039	4 / 0.1347	4 / 0.1932
Net Science	379 / 914	8 / 0.6045	19 / 0.8453	18 / 0.8499	21 / 0.8430	19 / 0.8512

(1) As to Karate club network, division results of the first four algorithms were the same where the network was divided into 2 communities as shown in Fig. 1; while IEM algorithm divided this network into 4 communities as shown in Fig. 2. The modularity was improved by 11.11% compared with the other four algorithms. Here, it should be noticed that in all figures of this paper, nodes in different communities were represented by different colors according to division results so as to express the results clearly. Additionally, it should be emphasized that the value of the network modularity and the number of communities of division results will be different in terms of different algorithms and different implementation methods. In general, the larger modularity means the relatively accurate number of communities and the better community structure that is closer to the real network.

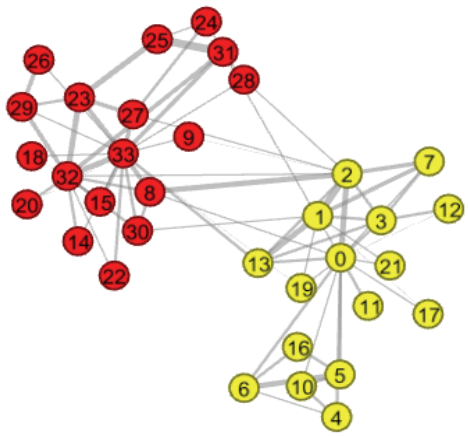

Fig. 1. CRMA for Karate Club.

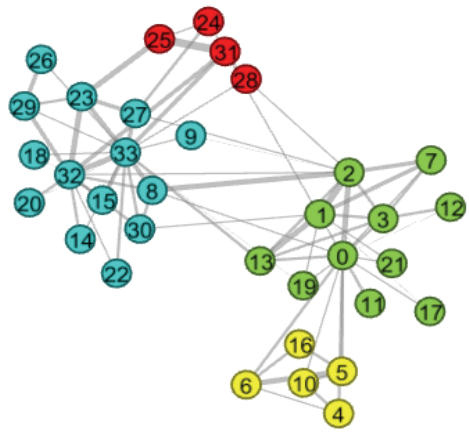

Fig. 2. IEM for Karate Club.

(2) As to Les Misérables network and Train Bombing network, division results of these five algorithms were all different. Among them, the division result of the weighted CN algorithm is the poorest because the weighted CN index only considered the influence of the weight of the edge connecting two nodes and their neighbors on the similarity. However, in the Les Misérables network, the weight represents the times of the two characters' appearance in the same scene simultaneously. And in the Train Bombing network, the weight represents the frequency of the contact of terrorists. Thus, most weights of edges in these two networks are 1, which led to the weighted CN similarity are 1 all the same. So in the clustering, one neighbor would be randomly selected and clustered into the community, which resulted in the deviation of community division. Overall, there are relatively small differences in division results of the latter four algorithms, and the CRMA and IEM algorithm have the relatively better performance. For these two networks, community division results of CRMA and IEM algorithms were shown in Figs. 3–6.

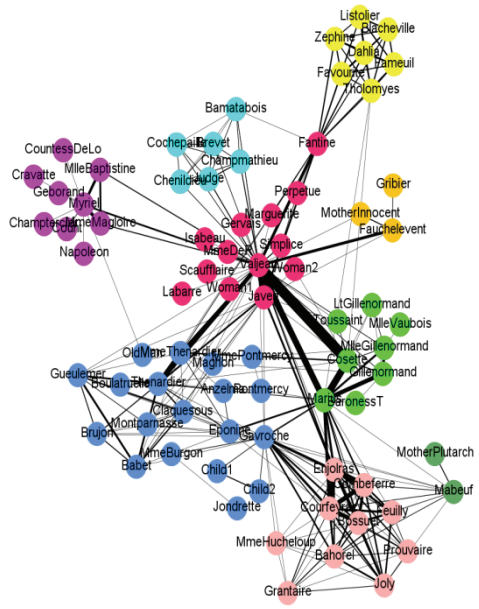

Fig. 3. CRMA for Les Misérables.

