A New Vertex Similarity Metric for Community Discovery: A Local Flow Model

Yueping Li, Yunming Ye and Xiaolin Du
Shenzhen Graduate School, Harbin Institute of Technology
Shenzhen 518055, China
Email: leeyueping@gmail.com, yym@hitsz.edu.cn, hermionedoo@vip.qq.com

***Abstract*—The hierarchical clustering methods based on vertex similarity have the advantage that global evaluation can be incorporated for community discovery. Vertex similarity metric is the most important part of these methods. However, the existing methods do not perform well for community discovery compared with the state-of-the-art algorithms. In this paper, we propose a new vertex similarity metric based on local flow model, called Local Flow Metric (LFM), for community discovery. LFM considers both the number of connecting paths and the local edge density which are essential measures in community structure. Compared with the existing metrics of vertex similarity, LFM outperforms substantially in community discovery quality and the computing time. Furthermore, our LFM algorithm is superior to the state-of-the-art algorithms in some aspects.**

***Index Terms*—hierarchical clustering, vertex similarity, community discovery, network flow**

I. INTRODUCTION

Many systems of current interest can be represented as graphs. Each of these graphs consists of vertices and their connecting edges, where the vertices indicate the individuals. Recent studies [1] suggest that many graphs in society and technology often exhibit hierarchical community structure. It is found that the communities correspond to some known sets of units dealing with related topics, such as citation networks [2], foob webs [3], and biochemical networks [4][5].

The problem of community discovery plays an important role for the identification and characterization of real networks [6]. Furthermore, the discovery of hierarchical community structure emerges as an essential task for capturing an in-depth understanding of networks.

Community discovery of graphs has been well studied. Early approaches include the Kernighan-Lin algorithm [7], spectral partitioning [8][9], or hierarchical clustering [10]. There are also many other kinds of methods based on different technologies such as spectral property of graph matrix [11][12][13], spin-spin interactions [14], random walks [15] and synchronization[16]. For more details, the reader can refer to the survey article by Fortunato [17].

The algorithms based on the vertex similarity compose a well-known branch of hierarchical clustering methods. Vertex similarity has been widely used in hierarchical clustering. However, there are just several metrics of vertex similarity in this field. In addition, the existing ones are not good for community discovery, for instance, the time complexity and the quality are not satisfied.

In this paper, we propose a new vertex similarity metric for community discovery. The metric is based on local flow model which enables it to evaluate both topological distance and local edge density. Thus, it can describe the similarity between two vertices better.

We implement the algorithm which employs the new metric in Java. Then we apply it to several well-known datasets, and compare it with the algorithms using existing vertex similarity metrics. In addition, we also compare our algorithm with the state-of-the-art algorithms: Girvan-Newman algorithm [18] and the improved Newman algorithm (CNM algorithm) [19]. The results show that our metric is better than the existing ones, and our algorithm is superior to the state-of-the-art algorithms in some aspects.

The rest of this paper is organized as follows. Section II formulates our problem, and proposes the measure of quality. Section III introduces existing metrics of vertex similarity. Our metric and algorithm are given in Sections IV and V. Experimental results are presented in Section VI. In Section VII, we summarize this work and point out the future work.

II. PROBLEM STATEMENT

Community structure has no universal accepted definition [1]. A common used one is that the division of vertices into groups such that there is a higher density of edges within groups than the edge density between them. In this paper, we consider simple graphs only, i.e., the graphs without self-loops or multi-edges. Given graph G, $V(G)$ and $E(G)$ denote the sets of its vertices and edges respectively. A community structure is a partition $P = C_1, C_2, .., C_k$ of graph G such that $C_1 \cup C_2 \cup .. \cup C_k = V(G)$ and $C_i \cup C_j = \varnothing$ for $i \neq j$.

Modularity of community structure is a quantitative measure of the quality of the partition [19]. It can be employed to evaluate the quality of different partitions of the same graph. Definition of modularity given below states that communities in a good partition have dense intra-community edges and less inter-community edges:

$$Q(P) = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j) \quad (1)$$

doi:10.4304/jsw.6.8.1545-1553

where A_{ij} is the adjacency matrix, m is the total number of the edges, and k_i is the degree of vertex i. The function δ yields one if vertices i and j are in the same community ($C_i = C_j$) while δ yields zero if vertices i and j belongs to different communities.

The task of our problem is to find an optimal partition P which makes the modularity $Q(P)$ maximum.

It is well know that our problem is an NP-hard problem [17]. Thus, there is no polynomial time algorithm for this problem unless $P = NP$. Most existing methods are approximation algorithms. In addition, the measure modularity has limits [17]. Thus, the resulting community structure will be favorable if it corresponds to the actual structure in real world.

III. Existing Metrics of Vertex Similarity

Hierarchical clustering methods have a main branch which contains a series of algorithms based on vertex similarity. The idea of these methods is to compute the similarity between each pair of vertices, firstly, no matter whether they are connected by an edge or not. Then, merge the vertex or the (temporary) community into the vertex or community most similar to it. Similarity metrics are the basis of these methods.

However, it appears that these algorithms perform well for specific types of problems, but work poorly in more general cases [20]. The reason is that existing vertex similarity metrics are designed for particular kinds of graphs. That is, these metrics are just well defined with respect to specific kinds of graphs. Thus, the algorithms based on these metrics cannot tackle a variety of graphs.

One common used vertex similarity metric for community discovery is the overlapping part between the neighborhoods of the vertices i and j, given by the ratio between the inter-section and the union of the neighborhoods.

The formula is as follows:

$$s_{NR}(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \tag{2}$$

where $\Gamma(i)$ is the neighbor set of vertex i.

Another metric is the number of independent paths between two vertices. Independent paths do not share any edge (vertex), and their number is related to the maximum flow that can be conveyed between the two vertices. It can be computed under the constraint that each edge can carry only one unit of flow.

