Hierarchical Community Detection Algorithm based on Local Similarity

Wei Weng[1], Shunzhi Zhu[1], Huarong Xu[2]
[1]Xiamen University of Technology
361024 Xiamen, China
[2]University of Arizona Tucson, Arizona
CO. 85721, U.S.A.
xmutwei@163.com

ABSTRACT: *Although community, one of the general characteristics of complex networks, has obvious hierarchical structure, in-depth research on its application in the current community detection algorithms is limited. In this paper we present a novel hierarchical community detection algorithm which starts from the node similarity calculation based on local adjacency in networks. Then we find the initial communities and joint them one by one as starting point from the bottom all the way to the top until all nodes fall within the same community. The concept of community similarity is defined for better community incorporation and each time, initial communities with the most similarity are integrated. Taking the advantage of time efficiency in local community mining, this kind of algorithm recognizes the community structure accurately when revealing the hierarchical structure of community, as the experiment shows.*

Categories and Subject Descriptors
I.5.3 [Pattern Recognition]: Clustering - Algorithms

General Terms: Algorithms, Experimentation

Keywords: Complex Network, Community Mining, Similarity

Received: 19 March 2014, Revised 28 April 2014, Accepted 28 April 2014

1. Introduction

Complex network, the abstract diagram of established individual relationships between nature and human society in many intricate systems, can be depicted by graph structure and its two basic elements-vertex and edges, representing members and the relationship between members respectively. Complex community exists everywhere, such as the reference relationship between WebPages, the friendship among microblog users and the e-mail communication correspondence relationship, etc. The basic idea of community mining is to divide the nodes in networks into several sets, where nodes are more closely linked to peers in the same set than those in different sets. As an effective means to detect the nature of complex network, community mining has drawn considerable attentions and enthusiasm from researchers, who have put forward many community mining algorithms up till now, such as overlapping community mining [1, 2] and local community mining [3, 4]. Community hierarchy can also been found everywhere, for example, each university student belongs to different departments and within the departments students are in different majors. In this paper, centering around the hierarchical community mining [5, 6] problem, we find a significant discovery that the combination and evolution process of community members can be detected when disclosing community hierarchy.

Hierarchical community mining has two types of top-down splitting strategy and bottom-up agglomerative strategy [7]. Splitting strategy views the network as a big community and then divides it by removing edges continuously until every node forms an independent community. On the contrary, the agglomerative strategy takes each node as an independent community initially and combines two of the communities together each time as per a certain strategy until all of them are joined together in a single community. Whichever strategy we prefer, the process of hierarchical community mining can be seen as the growing path of hierarchical clustering tree if the community is seen as node, community splitting as nodes to sub-nodes

and community agglomeration as nodes to father nodes. The traditional hierarchical community mining algorithms have a high computation complexity and it is also too time-consuming and insignificant in community division or agglomeration. To overcome these shortcomings, in this paper we introduce a new idea of hierarchical community mining algorithm based on agglomerative strategy. This new algorithm cooperates with the local community mining method to generate the initial communities which then are combined first of all, avoiding the hierarchical clustering tree generating from single node. This proposed algorithm greatly increases the time efficiency by virtue of local community mining and presents a high community mining quality as experiment shows.

2. Relative Researches

Girvan [5] proposed the concept of edge betweenness, which is defined as the amount of the shortest paths going through a certain edge. If there is a community in the complex network, the community can be spitted out by gradually removing these adjacent edges with a higher betweenness value that controls most of the information flow in the network because of its small quantity between communities. This is a kind of top-down hierarchical community mining algorithms. However, such algorithm may not be suitable in large-scale complex network for it requires the information about the entire network when calculating the edge betweeness. Reference [8] proposed an agglomeration approach to view each node as an independent community and then combine the two communities with the highest growth of modularity [7] value as per greedy principle. Nevertheless, this algorithm often falls into local optimum and needs a large amount of computation because its modularity computation also requires the overall network information.