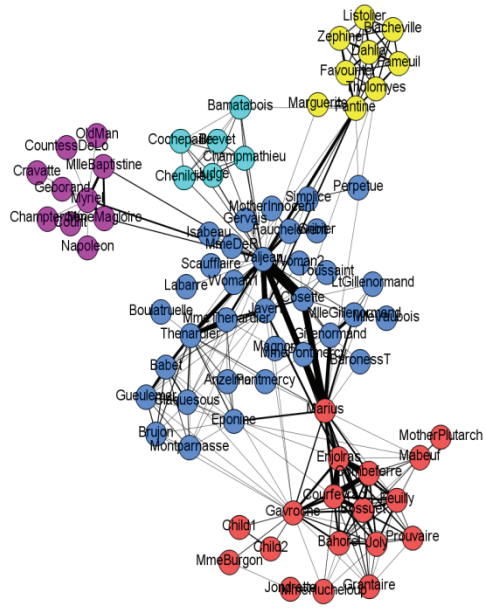

Fig. 4. IEM for Les Misérables.

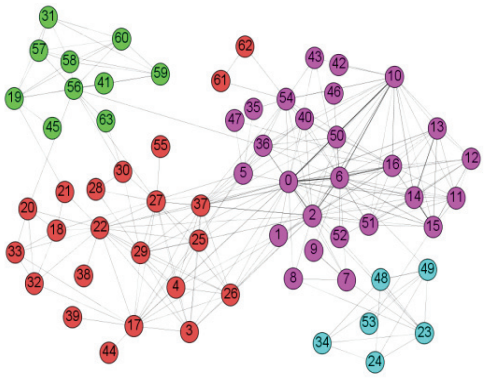

Fig. 5. CRMA for Train Bombing.

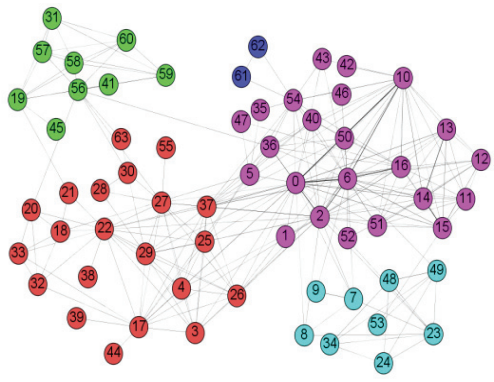

Fig. 6. IEM for Train Bombing.

(3) As to the US Airport network, division results of these five algorithms were all poor. In this network, 59.7% node pairs have no common neighbors. Moreover, among all node pairs with CNs, 46.5% node pairs have only one CN, which led to the lower modularity of each algorithm and poor quality of community division of weighted CN, weighted AA and weighted RA. Because these three similarity indexes just take into account of the influence of the degree or the strength of CNs on the similarity. However, IEM algorithm used the edge weight strength of the node pair to deal with the situation of having no CNs, so its community division quality is the highest. For this network, community division results of CRMA and IEM algorithms were shown in Figs. 7 and 8.

(4) As to the Net Science network, although it is sparse, the average weighted degree of nodes is 2.583, and the average clustering coefficient is about 0.798. So these five algorithms all have better performance on this network. Among them, the performance of the weighted CN is the worst, which had a large difference in the number of communities and the modularity compared with other four algorithms.

While the division results of the latter four algorithms have small differences. Of these, community division results of CRMA and IEM algorithms were shown in Figs. 9 and 10. In general, the community division quality of IEM algorithm is better than the other four algorithms. This further verified the correctness of IEM algorithm which defined the weighted similarity extensively combining with the edge weight, the degree, the intensity and the common neighbors of the two nodes.

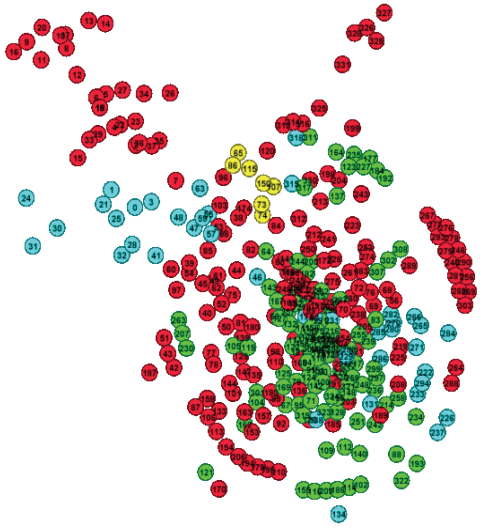

Fig. 7. CRMA for US Airport.