An alternate metric considers all paths running between two vertices. In this case, there is one problem that the total number of paths is infinite, but this can be avoided if one performs a weighted sum of the number of paths, where paths of length l are weighted by the factor α^l, with $\alpha < 1$.

In brief, the above metrics consider either edge density in neighbors or the number of connecting paths. Since the community structure is affected by the local edge density and the number of connecting paths in global, these metrics are not good enough for community discovery.

IV. A New Vertex Similarity Metric Based on Local Flow Model

A. Definition and Properties

Our local flow model aims to compute the similarity metric in small part of the graph taking edge density into account. The metric is evaluated by a max network flow of the (local) subgraph induced by a small diameter of the considering vertex. Furthermore, for each edge of the subgraph, its capacity is determined by its distance from the considering vertex. The capacity decreases as the distance increases.

Let $G = (V, E)$ be a simple graph. Assume that two vertices i and j are supposed to compute similarity. Let $dis(x, y)$ be the distance between vertices x and y where $x,y \in V(G)$. Set $d_e(i) = \min\{dis(a,i), dis(b,i)\}$ where the edge $e = (a,b)$ and $a,b,i \in V(G)$.

Since network flow is employed, we introduce related definitions and formulas. A flow network is a real function $f : V(G) \times V(G) \to \Re$ such that the flow along any edge can not exceed its capacity. Hence, we show the determination of the capacities of edges. The maximum capacity of edge is $Cap_{\max}$. Let decreasing factor α be a small constant such that $0 < \alpha < 1$, and L_{rad} be a threshold which is a small positive integer. For each edge $e \in E(G)$ satisfies $d_e(i) \le L_{rad}$ or $d_e(j) \le L_{rad}$, the capacity e_{cap} is assigned with

$$Cap_{\max} \times \alpha^{\min\{d_e(i), d_e(j)\}} \tag{3}$$

The capacities of other edges are assigned with 0.

The value of flow is defined by $|f(i,j)| = \sum_{v \in V(G)} f(i,v)$, where i is the source (considering vertex). This value represents the amount of flow passing from the source i to the sink j.

Then, we define the similarity metric $s_{LFM} = (i, j)$ equals the maximum flow from i to j. We denote the value of the maximum flow by $F_{\max} = (i, j)$. An illustration example is given as follows:

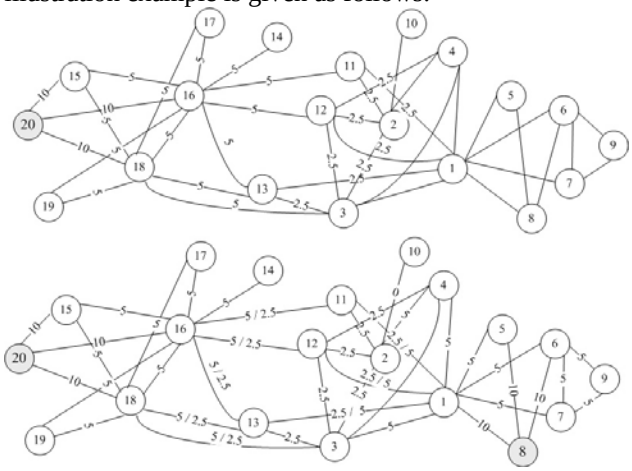


Figure 1. An example shows the capacities of edges when computing the similarity between vertex 20 and vertex 8.

TABLE I.
THE MAXIMUM LOCAL FLOW

Edge	(20,16)	(16,11)	(16,12)	(11,1)	(12,4)
Flow	5	2.5	2.5	2.5	2.5

Edge	(4,1)	(1,5)	(5,8)	(20,18)	(18,13)
Flow	2.5	5	5	7.5	5

Edge	(18,3)	(13,1)	(13,3)	(3,1)	(1,8)
Flow	2.5	2.5	2.5	5	7.5

The graph at the top of Fig. 1 shows the capacities when the vertex 20 is considered. The bottom graph illustrates the situation that the vertex 8 is also taken into account. The denotation $\frac{a}{b}$ of capacity points out that a is the capacity with respect to vertex 20 while b is that for vertex 8. We have the result that $F_{\max}(20,8)$ is 12.5. The flows of related edges are given in Table I.

The computation of our metric is symmetric. It is guaranteed by the following proposition.

Proposition 1. The value of the maximum flow from vertex i to vertex j is the same as the value of the maximum flow from vertex j to vertex i. Equivalently, $F_{\max}(i,j) = F_{\max}(j,i)$.

Proof : Without loss of generality, suppose $F_{\max}(i,j) > F_{\max}(j,i)$. Let flow f be the maximum flow outputs $F_{\max}(i,j)$. We construct a new flow f' as follows: $f'(u,v) = f(v,u)$ for any $u,v \in V(G)$. Then, we have $\sum_{v \in V(G)} f'(v,i) = -\sum_{v \in V(G)} f'(i,v) = F_{\max}(i,j)$. Based on the property of network flow, we have $\sum_{v \in V(G)} f'(j,u) = -\sum_{v \in V(G)} f'(v,i) = F_{\max}(i,j)$ where j is the source and i is the sink of flow f'. According to the determination of edge capacity, the capacities of all the edge of graph G are the same when computing $F_{\max}(i,j)$ and $F_{\max}(j,i)$. Thus, the flow in any edge is below the capacity with respect to flow f'. Therefore, flow f' is a flow with bigger value $F_{\max}(i,j)$ than the maximum flow with value $F_{\max}(j,i)$. A contradiction!

We give the necessary theorem for next proposition, first.

Theorem 1 (Max-flow min-cut theorem) If f is a flow in given graph $G(V,E)$ with source s and sink t, then the following conditions are equivalent:

1. $f(s,t)$ is a maximum flow in G;
2. $|f(s,t)| = \sum_{e=(a,b), a \in S, b \in T} e_{cap}(S,T)$

where (S,T) is a cut of G such that $s \in S$ and $t \in T$.

Let $|P_{(i,j)}(d)|$ be the number of the paths between vertices i and j in which the length of each path is d. Then we have the following proposition which indicates the relationship between local edge density and $F_{\max}(i,j)$.