Since the large scale of highly overlapped networks in reality make it difficult to acquire the entire information, community mining based on local information has attracted researchers' attention increasingly. Clauset presented an R-based local community mining algorithm [9]. Firstly, he defined the local modularity R and then added the adjacent nodes that would increase R value to the maximum iteratively until the local community reached the predefined threshold. Similar algorithms can be seen in Reference [3,10].

Our research focuses on the hierarchical community detection algorithm based on local community mining, which is usually neglected though possesses obvious advantage in the hierarchical community mining.

3. Algorithm Description

3.1 Relative definitions

Given complex network $G = (V, E)$, among which $V = \{v_1, v_2, v_m\}$ is the node set and $E = \{e_1, e_2, ..., e_n\}$ is the edge set. Set the weight matrix W and for any $w_{ij} \in W$. If there's a edge between v_i and v_i, $w_{ij} = 1$; otherwise, $w_{ij} = 0$. The degree of v_i can be defined as $K(v_i) = w_{i1} + w_{i2} + ...+ w_{im}$ and the nodes connected to node through edges are called the neighbors of v_i, the set of these neighbors is marked as $N(v_i)$, that is $K(v_i) = \{v_i \mid w_{ij} = 1\}$. The intersection of v_i and $N(v_i)$ is called as the star neighbourhood of v_i, recorded as $St(v_i)$.

1) Similarity calculation of nodes

The similarity of nodes v_1, v_j is calculated by the method adopted in Reference [11]:

$$S = (v_i, v_j) = \frac{\sum_{v_l \in st(v_i) \cap st(v_j)} \frac{1}{K(v_l)}}{\sqrt{\sum_{v_m \in st(v_i)} \frac{1}{K(v_m)}} \sqrt{\sum_{v_n \in st(v_j)} \frac{1}{K(v_n)}}} \quad (1)$$

Formula (1) calculates the node similarity as per the local information and the similarity of nodes v_i and v_j is related with the degree of nodes in their star-neighbourhood intersection. The star neighbourhood is formally defined here because if v_i and v_j, and are connected by an edge, $st(v_i) \cap st(v_j)$ will cover and v_i, v_j and obviously this edge couldn't be neglected when calculating the similarity of nodes (v_i, v_j). The greater degree of nodes in the star-neighbourhood intersection, the less contribution it makes to the similarity of nodes (v_i, v_j), since their role in connecting v_i and v_j are decentralized. Because the denominator plays a role of uniformization, for $S(v_i, v_j) \in [0, 1]$, if $S(v_i, v_j) = 0$, node v_i and node v_j neither connect with each other nor share the same neighbors; if $S(v_i, v_j) = 1$, $St(v_i)$ is equal to $St(v_j)$.

2) Community similarity

Since the hierarchical community mining needs to merge the communities from the bottom up, our integration aimed at those communities with "*highest similarity*", which can be calculated out by the formula (2):

$$S = (C_i, C_j) = \frac{\sum_{v_i \in C_i, v_j \in C_j} S(v_i, v_j)}{sumEdges(C_i, C_j)} \quad (2)$$

$sumEdges(C_i, C_j)$ in formula (2) represents the number of common edges of C_i and C_j. From the formula we can see that the similarity of each two communities is determined by the mean of the similarity of their common edges.

3) Dense vertex set

Although the traditional hierarchical community mining algorithms begin the bottom-up integration from an individual node, our algorithm starts from those "*initial communities*" that are extended from the core pairs. A node pair (u, v) having the highest similarity among their adjacent nodes is called as a dense pair [12], namely, $\sigma(u, v) = \max\{S(x, y) \mid (x = u, y \in N(u)) \vee (x = v, y \in N(v))\}$,

denoted by $u \overset{\varepsilon}{\leftrightarrow} v$, in which $\varepsilon = \sigma(u, v)$. Find all the dense pairs and see each pair as a set, if two sets have the intersection part, combine them until all sets are independent. These remaining sets are called as dense sets.