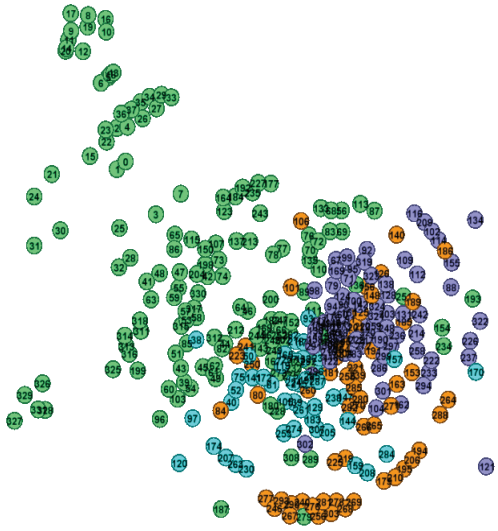

Fig. 8. IEM for US Airport.

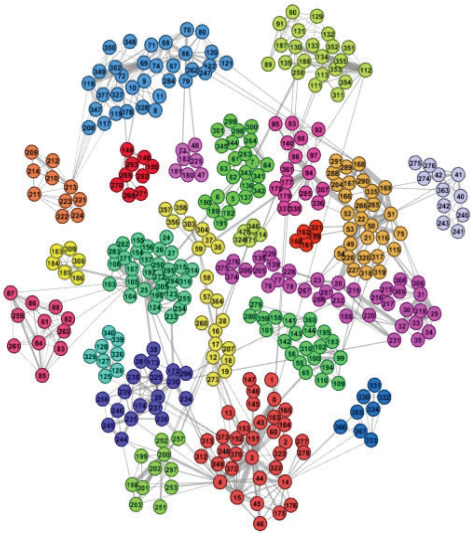

Fig. 9. CRMA for Net Science.

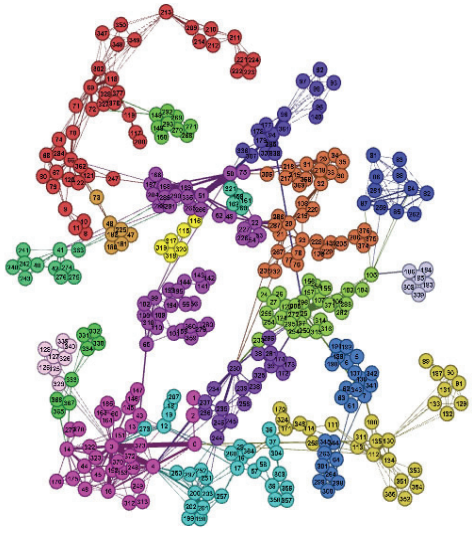

Fig. 10. IEM for Net Science.

From experimental results we know that IEM outperforms the other three weighted similarity indices, which could further verify its correctness and higher division quality. Additionally, IEM is better than CRMA algorithm. CRMA was the improvement of AGMA algorithm which took into

account of the sign of the edge in order to get better division results in signed networks. Though it is still applicable to traditional networks that only have positive links, each algorithm was designed to achieve its underlying goal, which naturally brought different results with regard to different networks. In terms of community division in traditional weighted networks, IEM is more reasonable and effective than CRMA algorithm.

Overall, though community division results of IEM for all these datasets are different from the other four algorithms, the modularity of its divisions are all the highest, which can show its superiority. Moreover, division results of IEM for networks of Karate club, US Airport and Net Science are basically consistent with the fast greedy, betweenness and other classical algorithms described in [13] in the number of communities and the modularity. This can also further verify the correctness of IEM algorithm.

4.5 Complexity Analysis

For the network $G=(V,E,W)$ where $|V|=n$, if we directly use the method based on the optimization of the modularity to cluster the nodes, the complexity would be very high. Therefore, the algorithm proposed in this paper first realized the fast clustering of nodes by defining the weighted similarity index, and then the initial division result can be got which consists of p communities. After that, it optimized the division result by merging communities to maximize the modularity of the network so as to get the best performance.

In IEM algorithm, we first calculate the similarity of any nodes pairs and save these values into the matrix, so the computational complexity is $O(m)$. Secondly, we traverse the matrix and use list (i, arraylist) to store the label of the node that has the highest similarity with node v_i, so the computational complexity is $O(n\log_2 n)$. Lastly, we calculate the modularity of the network after merging any two communities among p communities, so the computational complexity is $O(p^2)$. Compared with other algorithms, the computational complexity of IEM is slightly increased because of the additional computation of merging communities. However, for large scale networks, p is far less than n. So it can also guarantee the feasibility and effectiveness of the time on the premise of achieving a higher accuracy.