Proposition 2. Let L_{rad} be the chosen threshold, $Cap_{\max}$ be the max capacity, and α be the constant of decreasing factor when computing the $F_{\max}(i,j)$ in graph G where $i, j \in V(G)$. Then, we have

$$\sum_{d \le 2 \times L_{rad}} (|P_{(i,j)}(d)| \times Cap_{\max} \times \alpha^{d/2}) \ge F_{\max}(i,j) \quad (4)$$

and

$$|\{e \in E(G) \mid d_e(i) \le L_{rad} \vee d_e(j) \le L_{rad}\}| \ge \sum_{d \le 2 \times L_{rad}} |P_{(i,j)}(d)| \quad (5)$$

Proof: By Theorem 1, we can conclude that there is an (S,T) cut satisfying $F_{\max}(i,j) = \sum_{e=(a,b), a \in S, b \in T} e_{cap}(S,T)$ where $i \in S$ and $j \in T$.

Since for any edge $e=(a,b)$ such that $a \in S$ and $b \in T$, the edge e lies in a path between i and j, and the length of the path is less than $2 \times L_{rad}$. Then, Formula (4) holds.

If edge e is an edge of a path from i to j, then we have $d_e(i) \le L_{rad}$ or $d_e(j) \le L_{rad}$. Since one path has at least one edge different from other paths, Formula (5) can be concluded.

B. Comparison with Existing Vertex Similarity Metrics

In this subsection, we compare our metric with the existing ones on several pairs of the graph modeled by Zachary's karate club illustrated in Fig. 2.

Denote the metric of the number of independent paths by s_{NIP}, and the weighted sum of the number of paths by s_{WS}. Choose the weight factor to be 0.5 when computing the weighted sum. For our local flow model, L_{rad} is set to 4, $Cap_{\max}$ is 100 and decreasing factor $\alpha = 0.2$. We select a couple of pairs to vertices, and the values are given in Table II.

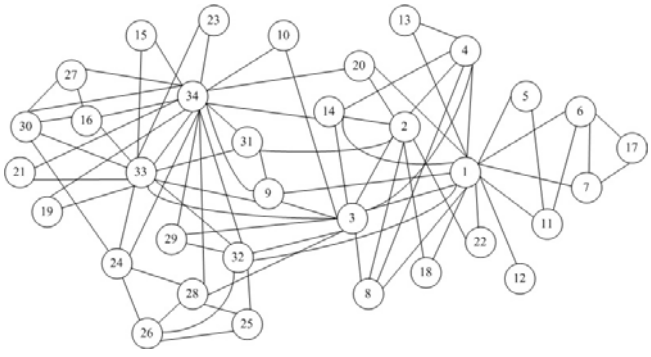


Figure 2. Zachary's karate club.

TABLE II.
METRICS COMPARISON IN KARATE CLUB GRAPH

Edge	(20,16)	(16,11)	(16,12)	(11,1)	(12,4)
Flow	5	2.5	2.5	2.5	2.5

Edge	(4,1)	(1,5)	(5,8)	(20,18)	(18,13)
Flow	2.5	5	5	7.5	5

Discussion: The metric of neighbor ratio s_{NR} cannot distinguish the similarity of pairs (6,2) and (6,30), since they have no common neighbors. In addition, it cannot describe the topology distance of a pair of vertices. The metric s_{NIP} cannot evaluate the topology distance either, concluded by the values of $s_{NIP}(6,2)$ and $s_{NIP}(6,30)$. The metric s_{WS} outputs wrong evaluation of pairs (1,12) and (2,12). The reason is that there are more paths from vertex 2 to vertex 12 than that from vertex 1 to vertex 12. The values indicate that our local flow metric is better in describing the similarity compared with the existing ones.

V. ALGORITHM DESCRIPTION

Community Discovery Algorithm Based on Local Flow Model

Input: a simple, undirected and unweighted graph G
Output: a community structure

1. Choose the diameter L_{rad}, the maximum capacity $Cap_{\max}$ and the decreasing factor α to be a small positive integer.
2. For each vertex pair (i, j) where $i \neq j$ and $dis(i, j) \leq 2 \times L_{rad}$ Do
3. Begin
   set the capacities with respect to vertex i and vertex j, respectively
4. Compute s_{LFM} by means of maximum flow
5. Clear the capacities set in Step 2
6. End //foreach
7. For each vertex pair (i, j) where $i \neq j$ and $dis(i, j) > 2 \times L_{rad}$ Do
8. Set $s_{LFM} := 0$;
9. Let Max be the maximum value of $s_{LFM}(i, j)$, for all $i, j \in V(G)$ and $i \neq j$
10. For each vertex pair (i, j) where $i \neq j$ and $dis(i, j) \leq 2 \times L_{rad}$ Do
11. Set $s_{LFM} := \frac{Max - s_{LFM}(i, j)}{Max}$;
12. Use the classical average linkage methods to find the community structure and output it.

Complexity analysis: Steps 2-6 employ the algorithm to search a maximum flow. Thus, the computing time is bounded by $(2 \times |V(G)| \times L_{rad}) \times (|E(G)| \times Cap_{\max}) \in O(|V(G)||E(G)|)$. Similarly, we conclude that the running time of steps 10-11 is $O(|V(G)| \times L_{rad})$. Thus, the computation of metrics needs $O(|V(G)||E(G)|)$ time. It is known that the time complexity of average link methods is $O(|V(G)|^3)$. Hence, the total time complexity of our algorithm is $O(|V(G)|^3)$.

A. Hierarchical Community Extraction Algorithm

In this subsection, we describe Step 12 in our algorithm. That is, given a set of vertices, once we obtain the quantity of the similarity between them (in the form of similarity matrix), a widely used method to uncover hierarchical organization is hierarchical clustering algorithm [22]. Usually, five different measures named 'single linkage clustering', 'complete linkage clustering', 'average linkage clustering', 'centroid linkage clustering', and 'ward linkage clustering' can be employed to define cluster-to-cluster similarity. We next present the definition of these five criteria.

