ICIC Express Letters
Volume **10**, Number **1**, January **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 35–41

AN IMPROVED COMMUNITY DISCOVERY ALGORITHM IN WEIGHTED SOCIAL NETWORKS

Jingfeng Guo[1,3], Miaomiao Liu[1,2,3,*], Linlin Liu[1,3] and Xiao Chen[1,3]

[1]College of Information Science and Engineering
Yanshan University
[3]The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province
No. 438, West Hebei Ave., Qinhuangdao 066004, P. R. China
[*]Corresponding author: liumiaomiao82@163.com

[2]College of Information Science and Engineering
Northeast Petroleum University
No. 199, Fazhan Road, Daqing 163318, P. R. China

Received June 2015; accepted September 2015

Abstract. *The algorithm AGMA (Automatic Graph Mining Algorithm) is improved and a novel community discovery algorithm, namely CRMA (Clustering Re-clustering Merging Algorithm) is proposed which can realize more reasonable community division for weighted social networks. Firstly, in the re-clustering stage, AGMA neglected the clustering situation of neighbors of current node and it did not take the weights of edges connecting the current node and its unclustered neighbors into account, which led to that some nodes cannot be clustered or their clustering was unreasonable. Aiming to this, the concept of the connection compactness between the node and the community is introduced, and then through clustering and re-clustering each node would be divided into the community that has the largest connection compactness with the node according to the node weight, the edge weight and the friend coefficient between the node and the community. Secondly, for community division in some unweighted networks and signed networks, if there are many nodes whose clustered neighbors are less than its unclustered neighbors, the division results of AGMA tend to result in the lower modularity. In view of this problem, the concepts of positive edge weight density, the cluster density and the connection coefficient between clusters are proposed. Lastly, communities are merged based on the above three concepts after re-clustering and the modularity is effectively enhanced. The higher correctness and better generality of the improved algorithm are verified through experiments.*

Keywords: Community discovery, Weighted social networks, Clustering, Signed networks

1. **Introduction.** Community discovery has always been a hot problem in the research of social networks. At present, most algorithms are for unweighted social networks that only contain positive links. However, many real networks are weighted networks and there are always positive and negative interactions between entities in these networks. This paper focuses on the community discovery in weighted networks and its extended application in unweighted networks and signed networks. A weighted social network can be represented as $G = (V, E, W)$, where V is the set of nodes, E is the set of edges and W is the set of edges' weights. w_{ij} represents the weight of the edge connecting v_i and v_j.

Related scholars have done a lot of research on the community discovery in weighted social networks. [1] proposed a framework on the basis of the similarity of interest for community discovery in weighted graphs. [2] proposed a community mining method based on the density. In the paper, twitters were nodes and the number of retweets was the weight. They clustered nodes based on variable density. [3] put forward ABCD (Attractiveness-Based Community Detection) algorithm for clustering of weighted graphs in large social networks based on the attractiveness between communities. [4] proposed AGMA algorithm for community mining in weighted signed networks. However, in the re-clustering stage, the algorithm neglected the situation that all neighbors of the current node have not yet been clustered. Meanwhile, it did not take account of the weight of the edge connecting the current node and its unclustered neighbor, which result in that some nodes cannot be clustered or their clustering was unreasonable. In addition, for some unweighted networks and signed networks, when the number of clustered neighbors of the current node was less than its unclustered neighbors, it will produce many communities consisting of these unclustered nodes according to AGMA, which resulted in the lower modularity.

In view of limitations of these algorithms and problems of AGMA, an improved algorithm CRMA is proposed. Firstly, in the re-clustering, we took overall consideration of the clustering situation of neighbors of the current node, the largest friend coefficient between the current node and the community which its neighbor was clustered into and the largest weight between the node and its unclustered neighbors. Then we clustered each node into the community which had the largest connection compactness with itself based on the size of the largest friend coefficient and the largest weight. Secondly, aiming to the problem of the lower modularity, the concepts of positive edge weight density, the cluster density and the connection coefficient between clusters were proposed. Finally, we merged these communities based on the calculation of these three values. The modularity was effectively enhanced and the community structure was more reasonable. The improved algorithm was also appliable to unweighted networks and signed networks. In the second part, the problems of AGMA are analyzed. The third part is CRMA which describes the relevant definitions and the implementation. The fourth part is experiments. The comparision and analysis on the communitity division results and modularity between the AGMA and CRMA were done which verified the higher accuracy and better generality of the improved algorithm.

