A New Vertex Similarity Metric for Community Discovery: A Distance Neighbor Model

Yueping Li

Shenzhen Graduate School, Harbin Institute of Technology, Xili
Shenzhen, 518055, China
leeyueping@gmail.com

Abstract. The hierarchical clustering methods based on vertex similarity can be employed for community discovery. Vertex similarity metric is the most important part of these methods. However, the existing metrics do not perform well compared with the state-of-the-art algorithms. In this paper, we propose a new vertex similarity metric based on distance neighbor model, called Distance Neighbor Ratio Metric (DNRM), for community discovery. DNRM considers both distance and nearby edge density which are essential measures in community structure. Compared with the existing metrics of vertex similarity, DNRM outperforms substantially in community discovery quality and the computing time. The experiments are designed rigorously involving both well-known social networks in real world and computer generated networks.

Keywords: Hierarchy clustering, vertex similarity, community discovery, modularity, time complexity.

1 Introduction

1.1 Background

Graph mining attracts much attention in both academy and industry. 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 and the edges represent the relations. Recent studies [1] reveal that many graphs in society often exhibit hierarchical *community structure*. In addition, the communities correspond to known sets of units dealing with related topics, such as citation networks [2], food webs [3] , and biochemical networks [4,5]. Thus, community discover plays an essential role for the identification and characterization of real networks [6]. Furthermore, uncover hierarchical community structure emerges as a premise task for capturing an in-depth understanding of networks.

In the literature, community discovery algorithms have been well studied. Generally, they can be divided into three categories: graph partitioning, hierarchical clustering, and methods for hyperlink-based network. Graph partition algorithms include the Kernighan-Lin algorithm [7], spectral partitioning

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 228–237, 2011.

[8,9]. Hierarchical algorithms contain two classes: agglomerative methods based on the optimization of modularity or the similarity metrics, and the divisive methods based on betweenness metrics such as Girvan-Newman (GN) algorithm [10], Tyler algorithm [11], and Radicchi algorithm [12]. The methods for detecting hyperlink-based Web communities such as the maximum flow communities (MFC) algorithm [13], the hyperlink- induced topic search (HITS) algorithm [14], the spreading activation energy (SAE) algorithm [15].

There are also many other kinds of methods based on different technologies such as spectral property of graph matrix[16][17][18], spin-spin interactions [19], random walks [20] and synchronization [21]. For more details, the reader can refer to the survey article by Fortunato [22].

1.2 Related Metrics of Vertex Similarity

In essence, similarity metric is the most important part of agglomerative methods 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.

However, it appears that these methods perform well for specific types of problems , but work poorly in more general cases [23]. The reason is that existing vertex similarity metrics are designed for particular kinds of graphs. Thus, the algorithms based on these metrics cannot tackle a variety of graphs.

Next, we present several classical metrics of vertex similarity for community discovery. Then, we will show their limitations and shortcomings.

One well-known similarity measure is Jaccard Index [24], which is defined as follows. For a vertex $u \in V(G)$, let $\Gamma(u)$ be the the set of neighbor vertices of u. It is natural that two vertices u and v are more likely if they share more common neighbor vertices. In addition, if the shared neighbors take up more proportion of all their neighbors, it also shows more similarity between these two vertices. The formula is defined as follows:

$$s_{JI}(i,j) = \frac{|\Gamma(i) \bigcap \Gamma(j)|}{|\Gamma(i) \bigcup \Gamma(j)|} \tag{1}$$

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

Another metric is the number of independent paths between two vertices, denoted by s_{NIP}. Independent paths do not share any edge (vertex). This metric indicates 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, and the flow value is an integer.

An alternate metric evaluates the number of the paths between two vertices. In this case, the problem is that the total number of paths is huge. But one method solving this problem is to compute a weighted sum of the number of paths, where paths of length l are weighted by the factor α^l, with $\alpha < 1$. It is clear that the paths with longer length can be neglected due to its tiny weight. Denote this weighted sum metric by s_{WS}.

There are several metrics defined by the distance in an n-dimensional Euclidean space, by assigning a position to each vertex of the given graph. Thus, the positions of the vertices have to be determined before computing these metrics. One approach to determine position of each vertex is the spectral bisection method which is based on the properties of the Laplacian matrix [25]. Due to its high computation cost, these metrics are seldom used in community discovery. Thus, they are not considered in our paper.