3.2 Algorithm Process

The leaf node of the hierarchical clustering tree generated by the proposed algorithm represents each of the communities (node sets) instead of nodes. Dense node set is not a leaf node yet because in the complex network some nodes are independent from any dense vertex sets. Therefore, the dense node set should be expanded initially to constitute the initial communities - those leaf nodes of the hierarchical clustering tree. Below is the generating process of initial community:

1) Calculate the similarity of adjacent nodes as per formula (1);

2) Judge and find all those dense pairs, then consolidate those adjacent ones to the dense vertex set (maybe several sets exist). These sets are the prototype of initial communities and through the following steps the final initial communities will be generated.

3) If all nodes are put into the community, turn to step 5), otherwise, take them out one by one and suppose the node taken out is v_i;

4) Find the neighboring node v_j which has the highest similarity with v_i. If v_j remains independent from any communities, go to step (3); if has been in the community, divide v_i into v_j' s community and then turn to step 3);

5) The resulting community $C = C_1, C_2,....C_p$ is the leaf nodes on the hierarchical clustering tree.

Subsequently, the process of community integration is as follows:

1) If C only has one initial community, the algorithm ends, otherwise, calculate the similarity between communities in as per formula (2) and find the community pair with the highest similarity.

2) Merge this community pair and then return to step 1).

4. Experimental Analysis

4.1 Experimental Data

Our experiment is based on the Zachary Karate club network [13] and the American football network [14] that are widely used for testing the effectiveness of community mining. At the initial stage, Zachary Karate club network covers 34 nodes and 78 edges but after a fight, the club members were split into two groups: one is guided by the supervisor and the other is guided by the principal. The American football network, with data from football games among 115 American universities in 2000, consists of 115 nodes and 616 edges, among which node stands for competitor and edge represents game between two teams. It has been divided into 12 communities, representing 12 alliances in the league tournament. The community division of these two complex networks is known and in Table 1 we have listed the standard dividing result.

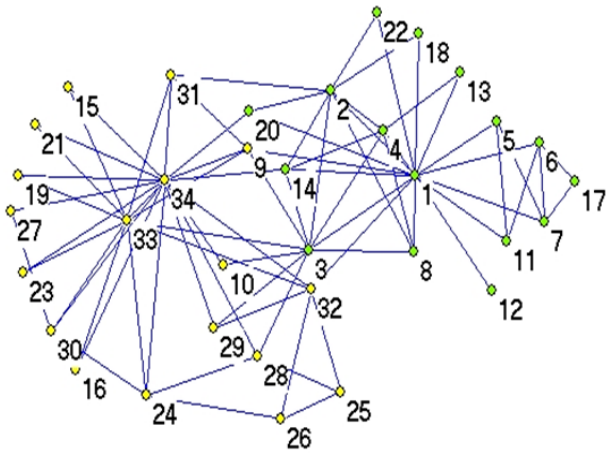

Figure 1. Zachary Karate club network

4.2 Evaluation Criteria

We adopted the result comparison of traditional and standard division to measure community quality of the hierarchical community mining because there is no method described in the literatures we review. Since the Zachary Karate club network is correctly divided into two communities, in our experiment we chose a 2-community hierarchy to compare our hierarchical community mining results with the standard results. Similarly, we chose a 12-community hierarchy in the American football network to for the comparison with standard results. Because hierarchical community mining is the process of bottom-up community integration from leaf nodes, the incorrect bottom combination will mislead the subsequent combinations. On the other hand, we can judge the quality of our hierarchical community mining algorithm by the comparison result and positive result will mean a good overall effect. In this paper, we use the P (precision), R (recall) and $F1$(F-Measure score) that were defined in Reference [10]. Suppose T is the set of node pairs (v_i , v_j) in standard community division (v_i and v_j are in the same community) and S is the set of node pairs in the proposed hierarchical community division (similarly, both nodes are in the same community), then,

$$P = \frac{|S \cap T|}{|S|} \tag{3}$$

$$R = \frac{|S \cap T|}{|S|} \tag{4}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

P is the proportion of the number of correct node pairs in the final dividing scale and R stands for the fraction of correctly classified node pairs in real community. Both P and R reflect the algorithm in one particular aspect, but they have no proportional relationship. In other words, high P doesn't necessarily go along with high R. $F1$ is more