2. **AGMA and Problem Analysis.**

The algorithm AGMA [4] had two problems. Problem 1: In the re-clustering, the algorithm directly divided the current node into the community having the largest friend coefficient with it. If all neighbors of the current node have not been clustered, it would result in that the current node cannot be clustered. If part of its neighbors have been clustered while others have not been clustered, and the weight between the current node and its unclustered neighbor was larger than the largest friend coefficient, it would lead to unreasonable clustering of these nodes.

Figure 1 is an artificial weighted network in [5]. In that paper the AOC (Autonomy Oriented Computing) algorithm divided the graph into 7 communities, namely A, B, C, D, E, F and G, shown as dashed ovals in Figure 1. Division results of AGMA were: $\{4, 10, 34, 35, 19\}$; $\{23, 31, 18, 6, 13, \underline{\mathbf{17}}, \underline{\mathbf{25}}\}$; $\{2, 9, 11, 26, 29, 5, \underline{\mathbf{24}}\}$; $\{7, 33, 8, 20\}$; $\{21, 27, 36, 15\}$; $\{14, 16, 28, 30, 32\}$ and unclustered nodes $\underline{\mathbf{1}}$, $\underline{\mathbf{3}}$, $\underline{\mathbf{12}}$ and $\underline{\mathbf{22}}$. From that we know v_1, v_3, v_{12} and v_{22} (shown as square nodes in Figure 1) cannot be clustered finally. The reason was that in the re-clustering, the only neighbor node of v_1, v_3 and v_{22} was v_{12}. However, v_{12} had yet not been clustered. In terms of the situation that all neighbors of the current unclustered node have not been clustered, AGMA did not give any processing. In addition, the clustering of v_{25}, v_{17} and v_{24} (shown as colorless circular nodes in Figure 1) were obviously unreasonable. When clustering v_{17}, its neighbors v_{31} and v_4 have been clustered while its two other neighbors v_{12} and v_{25} have yet not been clustered. AGMA directly divided v_{17} into the community G where its neighbor v_{31} was clustered. However, the weight of the edge connecting v_{17} and v_{25} was 0.33 and it was far greater than the friend coefficient, namely 0.08 between v_{17} and the community G. The similar problems also exist in the clustering of v_{24} and v_{25}.

Problem 2: [4] pointed out AGMA can also be applied to unweighted networks. However, for some networks, if there are more nodes whose number of unclustered neighbors