5. Conclusions

The existing weighted similarity indices only considered the influence of the weight information of common neighbors on the similarity, which may lead to poor community division results for some special networks. In view of this, a new algorithm IEM was proposed so as to achieve a more reasonable division of weighted networks. The algorithm consisted of three stages, namely, forming the initial community, expanding communities and merging them. In the first two stages, we focused on the influence of common neighbors to the similarity of the two nodes, and the weighted similarity of the two nodes based on the degree, the strength and the weight information of their common neighbors was defined. Moreover, the situation of the two nodes having no CNs was also taken into account, and the edge weight strength was defined as their similarity. Then the most closely related nodes were clustered fast according to their similarity to form the initial community and expand it. In the third stage, for those small communities consisting of only two nodes that may emerge in the first two stages, we merged these

communities by maximizing the weighted modularity of the network, thus the more reasonable and accurate community division results could be got. The weighted similarity index proposed improved weighted CN, weighted AA, and weighted RA. In addition, for the traditional weighted network containing only positive links, the IEM algorithm was more efficient than CRMA algorithm. The experimental results showed its effectiveness and high quality for community division of weighted networks. For large scale networks, how to reduce the computational complexity so as to improve the efficiency of the algorithm is the further research.

Acknowledgement

This paper is supported by Science Foundation for Young Scientists of Northeast Petroleum University (No. 2018QNQ-01) and Heilongjiang Natural Science Foundation (No. LH2019F042).

References

[1] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, article no. 056131, 2004.

[2] K. Subramani, A. Velkov, I. Ntoutsi, P. Kroger, and H. P. Kriegel, "Density-based community detection in social networks," in *Proceedings of 2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, Bangalore, India, 2011, pp. 1-8.

[3] R. Liu, S. Feng, R. Shi, and W. Guo, "Weighted graph clustering for community detection of large social networks," *Procedia Computer Science*, vol. 31, pp. 85-94, 2014.

[4] T. Sharma, "Finding communities in weighted signed social networks," in *Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Turkey, 2012, pp. 978-982.

[5] Z. Lu, Y. Wen, and G. Cao, "Community detection in weighted networks: algorithms and applications," in *Proceedings of 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, San Diego, CA, 2013, pp. 179-184.

[6] K. Wang, G. H. Lv, Z. W. Liang, and M. Y. Ye, "Detecting community in weighted complex network based on similarities," *Journal of Sichuan University (Natural Science Edition)*, vol. 51, no. 6, pp. 1170-1176, 2014.

[7] W. Q. Lin, F. S. Lu, Z. Y. Ding, Q. Y. Wu, B. Zhou, and Y. Jia, "Parallel computing hierarchical community approach based on weighted-graph," *Journal of Software*, vol. 6, no. 23, pp. 1517-1530, 2012.

[8] S. Wang, "Community detection based on the interaction modularity on weighted graphs," Yunnan University, Kunming, China, 2014.

[9] P. Zhan, "Implementation of parallelized method for local community detection in weighted complex networks," *South* China University of Technology, Guangzhou, China, 2013.

[10] J. Zhao and J. An, "Community detection algorithm for directed and weighted network," *Application Research of Computers*, vol. 31, no. 12, pp. 3795-3799, 2014.

[11] Z. Yao, "The analysis and prediction of weighted complex networks," Qingdao Technological University, Qingdao, China, 2012.

[12] J. Guo, M. Liu, L. Liu, and X. Chen, "An improved community discovery algorithm in weighted social networks," *ICIC Express Letters,* vol. 10, no. 1, pp. 35-41, 2016.

[13] X. Liu, "Community structure detection in complex networks via objective function optimization," National University of Defense Technology, Changsha, China, 2012.

Miaomiao Liu https://orcid.org/0000-0002-1667-9519

She was born in 1982 and she is currently an associate professor of Northeast Petroleum University in China. She received the master's degree from Ocean University of China in 2006 and got her doctorate at Yanshan University in 2017. Her main research interests include community discovery and link prediction in social networks.

Jingfeng Guo https://orcid.org/0000-0003-2373-6523

He is currently a professor and Doctoral supervisor of Yanshan University in China. His research interests include database theory, data mining and social network analysis.

Jing Chen https://orcid.org/0000-0003-2373-6523

She is currently an associate professor and Master supervisor of Yanshan University in China. Her research interests include community discovery and information dissemination in social networks.