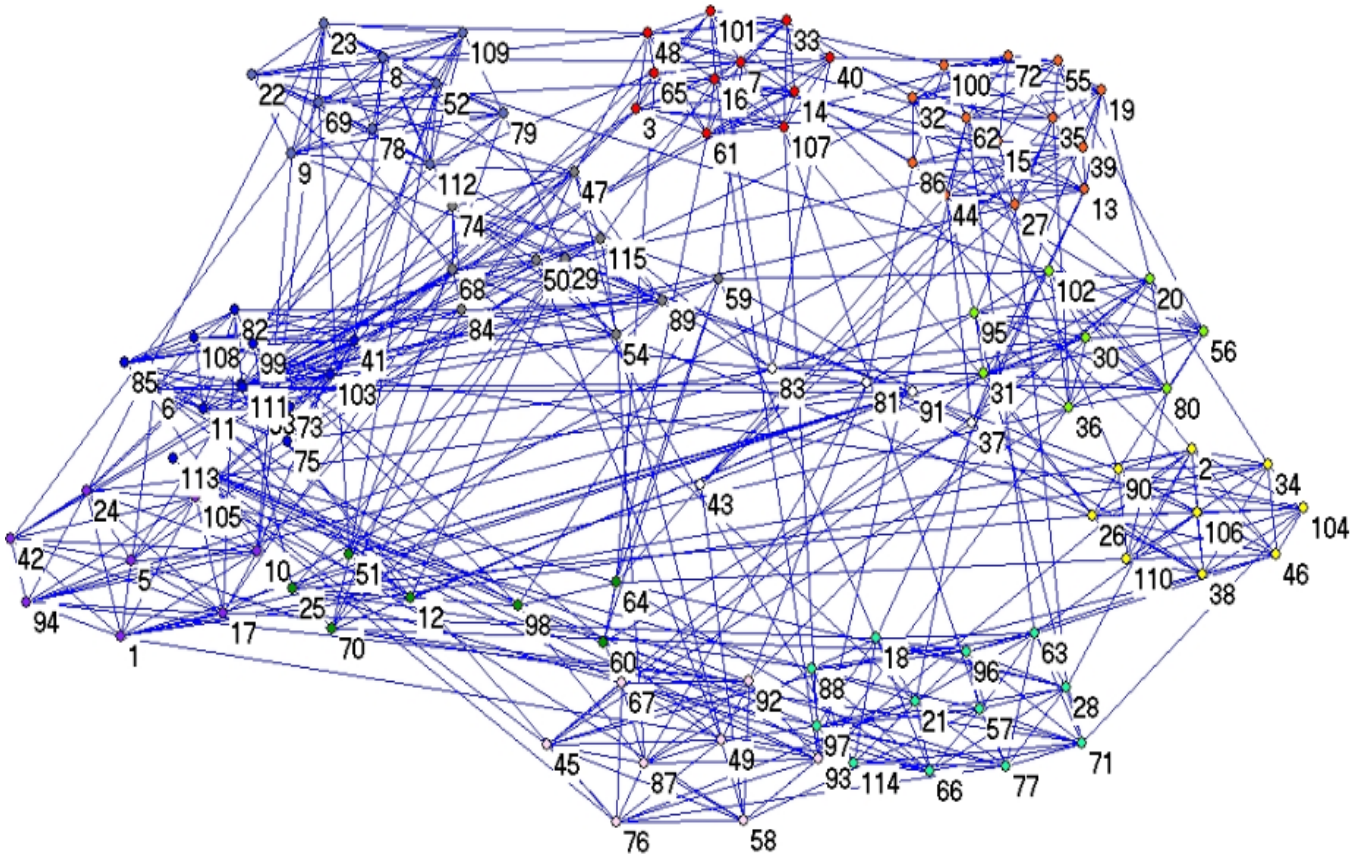

Figure 2. American football network

Dataset	Standard division
Zachary Karate club network	*SK1*:{1, 2, 3, 4, 5, 6,7, 8,11,12,13,14,17,18, 20, 22},*SK2*:{9,10,15,16,19, 21, 23, 24, 25, 26, 27,28,29, 30,31, 32,33,34}
American football network	*SF1*:{2, 26, 34, 38, 46, 90,104,106,110}, *SF2*:{20, 30, 31, 36, 56, 80, 95,102}, *SF3*:{3,7,14,16, 33, 40, 48, 61, 65,101,107}, *SF4*:{4, 6, 11, 41, 53,73,75, 82, 85, 99,103,108,111,113}, *SF5*:{45, 49, 58, 67, 76, 87, 92, 93}, *SF6*:{37, 43, 81, 83, 91}, *SF7*:{13,15, 19, 27, 32, 35, 39, 44, 55, 62, 72, 86, 100}, *SF8*:{1, 5, 10, 17, 24, 42, 94,105}, *SF9*:{8, 9, 22, 23, 52, 69, 78,79,109,112}, *SF10*:{18, 21, 28, 57, 63, 66,71,77, 88, 96, 97,114}, *SF11*:{12, 25, 51, 60, 64,70, 98}, *SF12*{29, 47, 50, 54, 59, 68,74, 84, 89,115}

Table 1. Standard division of two datasets

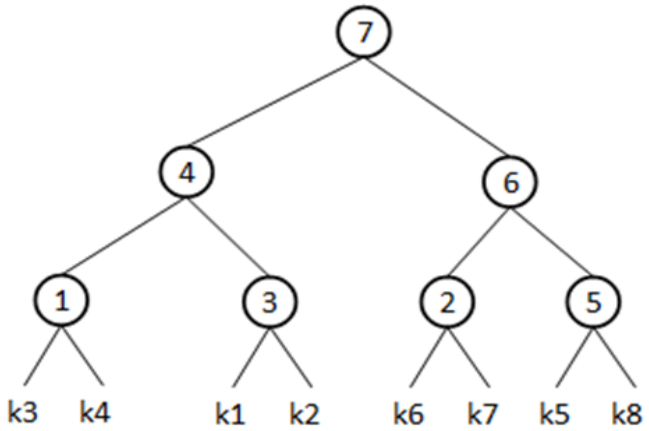

Figure 3. Hierarchical community structure of Zachary Karate club network

reasonable since it is the harmonic value of P and R and only when both P and R are high, can it yield a high result.

4.3 Experimental Analysis

Table 2 reflects the node distribution of initial communities in this algorithm. The Zachary Karate club network has 8 initial communities (K1 ~ K8) and the American football network has 32 initial communities (F1 ~ F32). Bottom-up community integration just starts from the initial communities. It is easy to discover that these initial communities are either the community subsets of standard division or in highly consistence with standard divided communities. It also implies that the community largely consistent with the standard divided ones can be obtained in the process of subsequent community combination. The hierarchical community structures of these two datasets are shown in Fig.3 and Fig.4 respectively, where the letter in the non-leaf nodes represents the merging sequence.

As shown in Table 3 and Table 4, a further comparison on P, R and $F1$ is made between our algorithm and other local community mining algorithms (Clauset-R algorithm[9], LWP algorithm [3] and LMD algorithm [10]).

As can be seen from Table 3 and Table 4, the accuracy of our algorithm is slightly lower than LWP while it is higher than other local community mining algorithms of the same kind. The proposed algorithm is far superior to other algorithms in terms of recall rate except the LMD algorithm.

Evaluation index	*P*	*R*	*F1*
Our algorithm	0.9412	0.9377	0.9394
Clauset-R	0.9088	0.5607	0.6515
LWP	0.9696	0.2757	0.4006
LMD	0.9227	0.6556	0.7368

Table 3. Comparison of Different Algorithm Results on the Zachary Karate Club Network

Note: Test results of Clauset-R, LWP and LWD are collected from Reference [10].