Single linkage (*SL*): uses the smallest distance between nodes in the two clusters.

$$d_{SL}(R,S) = \min\{d(s,r)\}, s \in S, r \in R \quad (6)$$

Complete linkage (*CL*): uses the largest distance between nodes in the two clusters.

$$d_{CL}(R,S) = \max\{d(s,r)\}, s \in S, r \in R \quad (7)$$

Average linkage (*AL*): uses the average distance between nodes in the two clusters.

$$d_{AL}(R,S) = \min\{d(s,r)\}, s \in S, r \in R \quad (8)$$

Centroid linkage (*CenL*): uses the distance between the centroids of the two clusters.

$$d_{CenL}(R,S) = d(\frac{1}{n_R}\sum_{1}^{n_R} r, \frac{1}{n_S}\sum_{1}^{n_S} s), s \in S, r \in R \quad (9)$$

Ward linkage(*WL*): uses the incremental sum of squares; that is, the increase in the total with-in cluster sum of squares as a result of joining cluster R and S.

Our paper employs the average linkage clustering algorithm as it offers a compromise between the single linkage and complete linkage and is more robust [23].

VI. EXPERIMENTAL RESULTS

We implement the algorithm in Section V and perform experiments in several well-known datasets including two computer generated networks and 4 social networks in real world.

A. Computer Generated Networks

Firstly, we focus on the computer generated networks with nested hierarchical structure that were proposed by Sales-Pardo et al. [6], which is an extension of the benchmark presented by Girvan and Newman [21]. This network is designed to have two levels. For example, a network with 512 nodes has four modules at the first level comprising 128 nodes each. Each of these four modules will comprise four sub-modules with 32 nodes each.

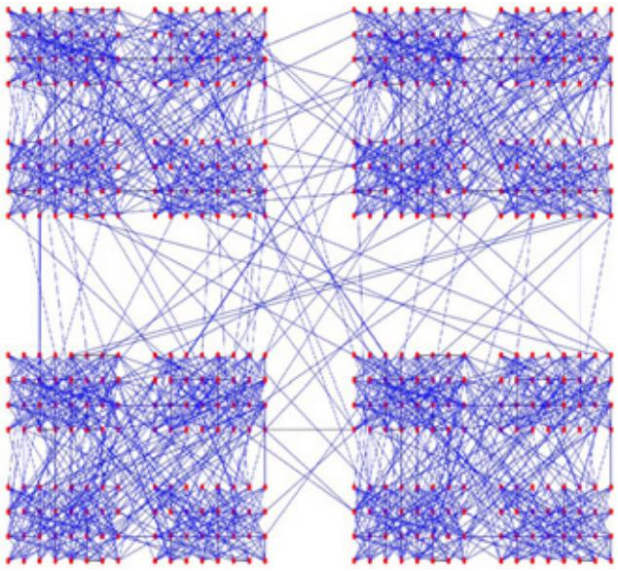


Figure 3. Hierarchical network of 512 nodes.

The edge set is determined as follows: an edge between a pair of nodes (i, j) with probability p_2, if (i, j) are in the same module at the second level; p_1, if (i, j) are in the same module at the first level; and p_0, otherwise. In addition, we define that $p_2 > p_1 > p_0$ so that the resulting network will have a larger density of connections between nodes grouped in the same sub-module at the second level, a smaller density of connections between groups of nodes grouped in the same module at the first level, and an even smaller density of connections between nodes grouped in separate modules at the top level. Thus, the computer generated network has been constructed with a nested hierarchical structure which is illustrated in Fig. 3. The graph is generated with $p_0 = 0.002$, $p_1 = 0.03$ and $p_2 = 0.5$.

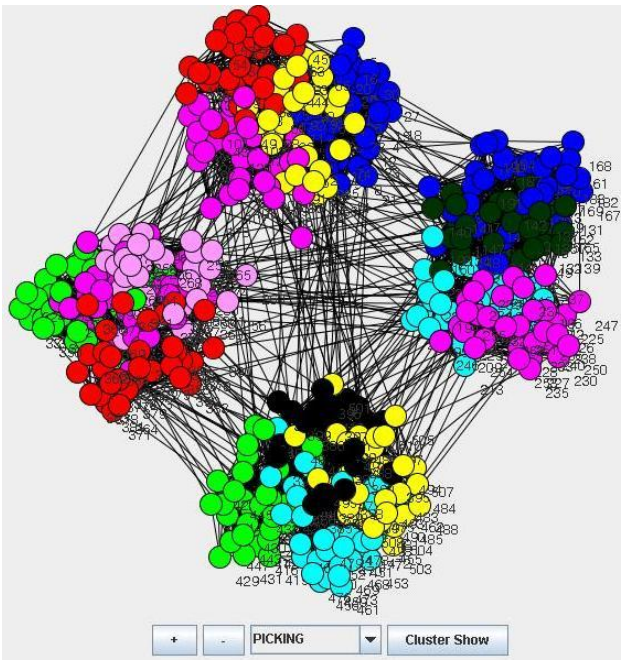


Figure 4. Communtiy structure of 512 nodes.

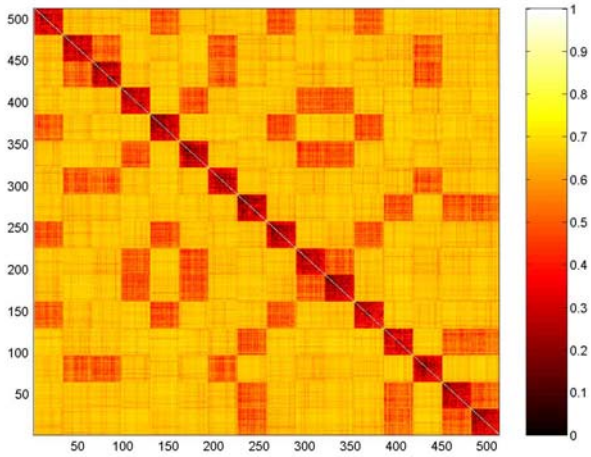


Figure 5. Similarity matrix of 512 nodes.