In brief, the metrics above consider either edge density in neighbors or the number of connecting paths. Since the similarity of one vertex pair is affected by the following features: distance, local edge density and number of disjoint connecting paths in global, these metrics are not well defined shown in Table 1. In addition, the existing metrics are not good for community discovery, for instance, the time complexity and the quality is not satisfied.

Table 1. Features of existing metrics

Metric	distance	edge density	disjoint paths	quality	time complexity
s_{JI}	×	√	×	mediate	$O(\|E(G)\|)$
s_{NIP}	×	partial	√	bad	$O(\|V(G)\|\|E(G)\|)$
s_{WS}	√	partial	partial	mediate	$O(2^{\|V(G)\|})$

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

The rest of this paper is organized as follows. Section 2 formulas our problem, and propose the measure of quality. Section 3 introduces our metric. Section 4 gives our algorithm. Experimental results are presented in Section 5. Finally, in Section 6, we summarize this work and point out the future work.

2 Problem Statement

Our problem is to divide the considered graph into communities in certain application scenario. Unfortunately, community structure has no universal accepted definition [1]. One common used one is that the division of vertices into groups such that there is a higher edge density within groups while less edge density between them.

This paper considers simple graphs only, i.e., the graphs without loop or multi-edges. Given graph G, $V(G)$ and $E(G)$ denote the sets of its vertices and edges respectively. In addition, our paper considers unweighted graphs, that is, all edges are unweighted.

A community structure is a partition $\mathcal{P} = C_1, C_2, \ldots, C_k$ of graph G such that $C_1 \bigcup C_2 \bigcup \ldots \bigcup C_k = V(G)$ and $C_i \bigcap C_j = \emptyset$ for $i \neq j$.

It appears that the number of the partitions of one graph is huge. One measure is necessary for evaluating the quality of one partition with respect to the

community in scenario. One common used quantitative measure is ***modularity*** [26]. Definition given below states that communities in a good partition has high intra-community edge density and less inter-community edge density:

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

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 ($C_i = C_j$), zero otherwise.

We are supposed to find an optimal partition $\mathcal{P}$ which makes the modularity $Q(\mathcal{P})$ maximum. When one partition has modularity larger than 0.3∼0.4, it can be concluded that this partition has community structure. The larger the modularity is, its community structure is more prominent and clear. However, it is well-known that our problem is an NP-hard problem [22]. Thus, there is not polynomial time algorithm for this problem unless P=NP. Most existing methods are approximation algorithms. Furthermore, modularity has limits [22]. Thus, the result community structure is favorable only if it corresponds to the actual structure in real world. If there is no in-advance structure information, the result will recommend a probable community structure but not an optimal one.

3 A New Vertex Similarity Metric Based on Distance Neighbor Model

3.1 Definition and Properties

Our model is designed based on the fact that the probability that two vertices have the same neighbors when they are in the same community is larger than the case they lie in different ones. This motivation arises from the neighbor ratio metric s_{JI}. But we extend this model by allowing not only adjacent neighbors but also neighbors within certain distance, which is a threshold denoted by d. We next present the definition of our distance neighbor model.

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 $x, y \in V(G)$.

The idea of the neighbor ratio metric s_{JI} is that if the proportion of the common neighbor over all neighbors is larger, then these two vertices i and j is more similar. Since we consider the non-adjacent neighbors within the distance d, it is natural that these neighbors' contribution to the similarity is less than the adjacent neighbors. Thus, it can be concluded that the contribution decreases when the distance from the neighbor to i or j increases. Therefore, the metric s_{DNM} of our distance neighbor model is defined as follows:

$$\Gamma(i, d) = \{u | dis(i, u) = d, \forall u \in V(G)\} \tag{3}$$

$$s_{DNM}(i, j) = \sum_{k=1}^{d} \frac{|\Gamma(i, k) \cap \bigcup_{c=1}^{k} \Gamma(j, c)| + |\Gamma(j, k) \cap \bigcup_{c=1}^{k} \Gamma(i, c)|}{k \times \sum_{c=1}^{k} |\Gamma(i, c) \cup \Gamma(j, c)|} \tag{4}$$