Evaluation index	*P*	*R*	*F1*
Our algorithm	0.8701	0.8832	0.8766
Clauset-R	0.6908	0.7392	0.7082
LWP	0.8908	0.7609	0.8150
LMD	0.8093	0.9061	0.8489

Table 4. Comparison of Different Algorithm Results on U.S. Football Network

Note: Test results of Clauset-R, LWP and LWD are collected from Reference [10].

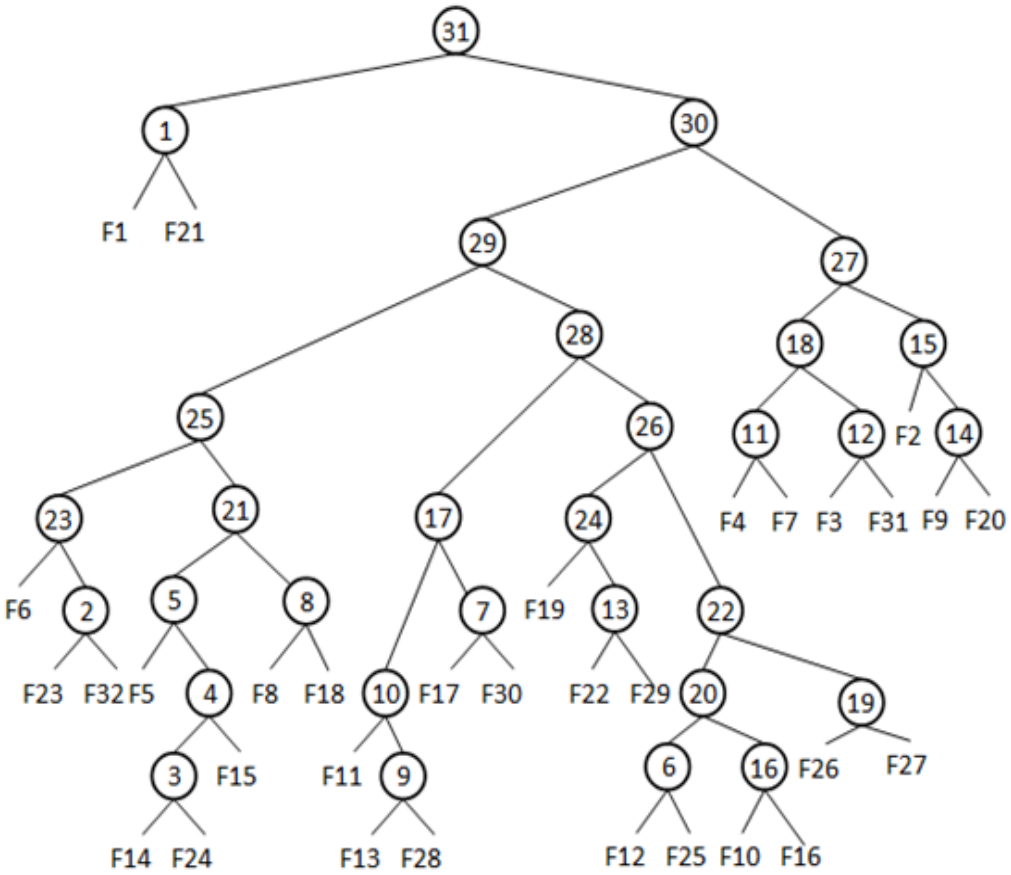

Figure 4. Hierarchical community structure of American football network

Dataset	Initial communities			
Zachary Karate club network	$k1$:{1, 2,12,18,20,22},	$k2$:{3, 4, 8,10,13,14},	$k3$:{5,11},	
	$k4$:{6, 7,17},	$k5$:{9, 31},	$k6$:{25, 26, 28,29,32},	
	$k7$:{24, 27, 30},	$k8$:{15,16,19, 21, 23, 33, 34}		
American football network	$F1$:{1, 25, 33, 89,103,109},	$F2$:{2, 64},	$F3$:{3, 52,74},	$F4$:{5, 84}, F5:{7,77}
	$F6$:{0, 4, 9, 16, 23, 41, 93,104},	$F7$:{10, 81, 98,107},	$F8$:{11, 24, 90},	$F9$:{13, 60,106},
	$F10$:{14, 38, 85},	$F11$:{17,87},	$F12$:{18, 34},	$F13$:{20,113}
	$F14$:{21,111},	$F15$:{8, 22,108},	$F16$:{12, 26, 36, 42, 43},	$F17$:{27, 56},
	$F18$:{28, 50, 69},	$F19$:{19, 29, 30, 35, 55,79, 80, 82, 94,101},	$F20$:{6,15, 32,39, 47,100},	
	$F21$:{37, 45,105},	$F22$:{48,86,91},	$F23$:{49, 53, 83},	$F24$:{51,68,78},
	$F25$:{31, 54, 61, 71, 99},	$F26$:{58,59},	$F27$:{63, 97},	$F28$:{65,96}, F29:{44, 57,66,75, 92,112},
	$F30$:{62, 70, 76, 95},	$F31$:{40, 72,102},	$F32$:{46,67,73,88,110,114}	

Table 2. Initial Communities on Two Datasets

It has a much higher recall rate than that of LMD in Zachary Karate club network but in the network for U.S. football network, it is slightly lower. In regards to $F1$, our algorithm always gets the upper hand in all algorithms of the same kind, especially in the Zachary Karate club network.

5. Conclusion

Local algorithm is computationally efficient when dealing with large and complex network, the algorithm we proposed is thus applicable to the complex networks involving tens of thousands of nodes. Various experimental results demonstrate that the quality of the community structure detected by our algorithm is significantly superior to other local community mining algorithms when disclosing the hierarchical structure of community. We will be committed to the popularization of the proposed hierarchical community detection algorithm based on local similarity in the weighted complex network and directed complex network that are widespread in reality. In addition, we will combine the linear time algorithms [15] to reduce the time complexity.

6. Acknowledgment