For our algorithm, we set the diameter $L_{rad} = 4$ and the decreasing factor $\alpha = 0.2$. The resulting structure is shown in Fig. 4 which consists of 16 communities, and these communities compose 4 bigger ones. It appears that our result corresponds to the hierarchical structure of the network.

The modularity we obtained is 0.7449. In addition, the similarity matrix obtained by our LFM is given in Fig. 5, in which we can easily find the 16 communities. Note the value 0 indicates the node pair is the most similar, while the value 1 shows the most dissimilarity.

Another hierarchical network we used is proposed by Ravasz et al. [24]. As Ravasz et al. pointed out [25], conventional network clustering methods are hard to uncover in the hierarchical community structure of such a network. However, our method performs well. We can see that the basic hierarchical organizations of the network are clear, which is illustrated in Fig. 6. The modularity we obtained is 0.5271. In addition, we can distinguish the communities clearly from our metric shown in Fig. 7.

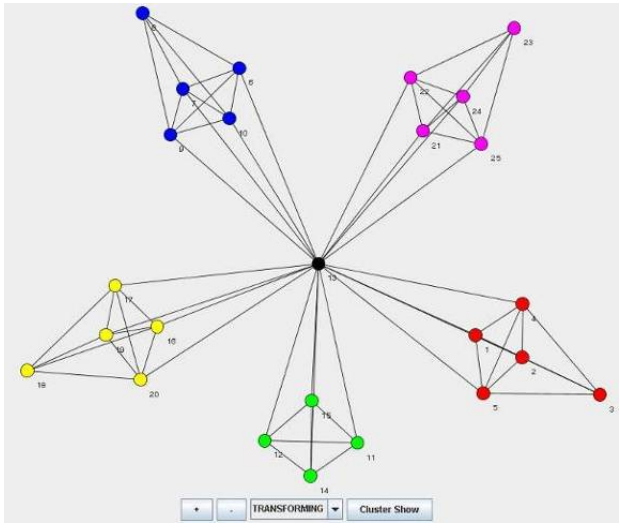


Figure 6. Community structure of Ravasz-Barabasi network.

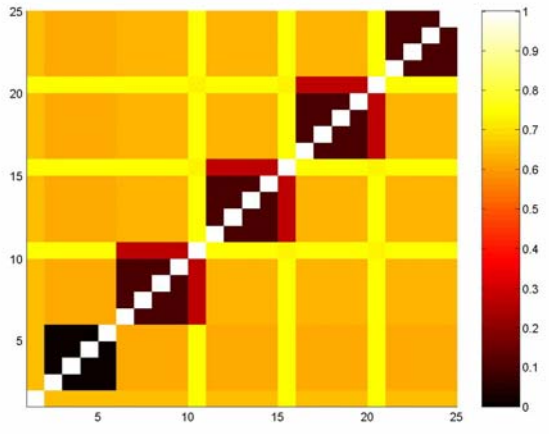


Figure 7. Similarity matrix of Ravasz-Barabasi network.

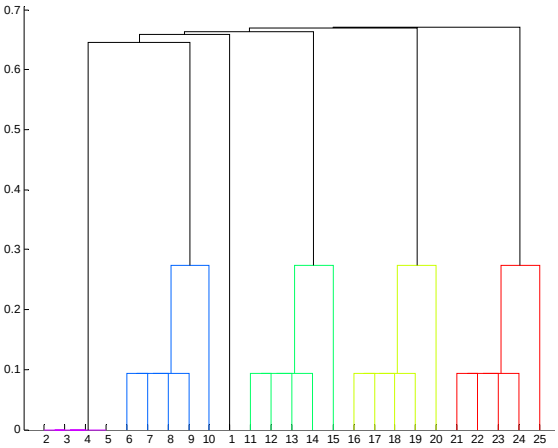


Figure 8. Dendrogram of Ravasz-Barabasi network.

In addition, the merge sequence is proposed by the dendrogram in Fig. 8.

B. Social Networks in Real World

At first, we test the classical social network of Zachary's karate club shown in Fig. 2. The famous karate club network analyzed by Zachary is widely used as a test benchmark for the methods of detecting communities in complex networks [20].

The network consists of 34 members of a karate club as nodes and 78 edges representing friendship between members. Due to a disagreement between the club's administrator and the club's instructor, the club split into two smaller ones. The challenge is whether we can extract the potential hierarchical structures of the network and detect the correct communities.

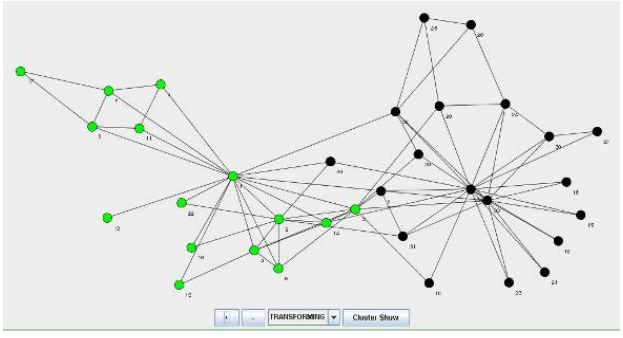


Figure 9. Community structure of karate network.

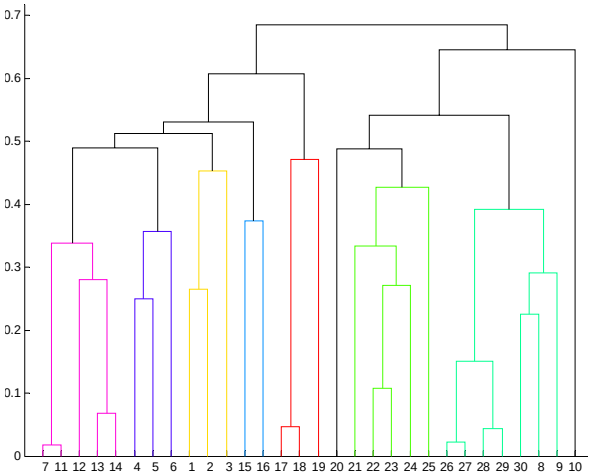


Figure 10. Community structure of karate network.