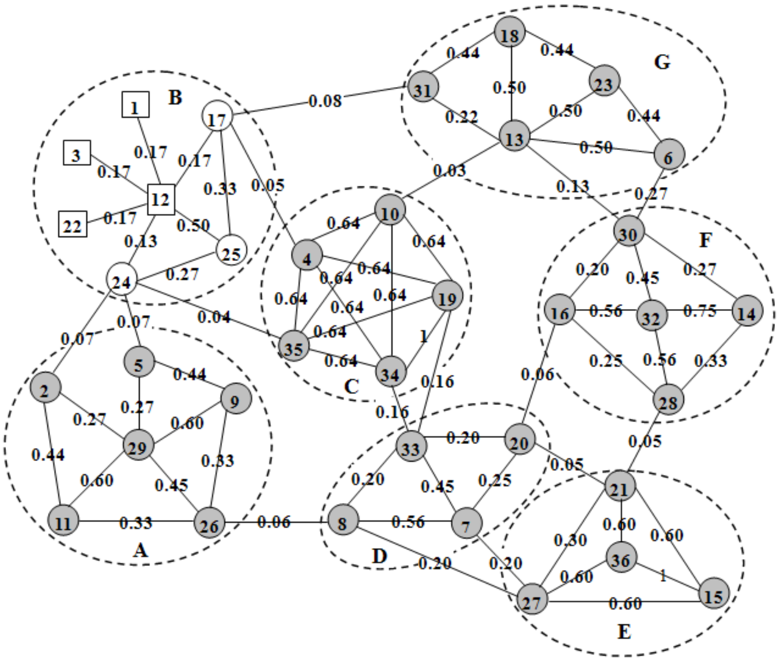


FIGURE 1. An artificial weighted network

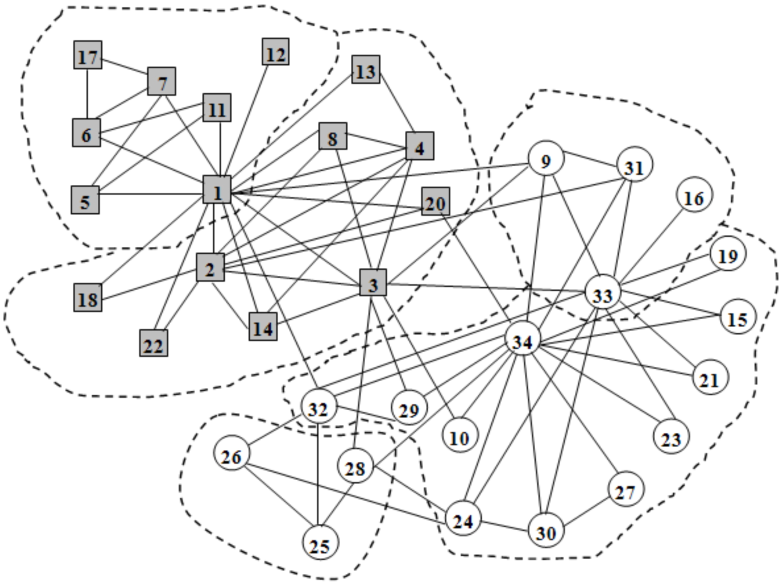


FIGURE 2. Zachary's Karate club network

is larger than its clustered neighbors, AGMA directly clustered the current node with its unclustered neighbors together to form a new community. If there were many nodes of this type in the network, these small communities generated in the clustering would result in the lower modularity.

Zachary's Karate club [6] is an unweighted network consisting of two communities in which nodes were represented by gray squares and colorless circulars respectively in Figure 2. Community division results of AGMA for this dataset were shown in Figure 2 by dotted areas. From that we know AGMA divided the dataset into 5 communities and the final community modularity was relatively lower in comparison with the real community structure of this dataset.

3. CRMA.

3.1. Problem formulation.

$G = (V, E, W)$, $v_i \in V$. Let $Nav(v_i)$ represent the set of neighbors of v_i. Let $v_i.wtcounter$ represent the weight counter of v_i and it equals the sum of weights of edges passing through when traversing v_i. Let lcv and lucv respectively denote the list of clustered and unclustered neighbors of the current node. Let $v_i.ncv$ and $v_i.nucv$ respectively denote the number of clustered and unclustered neighbors. Let $v_i.c$ represent whether v_i has been clustered or not. If v_i has been clustered, the value of $v_i.c$ is 0, or else it equals 1.

Definition 3.1. ***Node Weight***. *Let $G = (V, E, W)$, $\forall v_i \in V$, the weight of v_i is defined as:*

$$v_i.weight = \sum w_{ij} \quad (v_j \in Nav(v_i)) \tag{1}$$

Definition 3.2. ***Network Average Weight***. *Let $G = (V, E, W)$, the average weight of a network is defined as:*

$$avgwt = \sum\nolimits_{i=1}^{|V|} v_i.weight/|V| \quad (v_i \in V) \tag{2}$$

Definition 3.3. ***Friend Coefficient***. *Let $G = (V, E, W)$, the known community $C_p(V_p, E_p, W_p)$ where $V_p \subseteq V \wedge E_p \subseteq E \wedge W_p \subseteq W$. $\forall v_i \in V$, the friend coefficient between v_i and C_p is defined as:*

$$FC_p(v_i) = FC[v_i, C_p] = \sum w_{ij} \quad (v_j \in Nav(v_i) \cap V_p) \tag{3}$$

Definition 3.4. ***Connection Compactness***. *Let $G = (V, E, W)$ and the known community $C_p(V_p, E_p, W_p)$. $\forall v_i \in V$, the connection compactness between v_i and C_p is defined as:*

$$CC_P(v_i) = CC[v_i, C_p] = \frac{\sum w_{ix}}{\sum_{y=1}^{|V|} |w_{iy}|} = \frac{FC_p(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|} \quad (v_x \in V_p \wedge v_y \in V) \tag{4}$$

Definition 3.5. ***Positive Edge Weight Density***. *Let $G = (V, E, W)$, $\forall v_i \in V$, the positive edge weight density of v_i is defined as:*

$$v_i.dense^+ = \frac{\sum w_{ij}}{\frac{1}{2} * \sum_{x=1}^{|N|} \sum_{y=1}^{|N|} w_{xy}} \quad (v_j \in Nav(v_i) \wedge v_x, v_y \in V \wedge w_{ij} > 0 \wedge w_{xy} > 0) \tag{5}$$