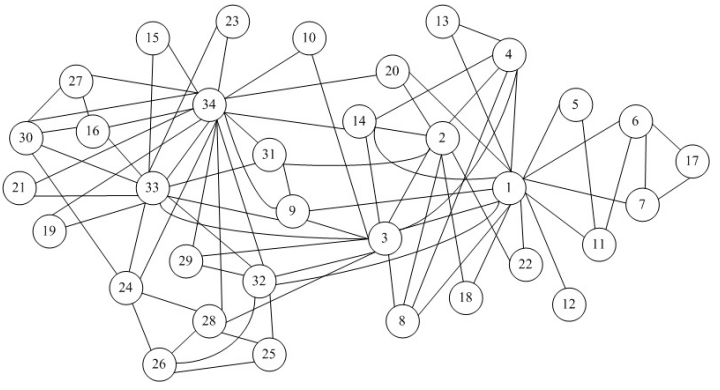

Fig. 1. Karate network

For better understanding of our metric, an illustration example using the well-known dataset karate club network is given. The topology structure is shown in Fig. 1. The statistic data of vertex pair is presented in Table 2.

Table 2. The statistic data of vertex pair $(6, 3)$

d	1	2
$\Gamma(6,d)$	1,7,11,17	3,4,5,8,9,11,12,13,14,18,20,22
$\Gamma(3,d)$	1,2,4,8,9,10,14,28,29,32	5,6,7,11,12,13,20,22,24,25,26,31,33,34
$s_{DNM}(6,3)$	$\frac{1}{13}$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}$
d	3	4
$\Gamma(6,d)$	2,10,28,29,31,32,33,34	15,16,19,21,23,24,25,26,27,30
$\Gamma(3,d)$	15,16,17,19,21,27,30	0
$s_{DNM}(6,3)$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}+\frac{1}{3}\times\frac{9}{34}$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}+\frac{1}{3}\times\frac{9}{34}+\frac{1}{4}\times\frac{10}{34}$

According to the definition, it can be concluded that our metric is symmetric, that is, $s_{DNM}(i,j) = s_{DNM}(j,i)$ for $\forall i,j \in V(G)$.

Next, we propose the relationship between local edge density and $s_{DNM}(i,j)$. Set $d_e(i) = \min\{dis(a,i), dis(b,i)\}$ where the edge $e = (a,b)$ and $a,b,i \in V(G)$. Let $P_{(i,j)}(k)$ be the set of the paths between vertices i and j in which the length of each path is no larger than k. Then we have the following proposition.

Proposition 1. *Let d be the chosen distance threshold, and $E(P_{(i,j)}(k))$ be the edge set of the path set $P_{(i,j)}(k)$ where $i,j \in V(G)$ and k is a positive integer. Set $E^d_{(i,j)}$ to be the edge set satisfying: (I) the distance from i and from j is no larger than d ; (II) lies in $E(P_{(i,j)}(2d))$. Then, we have*

$$\sum_{k=1}^{d} \frac{E^k_{(i,j)}}{k \times \sum_{c=1}^{k} |\Gamma(i,c) \cup \Gamma(j,c)|} \geq s_{DNM}(i,j) \tag{5}$$

3.2 Comparison with Existing Vertex Similarity Metrics

In this subsection, we compare our metric with the existing ones on several pairs of the graph modelled by Zachary's karate club illustrated in Fig. 1. Choose the distance threshold d to be 2. We select a couple of pairs to vertices, and the values are given in Table 3.

Table 3. Metrics comparison in karate club graph

Pairs	(1,12)	(1,6)	(1,2)	(6,7)	(6,2)	(6,30)	(17,30)	(2,12)
s_{JI}	0	2/17	7/16	1/4	0	0	0	0
s_{NIP}	1	4	9	4	4	4	2	1
s_{WS}	1.0	3.2	109.2	3.1	108.3	164.7	56.6	46.3
s_{DNM}	0.581	0.487	0.405	0.610	0.469	0.323	0.164	0.426

Discussion: The metric of neighbor ratio s_{JI} 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 evaluations 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 distance neighbor model is better in describing the similarity compared with the others.

Finally, we summary the features of our metric in Table 4.