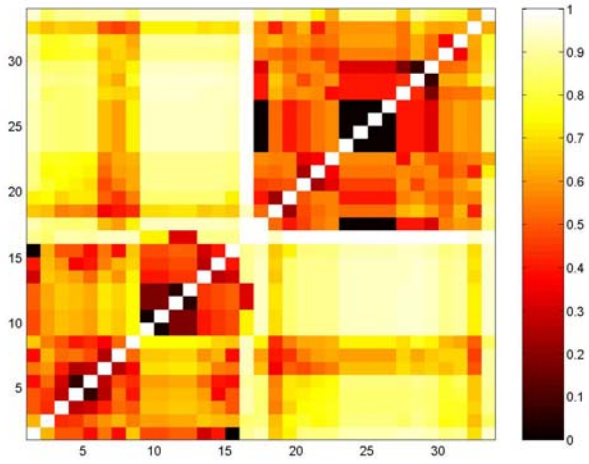


Figure 11. Similarity matrix of karate network.

Figure 9 shows that our result contains two communities. The similarity matrix is given in Fig. 11, in which the bottom-left corner is one community, and the top-right is the other. Thus, our metric works well in karate network. In addition, the modularity we obtained is 0.3542.

The second social network we test is the college football network which represents the game schedule of the 2000 season of Division I of the US college football league.

The nodes in the network represent the 115 teams, while the edges represent 613 games played in the course of the year. The teams are divided into conferences of 8-12 teams each and generally games are more frequent between teams of the same conference than between teams of different conferences. What we concern is to extract the potential hierarchical structures of the network, to detect the correct conferences.

From the matrix in Fig. 13, we can accurately distinguish that the number of conferences in colleague football league is about 11. Furthermore, the sizes of conferences output by our algorithm are in the range of 8-12, which corresponds to the actual situation.

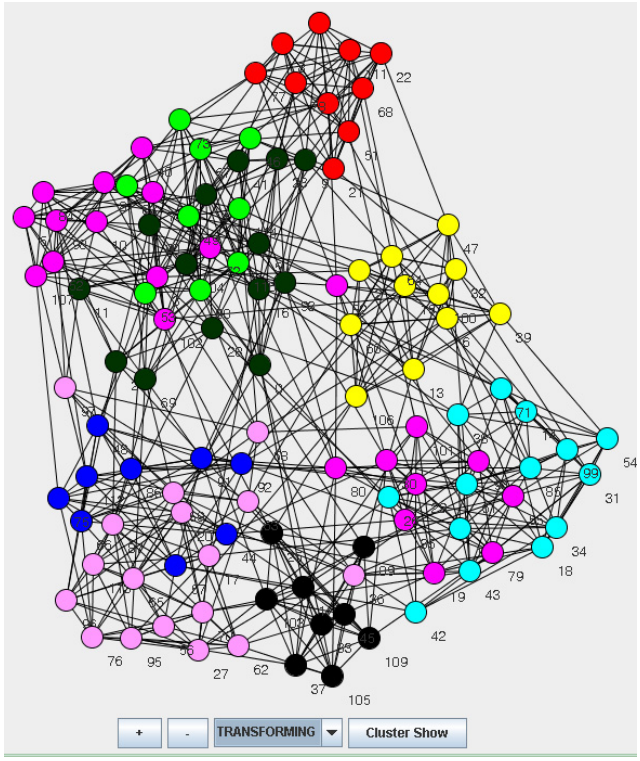


Figure 12. Community structure of colleage football network.

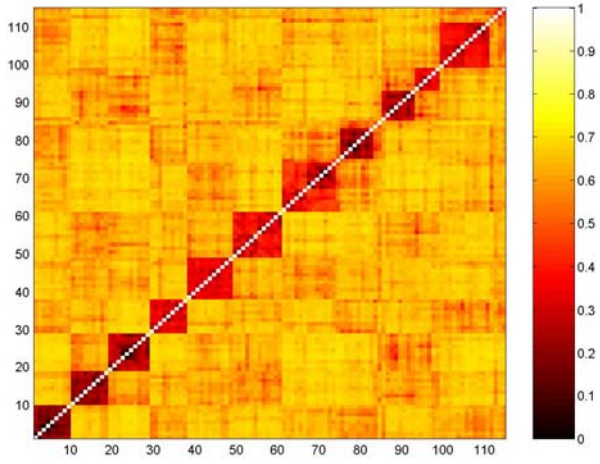


Figure 13. Similarity matrix of colleage football network.

Next, we will investigate the dolphin social network, representing the social interactions of bottlenose dolphins living in Doubtful Sound, New Zealand.

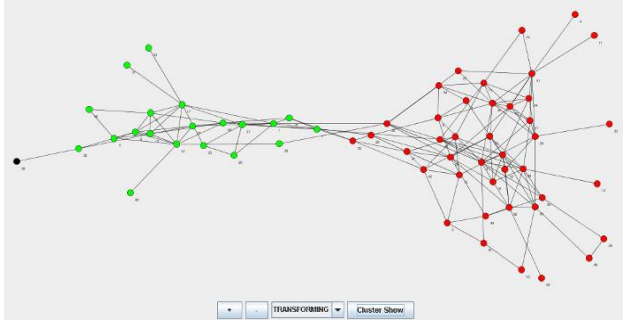


Figure 14. Community structure of dolphins social network (LFM).