Definition 3.6. ***Cluster Density***. *Let $G = (V, E, W)$ and the known cluster C_p. The density of C_p is defined as:*

$$DC_p = \sum v_i.dense^+/|V_p| \quad (v_i \in V_p) \tag{6}$$

Definition 3.7. ***Connection Coefficient***. *Let $G = (V, E, W)$, the known cluster $C_p(V_p, E_p, W_p)$ and $C_q(V_q, E_q, W_q)$. The connection coefficient between C_p and C_q is defined as:*

$$CC_{pq} = \frac{1}{2} * \sum\nolimits_{i=1}^{|V|} \sum\nolimits_{j=1}^{|V|} w_{ij}/(|V_p| * |V_q|) \quad (v_i \in V_p \wedge v_j \in V_q \wedge w_{ij} > 0) \tag{7}$$

Theorem 3.1. *Let $G = (V, E, W)$, the known k clusters $C_1(V_1, E_1, W_1)$, $C_2(V_2, E_2, W_2)$, $\ldots$, $C_k(V_k, E_k, W_k)$ and $v_i \in V \wedge v_i \notin V_1 \cup V_2 \cup \ldots \cup V_k$. $\forall 1 \le t, p \le k$, there is always satisfying $FC_t(v_i) \ge FC_p(v_i)$. Then if we let $C_{\max} = C_t \cup v_i$, it can certainly get the conclusion that $CC_{\max}(v_i) \ge CC_p(v_i)$.*

Proof: $v_i \in V \wedge v_i \notin V_1 \cup V_2 \cup \ldots \cup V_k$. $\forall 1 \le t, p \le k$, $FC_t(v_i) \ge FC_p(v_i)$. Now letting $C_{\max} = C_t \cup v_i$, then we can get $FC_{\max}(v_i) = FC_t(v_i) \ge FC_p(v_i)$ $(\forall 1 \le p \le k \wedge C_p \ne C_{\max})$. From Definition 3.4 we know:

$$CC_{\max}(v_i) = \frac{FC_{\max}(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|} = \frac{FC_t(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|}, \quad CC_p(v_i) = \frac{FC_p(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|}$$

$$\because FC_t(v_i) \geq FC_p(v_i) \text{ and } \sum\nolimits_{y=1}^{|V|} |w_{iy}| > 0 \ \therefore CC_{\max}(v_i) \geq CC_p(v_i).$$

So the theory is correct. On the contrary, if we let $C_{\max} = C_q \cup v_i$ $(1 \leq q \leq k \wedge q \neq t)$ we can also get $CC_{\max}(v_i) \geq CC_p(v_i)$, then we make $C_p = C_t$ and we can get:

$$CC_{\max}(v_i) = CC_q(v_i) \geq CC_p(v_i) \Rightarrow \frac{FC_q(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|} \geq \frac{FC_p(v_i)}{\sum_{y=1}^{|V|} |w_{iy}|}$$
$$\Rightarrow FC_q(v_i) \geq FC_p(v_i) = FC_t(v_i)$$

This is in contradiction with the known condition: $\forall 1 \leq q, t \leq k$, $FC_t(v_i) \geq FC_q(v_i)$. So the assumption does not hold and the theorem is correct. Namely, clustering the node into the community having the largest friend coefficient with it can certainly make sure the connection compactness between the node and this community equals or is larger than connection compactness between the node and any other communities.