Table 4. Features of our metric s_{DNR}

Metric	distance	edge density	disjoint paths	quality	time complexity
s_{DNR}	$\surd$	$\surd$	$\times$	good	$O(\|E(G)\|)$

4 Algorithm Description

Community Discovery Algorithm Based on Distance Neighbor Model

Input: a simple, undirected and unweighed graph G
Output: a community structure
1. Choose the diameter d.
2. Foreach vertex pair (i, j) where $i \neq j$ and $dis(i, j) \leq 2 \times d$ Do
3. Begin For $k := 1$ to d Do Search $\Gamma(i, k)$ and $\Gamma(j, k)$; //endfor
4. For $k := 1$ to d Do Compute $\Gamma(i, k) \cap \Gamma(j, k)$ and $\Gamma(i, k) \cup \Gamma_j, k)$; //endfor
5. Compute $s_{DNM}(i, j)$;
6. End //foreach
7. Use the classical average linkage method to find the community structure and output it.

Complexity analysis: Steps 3-6 employ a procedure which is a part of breadth-first-search. Thus, the running time is bounded by $|E(G)|$. The step 2 is a loop which repeats for each pair of vertices. Therefore, the computation of metrics needs $O(|V(G)|^2|E(G)|)$ 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 $O(|V(G)|^2|E(G)|)$. It is necessary to mention that the actual running time is much less than this worst time, since if the distance between two vertices is larger than $2d$, no computation is needed.

5 Experimental Results

We implement the algorithm is Section 4 in Java and perform experiments in several well-known datasets. We choose the distance threshold d to be four.

The first data set is Ravasz network [27]. As Ravasz et al. pointed out, conventional network clustering methods are difficult to find the correct community structure of such network. The community structure and the merge sequence (dendrogram) are presented in Fig. 2 and Fig. 3, respectively.

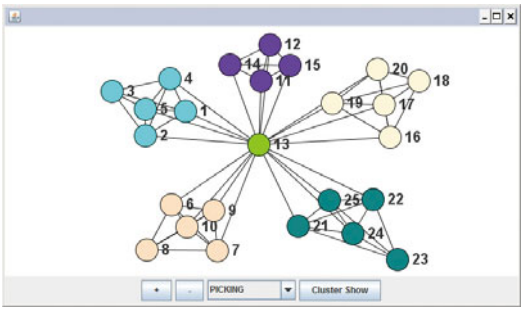

Fig. 2. Community structure using our metric in Ravasz network

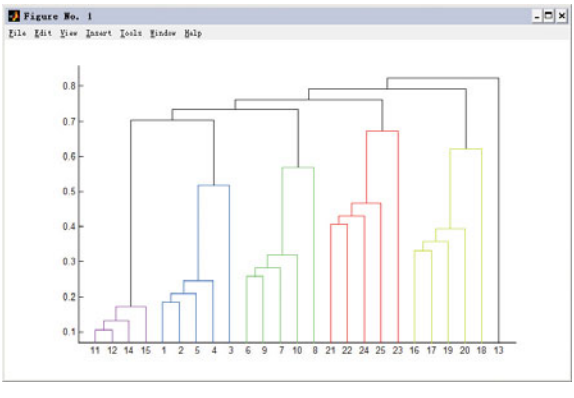

Fig. 3. Merge sequence using our metric in Ravasz network

Table 5. Modularity Results

	Ravasz	karate	football	dolphins	musician
s_{NIP}	0.3452	0.203	0.1300	0.018	0.1450
s_{JI}	0.1308	0.3400	0.6020	0.4280	0.4010
s_{WS}	0	0.2130	0.3230	0.3320	0.3010
s_{DNM}	0.5270	0.3628	0.6042	0.4278	0.4047
GN	0.5509	0.406	0.572	0.52	0.405
CNM	0.3326	0.302	0.402	0.353	0.439

We also test some famous datasets: karate club network [23], US college football league [10], Jazz musician network [28] and dolphin social network [29]. We compare our algorithm with the algorithms based on existing vertex similarity metrics. In addition, we also show the result of the-state-of-the-art algorithms Girvan-Newman algorithm [10] and CNM algorithm presented by Clauset, Newman and Girvan [1].

We propose the experimental results of modularity in Table 5, which shows that our metric is superior to the existing ones. Table 5 indicates that the Girvan-Newman algorithm outperforms in three datasets compared with our algorithm. However, in Ravasz network our metric finds the correct structure shown in Fig. 2, though GN gets high modularity; In dolphins society the modularity of best division is 0.478 ± 0.03 stated by Lusseau [29]. Our metric is better than GN in this dataset. In addition, the time complexity of (improved) Girvan-Newman algorithm $O(|E(G)|^2|V(G)|)$ is higher than our algorithm.