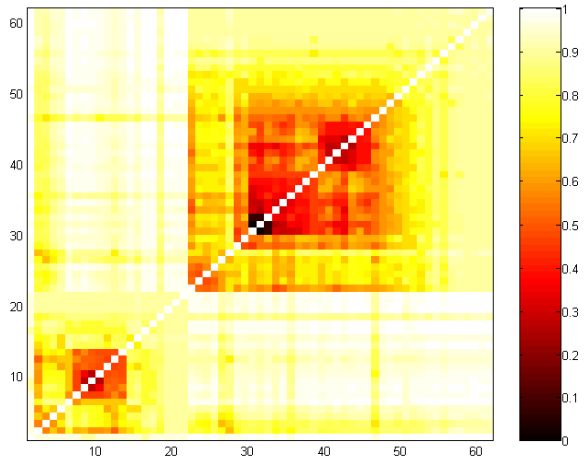


Figure 15. Similarity matrix of dolphins social network (LFM).

The network was studied by the biologist David Lusseau [26], who divided the dolphins into two groups according to their age. This network was constructed from observations of a community of 62 bottle-nose dolphins over a period of 7 years from 1994 to 2001.

Figure 14 shows that the structure of our result divides the network into three communities: two big communities and one isolated node (the leftmost node). Since the isolated node is adjacent (similar) with just one other node, it can be considered as one community. Thus, our result is nearly optimal.

The modularity of our result is 0.3742 which is much less than 0.5042 which is outputted by the metric of neighbor ratio s_{NR} defined by Formula (2).

However, Newman and Girvan [21] stated that the split into two groups appears to correspond to a known division of the dolphin community and its modularity is 0.38 ± 0.08.

Our result achieves this optimum expect one node (the leftmost one in Fig. 14). The community structure outputted by the metric s_{NR} is given in Fig. 16.

It is clear that the result obtained by the metric s_{NR} does not have significant community. The similarity matrix of s_{NR} also supports this conclusion which is given in Fig. 17.

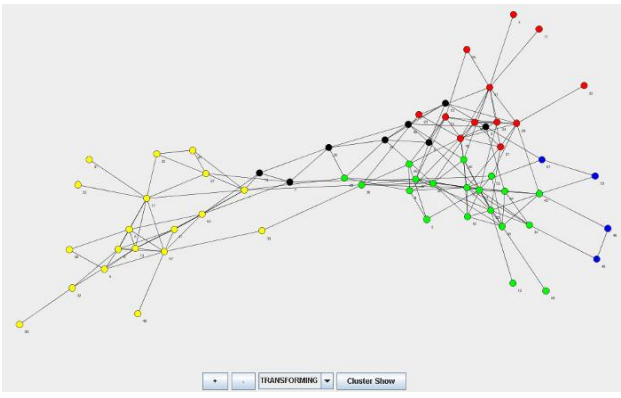


Figure 16. Community structure of dolphins social network (s_{NR}).

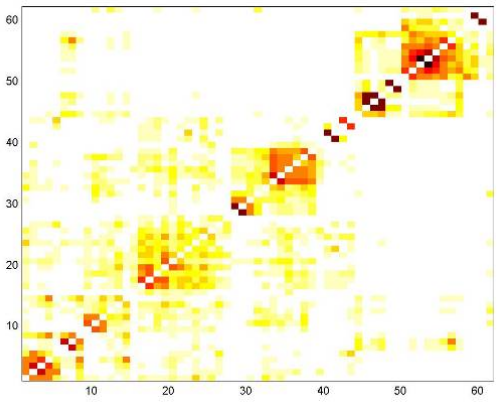


Figure 17. Similarity matrix of dolphins social network (s_{NR}).

Next, we discuss the parameter setting of the diameter L_{rad} and the decreasing factor α of our LFM model. Table III shows that the choice of the parameter L_{rad} affects the results more slightly than the choice of the decreasing factor α. Furthermore, the discovery quality is the best when $\alpha = 0.2$. This property also holds with respect to the two datasets of computer generated network. The statistic data is omitted, here.

TABLE III.
PARAMETERS SETTING OF LOCAL FLOW MODEL

karate club dataset			
L_{rad}	α	**Modularity**	**Number of Communities**
4	0.1	0.2003	17
	0.2	0.3542	2
	0.4	0.0664	15
	0.6	0.0621	14
	0.8	0	34
	1	0	34
3	0.1	0.2003	17
	0.2	0.3542	2
	0.4	0.0664	15
	0.6	0.0621	14
	0.8	0	34
	1	0	34
2	0.1	0.2003	17
	0.2	0.3542	2
	0.4	0.0664	15
	0.6	0.0621	14
	0.8	0	34
	1	0	34
1	0.1	0.2003	17
	0.2	0.3542	2
	0.4	0.0664	15
	0.6	0.0664	15
	0.8	0	34
	1	0	34

colleage football dataset			
L_{rad}	α	**Modularity**	**Number of Communities**
4	0.1	0.6057	9
	0.2	0.6021	9
	0.4	0.411	11
	0.6	0.0945	59
	0.8	0.043	87
	1	0.0390	86
3	0.1	0.6057	9
	0.2	0.6021	9
	0.4	0.411	11
	0.6	0.0945	59
	0.8	0.043	87
	1	0.0389	88
2	0.1	0.6057	9
	0.2	0.6021	9
	0.4	0.411	11
	0.6	0.0945	59
	0.8	0.043	87
	1	0.0389	88
1	0.1	0.6057	9
	0.2	0.6017	8
	0.4	0.5683	10
	0.6	0.515	13
	0.8	0.274	11
	1	0.1013	55

dolphins dataset			
L_{rad}	α	**Modularity**	**Number of Communities**
4	0.1	0.3973	19
	0.2	0.3742	3
	0.8	0.242	28
3	0.1	0.3973	19
	0.2	0.3742	3
	0.8	0.242	28
2	0.1	0.6057	9
	0.2	0.6021	9
	0.8	0.411	11
1	0.1	0.3973	19
	0.2	0.3742	3
	0.8	0.242	28

TABLE IV.
COMPARISON RESULTS

algortihm	**karate club**	**football colleage**	**dolphins society**
s_{NIP}	0	0.203	0.130
s_{NR}	0.34	0.602	0.5042
s_{WS}	0.213	0.323	0.332
s_{LFM}	0.3542	0.6021	0.3742
Girvan-Newman	0.406	0.572	0.52
CNM	0.302	0.402	0.353

Finally, we also show the comparison results of the-state-of-the-art algorithms Girvan-Newman algorithm [18][21] and CNM algorithm presented by Clauset, Newman and Girvan [19], which is given in Table IV.