3.2. **Implementation of CRMA.** The descriptions of algorithm CRMA are as follows.

```
Input: $G = (V, E, W)$
Output: Community division results of $G$, namely $C_1(V_1, E_1, W_1)$, $C_2(V_2, E_2, W_2)$, $\ldots$, $C_k(V_k, E_k, W_k)$
 1. Initialization: For each $v_i \in V$, initialize $v_i$.visited, $v_i$.weight, etc. Calculate avgwt.
 2. Create a queue $Q$. For each node $v_i \in V$ do
 3. If $v_i$.visited = false then $q$.add($v_i$).
 4. If $q$.isEmpty = false then $q$.remove($v_i$). Start to traverse $v_i$'s every neighbor node $v_j$ and update $v_j$.wtcounter.
 5. Forming lcv and lucv: If $v_j$.wtcounter > avgwt/2 then lcv.add($v_j$) else lucv.add($v_j$).
 6. Clustering: If $v_i$.wtcounter > ($v_i$.weight)/2 and ($v_i$.ncv + $v_i$.nucv) > $v_i$.nav/2 and nucv >= ncv then new cluster $C_p$ and $C_p = v_i \cup$lucv else $C_{\max} = v_i \cup C_{\max}$ where $FC_{\max}(v_i) >= FC_q(v_i)$ $(1 \leq p, q, \max \leq k)$.
 7. Repeat to execute step 3 to step 6 until $Q = \Phi$ and then we get $C_1$, $C_2$, $\ldots$, $C_p$. // Clustering end.
 8. For each node $v_i \in V \wedge v_i$.visited = true $\wedge v_i$.c = 0 do
    Find out $FC_{\max}(v_i)$ $(1 \leq \max \leq k)$ which is the largest friend coefficient between $v_i$ and the existing communities. Find out $e_{ij\max}$ which is the largest weight between $v_i$ and all its unclustered neighbors.
    if $Nav(v_i) \subseteq C_p$ then make $C_p = C_p \cup v_i$ else if $Nav(v_i) \subseteq C_1 \cup C_2 \cup \ldots \cup C_p$ then make $C_{\max} = C_{\max} \cup v_i$ else if $Nav(v_i) - Nav(v_i) \cap (C_1 \cup C_2 \cup \ldots \cup C_p) \neq \Phi$ then if $FC_{\max}(v_i) > w_{ij\max}$ then make $C_{\max} = C_{\max} \cup v_i$ else new cluster $C_{p+1}$ and $C_{p+1} = v_i \cup v_j$ else if $Nav(v_i) \cap (C_1 \cup C_2 \cup \ldots \cup C_p) = \Phi$ then $C_{p+1} = Nav(v_i) \cup v_i$. //Re-clustering end.
 9. For $1 \leq p, q \leq k$ do calculate $DC_p$ and $CC_{pq}$.
    $\forall 1 \leq p \leq k$, $\exists q$ satisfying $CC_{pq} >= DC_p + DC_q$ then let $\Delta = CC_{pq} - DC_p - DC_q$, find $q$ which can make the value of $\Delta$ be the largest, then let $C_p = C_p \cup C_q$.
10. Return to 9 and merge these communities iteratively until any two clusters cannot be merged. //Merging end.
```

4. Experiments and Ananysis.

4.1. **An artificial weighted network.** The community division results of CRMA to Figure 1 were as follows: $\{\{4, 10, 34, 35, 19\}$, $\{23, 31, 18, 13, 6\}$, $\{2, 9, 11, 26, 29, 5\}$, $\{7, 33, 8, 20\}$, $\{14, 16, 28, 30, 32\}$, $\{21, 27, 36, 15\}$ and $\{\mathbf{12}, \mathbf{1}, \mathbf{3}, \underline{\mathbf{25}}, \underline{\mathbf{17}}, \mathbf{22}, \underline{\mathbf{24}}\}\}$. In the re-clustering, v_1 was clustered with its only unclustered neighbor v_{12} together firstly. Then the clustering for nodes v_3, v_{25}, v_{17} and v_{24} was completed successively.

4.2. **Zachary's Karate club network.** In the clustering and re-clustering, community division results of CRMA were the same as that of AGMA. And the modularity was 0.3720. However, CRMA merged these communities in the third stage so the final community structures were consistent with the standard dataset and the modularity was 13% increased. Results showed that the improved algorithm had higher accuracy and rationality according to the evaluation index of the modularity defined in [7].

4.3. **Gahuku-Gama Subtribes network.** As to community division for signed networks, positive links should be within communities and negative links should be between different communities as far as possible [8]. And the frustration [9] is often used as evaluation index of the division results. Gahuku-Gama Subtribes [10] is an unweighted signed network and it consists of 3 communities shown as dotted ellipses in Figure 3. Community division results of AGMA and CRMA were shown in Table 1. The increased modularity and decreased frustration further validated the correctness of the improved algorithm.