6 Conclusions

In this paper, we have proposed a new vertex similarity metric for community discovery. This metric computes overlapping proportion of neighbors within a distance with respect to 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 advantage 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 which enables that our algorithm can tackle large scale graph. In addition, in the light of Clauset's local algorithm [30], 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.

Acknowledgements

This research is supported in part by NSFC under grant No.60603066, China National High-tech Program under grants No.2007AA01Z436, and Shenzhen

Science and Technology Program under grants No.NSKJ-200707, 08CXY-44, PCT200805190162A.

References

1. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101 (2008)
2. Price, D.: Networks of scientific papers. In: Kochen, M. (ed.) The Growth of Knowledge: Readings on Organization and Retrieval of Information, pp. 145–155. Wiley, Chichester (1965)
3. Dunne, J.A., Williams, R.J., Martinez, N.D.: Foodweb structure and network theory: The role of connectance and size. Proc. Natl. Acad. Sci. USA 99, 12917–12922 (2002)
4. Kauffman, S.A.: Metabolic stability and epigenesis in randomly connected nets. J. Theor. Bio. 22, 437–467 (1969)
5. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98, 4569–4574 (2001)
6. Sales-Pardo, M., Guimera, R., Moreira, A.A., Amaral, L.A.N.: Module identification in bipartite and directed networks. Proc. Natl. Acad. Sci. USA 104, 15224–15229 (2007)
7. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal 49, 291–308 (1970)
8. Fiedler, M.: Algebraic connectivity of graphs. Czech. Math. J. 23, 298–305 (1973)
9. Pothen, A., Simon, H., Liou, K.P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11, 430–452 (1990)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
11. Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as spectroscopy: automated discovery of community structure within organizations. In: Huysman, M.H., Wenger, E., Wulf, V. (eds.) Proceedings of the International Conference on Communities and Technologies, pp. 81–96. Springer, Heidelberg (2003)
12. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and indentifying communities in networks. Proc. Nat. Academy of Science (PNAS) 101(9), 2658–2663 (2004)
13. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.M.: Self-Organization and Identification of Web Communities. Computer 35(3), 66–71 (2002)
14. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. J. ACM 46(5), 604–632 (1999)
15. Pirolli, P., Pitkow, J., Rao, R.: Silk from a Sows Ear:Extracting Usable Structures from the Web. In: Proc. ACM Conf. Human Factors in Computing Systems (CHI), pp. 118–125 (1996)
16. Donetti, L., Münoz, M.A.: Detecting network communities: a new systematic and powerful algorithm. J. Stat. Mech., P10012 (2004)
17. Capocci, A., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: The scale-free topology of market investments. Physica A 352, 669 (2005)
18. Alves, N.A.: Unveiling community structures in weighted networks. Phys. Rev. E 76(3), 36101 (2007)

19. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a potts model. Phys. Rev. Lett. 93(21), 218701 (2004)
20. Zhou, H.: Distance, dissimilarity index, and network community structure. Phys. Rev. E 67(6), 061901 (2003)
21. Arenas, A., Diaz-Guilera, A., Peerez-Vicente, C.J.: Synchronization reveals topological scalses in complex networks. Phys. Rev. Lett. 96(11), 114102 (2006)
22. Fortunato, S.: Community detection in graphs, arXiv, 0906.0612 (2009)
23. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B 38, 321–330 (2004)
24. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Socit Vaudoise des Sciences Naturelles 37, 547–579 (1901)
25. Barnes, E.R.: An algorithm for partitioning the nodes of a graph. SIAM Journal for Algorithms and Discrete Methods 3, 541–550 (1982)
26. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004)
27. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabsi, A.L.: Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555 (2002)
28. Gleiser, P., Danon, L.: Community structure in Jazz. Adv. Complex Systems 6, 565–573 (2003)
29. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin commu-nity of doubful sound features a large problem of long-lasting associations. Behav. Ecol. Sociobiol. 54, 396–405 (2003)
30. Clauset, A.: Finding local community structure in networks. Phys. Rev. E 72, 026132 (2005)