VII. CONCLUSIONS

In this paper, we have proposed a new vertex similarity metric for community discovery. This metric employs a local flow in a small subgraph near the considered vertices. Thus, it considers both topological distance and local edge density. The experimental result shows that our metric is better than the existing ones. In addition, it appears that our algorithm based on this metric has several advantages to the Girvan-Newman algorithm and CNM algorithm in some aspects.

Since the computation of our metric is in a local part, distributed or parallel computation is available. It enables that our algorithm can tackle large scale graph. In

addition, in the light of Clauset's local algorithm [22], our algorithm can be extended to an incremental algorithm of which the running time is reduced dramatically. Furthermore, an online algorithm based on our model can be developed.

ACKNOWLEDGMENT

This research is supported in part by China Postdoctoral Science Foundation under Grant no.20100480060, NSFC under Grant no.61073195, and Shenzhen Science and Technology Program under Grant no. CXB201005250024A.

REFERENCES

[1] A. Clauset, C. Moore, and M.E.J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[2] D. Price, "Networks of scientific papers," In M. Kochen(Ed.), *the growth of knowledge: readings on organization and retrieval of information*, New York: Wiley, 1965, pp. 145–155.

[3] J. A. Dunne, R. J. Williams, and N. D. Martinez, "Foodweb structure and network theory: The role of connectance and size," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 12917–12922, 2002.

[4] S. A. Kauffman, "Metabolic stability and epigenesis in randomly connected nets," *J. Theor. Bio.*, vol. 22, pp. 437–467, 1969.

[5] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc. Natl. Acad. Sci. USA* , vol. 98, pp. 4569–4574, 2001.

[6] M. Sales-Pardo, R. Guimera, A.A. Moreira, and L.A.N. Amaral, "Extracting the hierarchicial structure of complex system," *Proc. Natl. Acad. Sci. USA*, vol. 104, pp. 15224–15229, 2007.

[7] B.W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell System Technical Journal*, vol. 49, pp. 291–308, 1970.

[8] M. Fiedler, "Algebraic connectivity of graphs," *Czech. Math. J.*, vol. 23, pp. 298–305, 1973.

[9] A. Pothen, H. Simon, and K.-P. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM J. Matrix Anal. Appl.*, vol. 11, pp. 430–452, 1990.

[10] J. Scott, *Social Network Analysis: A Handbook*. Sage, London, 2nd edition, 2000.

[11] L. Donetti, and M. A. Munoz, "Detecting network communities: a new systematic and powerful algorithm," *J. Stat. Mech.*, P10012, 2004.

[12] A. Capocci, V.D.P. Servedio, G. Caldarelli, and F. Colaiori, "Detecting communities in larger networks," *Physica A*, vol. 352, pp. 669, 2005.

[13] N. Alves, "Unveiling community structures in weighted networks," *Phys. Rev. E*, vol. 76, 036101, 2007.

[14] J. Reichardt and S. Bornholdt, "Detecting fuzzy community sturctures in complex networks with a Potts model," *Phys. Rev. Lett*, vol. 93, 218701, 2004.

[15] H. Zhou, "Distance, dissimilarity index, and network community structure," *Phys. Rev. E*, vol. 67, 061901, 2003.

[16] A. Arenas, A. Diaz-Guilera, and C. J. Peerez-Vicente, "Synchronization reveals topological scalses in complex networks," *Phys. Rev. Lett*, vol. 96, 114102, 2006.

[17] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75-174, 2010.

[18] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, 7821–7826, 2002.

[19] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, 066111, 2004.

[20] M. E. J. Newman, "Detecting community structure in networks," *Eur. Phys. J. B*, vol. 38, pp. 321–330, 2004.

[21] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, 026113, 2004.

[22] A. Clauset, "Finding local community structure in networks," *Phys. Rev. E*, vol. 72, 026132, 2005.

[23] R. Duda, P. Hart, D. Stork, *Pattern Recognition*, 2nd Edition, John Wiley and Sons, 2003.

[24] E. Ravasz and A.L. Barabási, "Hierarchical organization in complex network," *Phys. Rev. E*, vol. 67, 026112, 2003.

[25] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási, "Hierarchical organization of modularity in metabolic networks," Science, vol. 297, pp. 1551–1555, 2002.

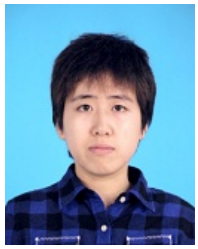
[26] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson, "The bottlenose dolphin commu-nity of doubful sound features a large problem of long-lasting associations," *Behav. Ecol. Sociobiol.*, vol. 54 pp. 396–405, 2003.

Yueping Li was born in Guangdong Province, China in Sep. 1980, and received his PhD in Computer Science from Sun Yat-sen University in 2008.

Currently, he is a post doctor in Shenzhen Graduate School, Harbin Institute of Technology. His research interests involve web mining, graph algorithm and optimization.

Yunming Ye was born in China in Sep. 1976, and received his PhD degree in Computer Science from Shanghai Jiao Tong University in 2004.

Currently, he is a professor in Shenzhen Graduate School, Harbin Institute of Technology. His research interests include Web mining, Web Search, and social computing.



Xiaolin Du was born in HeiLongjiang Province, China in Jan. 1983, and received her Master Degree in Computer Science from Harbin Institute of Technology in 2009.

Currently, she is a PhD candidate in Shenzhen Graduate School, Harbin Institute of Technology. Her research interests involve data mining, data visualization and social network discovering.