This work is partially supported by General Programs of Natural Science Foundation of China (61373147), the Xiamen University of Technology's International Cooperation and Exchange Project under Grant Nos. E201301300 and E201300200, the Fujian Province Department of Education Category A Projects under Grant Nos. JA13238 and the Project of Xiamen Science and Technology Program under Grant Nos. 3502Z20133041.

References

[1] Zhang, S., Wang, R. S., Zhang, X. S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physical A: Statistical Mechanics and its Applications*, 374 (1) 483-490.

[2] Palla, G., Derenyi, I., Farkas, I. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 (7043) 814-818.

[3] Luo, F., Wang, J. Z., Promislow, E. (2008). Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6 (4) 387-400.

[4] Chen, Q., Wu, T. T., Fang, M. (2013). Detecting local community structures in complex networks based on local degree central nodes. *Physica A: Statistical Mechanics and Its Applications*, 392 (3) 529-537.

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

[6] Ravasz, E., Baraba´si, A. L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67 (2) 026112,1-7.

[7] Newman, M. E. J., Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review*, 69 (2) 026113, 1-16.

[8] Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review*, 69 (6) 066133,1-5.

[9] Clauset, A. (2005). Finding local community structure in networks. *Physical Review*, 72 (2) 026132

[10] Chen, Q., Wu, T. T., Fang, M. (2013). Detecting local community structures in complex networks based on local degree central nodes. *Physica A: Statistical Mechanics and Its Applications*, 392 (3) 529-537.

[11] Liu, Xu., Yi Dong-Yun. (2011). Complex network community detection by local similarity. *ACTA AUTOMATICA SINICA*, 37 (12) 1520-1529. (in Chinese)

[12] Huang J., Sun H., Han J., (2010). SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks. *In*: Proceedings of the 19th ACM international conference on Information and knowledge management, p. 219-228.

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

[14] Girvan, M., Newman M.E.J. (2002). Community structure in social and biological networks. *In*: Proceedings of National Academy of Sciences of the United States of America, 99 (12) 7821-7826.

[15] Weng Wei, Zhang Nian, XU Huarong. (2014). Community Mining Method of Label propagation based on dense pairs. *Journal of Engineering Science and Technology Review*. 7 (3) 76-85. A. -L.