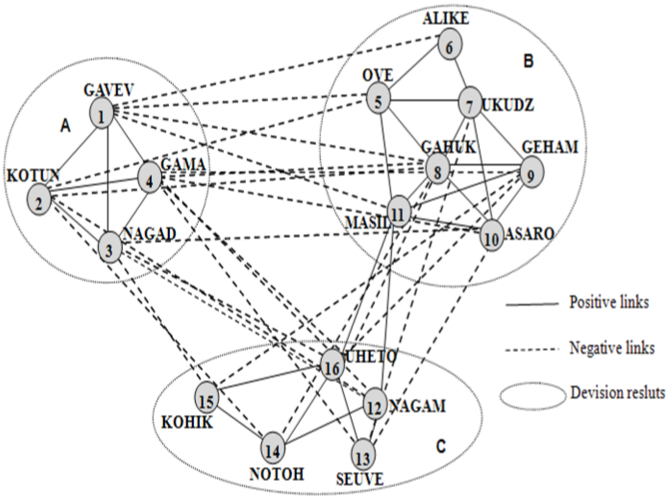
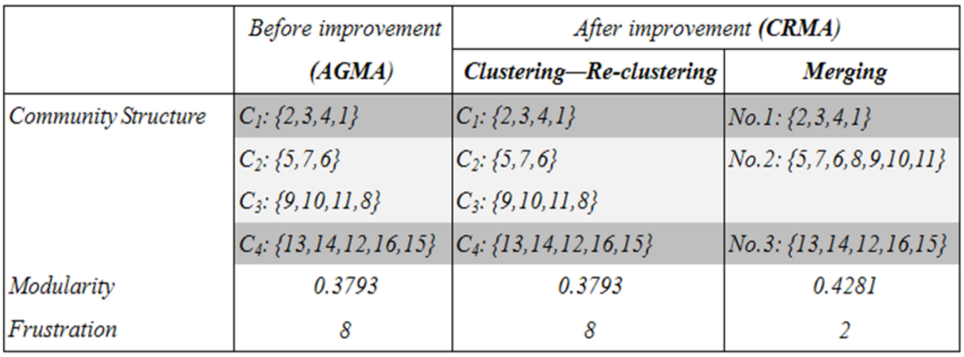


FIGURE 3. Gahuku-Gama Subtribes network

TABLE 1. Community division results to Gahuku-Gama Subtribes network



	Before improvement	*After improvement* ***(CRMA)***	
	(AGMA)	***Clustering—Re-clustering***	***Merging***
Community Structure	*C_1: {2,3,4,1}*	*C_1: {2,3,4,1}*	*No.1: {2,3,4,1}*
	C_2: {5,7,6}	*C_2: {5,7,6}*	*No.2: {5,7,6,8,9,10,11}*
	C_3: {9,10,11,8}	*C_3: {9,10,11,8}*	
	C_4: {13,14,12,16,15}	*C_4: {13,14,12,16,15}*	*No.3: {13,14,12,16,15}*
Modularity	*0.3793*	*0.3793*	*0.4281*
Frustration	*8*	*8*	*2*

5. **Conclusions.** The AGMA algorithm was improved and the algorithm CRMA was proposed which can achieve better community division for weighed social networks by clustering, re-clustering and merging. Firstly, the concept of the connection compactness was introduced and the clustering and re-clustering were done on the basis of the node weight, the edge weight and the friend coefficient. Then the concept of the positive edge weight density was proposed. Lastly, communities were merged based on the density and connection coefficient of these clusters which effectively increased the modularity and reduced the frustration. Results showed the higher accuracy and better rationality of the improved algorithm. How to optimize the efficiency of the algorithm is the next work.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61472340).

REFERENCES

[1] E. Jaho, M. Karaliopoulos and I. Stavrakakis, ISCoDe: A framework for interest similarity-based community detection in social networks, *IEEE Conference on Computer Communications Workshops*, 2011.

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

[3] R. Liu, S. Feng, R. Shi and W. Guo, Weighted graph clustering for community detection of large social networks, *The 2nd International Conference on Information Technology and Quantitative Management*, 2014.

[4] T. Sharma and Lucknow, Finding communities in weighted signed social networks, *International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[5] B. Yang and J. Liu, An autonomy oriented computing (AOC) approach to distributed network community mining, *The 1st International Conference on Self-Adaptive and Self-Organizing Systems*, 2007.

[6] W. W. Zachary and J. Anthropol, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, vol.33, pp.452-473, 1977.

[7] Y. Li, *The Community Detection for Signed Networks Based on the Evolutionary Algorithm*, Ph.D. Thesis, University of Electronic Science and Technology of Xi'an, 2013.

[8] S. Cheng, H. Shen, G. Zhang and X. Cheng, A survey of signed network research, *Journal of Software*, 2013.

[9] P. Anchuri and M. Magdon-Ismail, Communities and balance in signed networks: A spectral approach, *International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[10] M. P. S. Bhatia and P. Gaur, Statistical approach for community mining in social networks, *IEEE/SOLI*, 2008.