The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

A Weighted Parsimony Model for Community Detection in Complex Networks*

Junhua Zhang[1,2,†] Xiang-Sun Zhang[1]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190, China
[2]Key Laboratory of Random Complex Structures and Data Science,
Academy of Mathematics and Systems Science, CAS, Beijing 100190, China

Abstract Many real-world networks have a common feature of organization, i.e., community structure. Detecting this structure is fundamental for uncovering the links between the structure and the function in complex networks and for practical applications in many disciplines such as biology and sociology. In this paper we propose a weighted parsimony criterion for community detection in complex networks. This criterion relates communities with cliques (or complete subgraphs). Parsimony here means that as minimal as possible number of inserted and deleted edges is needed when we make the network considered become a disjoint union of cliques. A weight based on the topological features of the network is introduced to ensure the obtained subgraphs to be communities by balancing the inserted and deleted edges. Tests on real networks give excellent results.

Keywords Community detection; parsimony; cliques; complex networks

1 Introduction

Research on complex networks has attracted a great deal of attention in recent years, one main reason is that many real-world systems can be represented by networks composed of vertices and edges [1, 2, 3]. Among others, some typical examples are the Internet [4], social networks [5, 6], biological networks [7, 8] as well as the food webs [9]. Many such networks are characterized by a mesoscopic level of organization, with groups of nodes forming tightly connected units, called communities or modules, that are only weakly linked to each other [10, 11, 12, 13, 14]. Community detection and network partition are fundamental for uncovering the links between the structure and the function in complex networks and for practical applications in many disciplines such as biology and sociology.

A large number of papers related to community detection have emerged in recent years. By and large, such papers can be classified into two categories: the one is focusing on the partition criteria function (quality index of partition) design, the other puts interest in presenting algorithms that describe the dynamical process or a procedure resulting the network community structure. For the first class research, one can use different methodologies and algorithms to realize the partition criteria, while for the second

*This work is partially supported by the National Natural Science Foundation of China under grant No.60873205, Innovation Project of Chinese Academy of Sciences, kjcsyw-s7.
†Corresponding author. zjh@amt.ac.cn

class research one does not care too much for the potential corresponding criteria of the studied algorithms. Quality functions, such as modularity defined by Newman and Girvan [11], modularity density [15], entropy function form [16], and Potts model related methods [17, 18, 19], belong to the first category. Other approaches include clique percolation [20, 21], spectral [13], a continuous mapping to a conic optimization problem [22], and maximum likelihood [23]. Ref. [24] defines a measure of robustness of community structure based on random perturbations. While the paper [10], as well as the recent publications [25, 26, 27, 28, 29, 30], are related to detection algorithms. Plenty of methods regarding community detection and network partition in complex networks have been recently reviewed in [31, 32].

In this paper we propose a weighted parsimony criterion for community detection in complex networks. This criterion relates communities with cliques (or complete subgraphs). Parsimony here means that as minimal as possible number of inserted and deleted edges is needed when we make the network considered become a disjoint union of cliques. A weight based on the topological features of the network is introduced to ensure the obtained subgraphs to be communities by balancing the inserted and deleted edges. Tests on real networks give excellent results.

2 Community detection as a constrained optimization problem

2.1 Parsimony criterion for community detection

The first criterion for community identification was given by M.Grötschel and Y.Wakabyashi [33, 34] (GW). A summary statement of the GW criterion is presented in [25]:

"Identifying a community structure in a network is nothing but inserting and deleting edges in a somehow most parsimonious way so that the network becomes a target network, i.e., a disjoint union of complete subgraphs (or cliques)".

Here we give a closed mathematical formula for GW criterion. Let $A = (a_{ij})$ be an $n \times n$ symmetric adjacency matrix of a network $N(V;E)$, where $V = \{v_1, v_2, \cdots, v_n\}, E = \{(v_i, v_j) : a_{ij} = 1\}$. By defining B as a $s \times s$ matrix with all elements equal to 1, B represents a complete subnetwork with dimension $s (s = \sqrt{|B|})$. For an n-nodes network, a candidate target network is a set of non-overlapping complete subnetworks $B_1, \cdots B_k$, $\sqrt{|B_1|} + \cdots + \sqrt{|B_k|} = n$. With these notations, the parsimony criterion can be described by the following optimization model:

$$\begin{aligned} &\min_k \min_{\sum_{i=1}^k \sqrt{|B_i|}=n} \{\sum_{i=1}^k (|B_i| - |A_{B_i}| - \sqrt{|B_i|}) + (|A| - \sum_{i=1}^k |A_{B_i}|)\} \\ =&\min_k \{|A| - n + \min_{\sum_{i=1}^k \sqrt{|B_i|}=n} \sum_{i=1}^k |B_i| - 2\sum_{i=1}^k |A_{B_i}|\}, \end{aligned} \tag{1}$$

where A_{B_i} is a sub-matrix of A with elements corresponding to B_i. Equivalently, we optimize:

$$P : \min_k \min_{\sum_{i=1}^k \sqrt{|B_i|}=n} \{\sum_{i=1}^k |B_i| - 2\sum_{i=1}^k |A_{B_i}|\}. \tag{2}$$

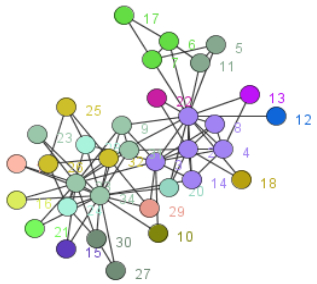

Figure 1: The community structure of the karate club network obtained by the parsimony model *PM*.

Formally we write the parsimony model as

$$PM: \min_k \bar{P}(k) = \min_k \min_{\sum_{i=1}^k \sqrt{|B_i|}=n} P(B_1, \cdots, B_k), \tag{3}$$

and the sub-optimization problem as

$$\begin{aligned} &\bar{P}: \min P(B_1, \cdots, B_k), \\ &s.t. \sum_{i=1}^{k} \sqrt{|B_i|} = n, \\ &P(B_1, \cdots, B_k) = \sum_{i=1}^{k} |B_i| - 2\sum_{i=1}^{k} |A_{B_i}|. \end{aligned} \tag{4}$$

2.2 Some problems

When we apply the above parsimony model to some real-world networks, unexpected results are encountered. The algorithm for partitioning the network we use is the simulated annealing algorithm for module identification [35], a widely used algorithm for community detection in complex networks.

The first network we investigate is the famous karate club network analyzed by Zachary [36], which is widely used as a test example for methods of detecting communities in complex networks [10, 12, 37]. The network consists of 34 members of a karate club as nodes and 78 edges representing friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and the club's instructor, the club split into two smaller ones.

However, 18 groups are obtained by using the parsimonious model (3) and (4) (Figure 1), in which several groups consist of only one node. Obviously it is not proper for us to take a single node as a community.

Similar result is obtained for another widely used network, the scientific collaboration network [10, 38]. Now we get 64 groups for the 200-node network.

Remarks. When the network is sparse, that is, it is far from a clique, the result is usually unsatisfactory. In this case many such groups that only contain one or two nodes are

detected, because the less the nodes in a group, the less the edges are needed to add into the group to make it a clique. In view of the complexity of networks, maybe it is proper to use a weight to balance the inserted term $\sum_{i=1}^{k}(|B_i|-|A_{B_i}|-\sqrt{|B_i|})$ and the deleted term $(|A|-\sum_{i=1}^{k}|A_{B_i}|)$ of the parsimony model (1).

3 A weighted parsimony criterion for community detection

3.1 The weighted model

Now we consider the weighted parsimony model as follows:

$$\begin{aligned}&\min_k \min_{\sum_{i=1}^{k}\sqrt{|B_i|}=n} \{w\cdot\sum_{i=1}^{k}(|B_i|-|A_{B_i}|-\sqrt{|B_i|})+(1-w)\cdot(|A|-\sum_{i=1}^{k}|A_{B_i}|)\}\\ =&\min_k\{(1-w)\cdot|A|-w\cdot n+\min_{\sum_{i=1}^{k}\sqrt{|B_i|}=n} w\cdot\sum_{i=1}^{k}|B_i|-\sum_{i=1}^{k}|A_{B_i}|\},\end{aligned} \tag{5}$$

where $0<w<1$ is a weight coefficient.

Equivalently, we optimize:

$$WP:\min_k \min_{\sum_{i=1}^{k}\sqrt{|B_i|}=n} \{w\cdot\sum_{i=1}^{k}|B_i|-\sum_{i=1}^{k}|A_{B_i}|\}. \tag{6}$$

Formally we write the weighted parsimony model as

$$WPM:\min_k \overline{WP}(k)=\min_k \min_{\sum_{i=1}^{k}\sqrt{|B_i|}=n} WP(B_1,\cdots,B_k), \tag{7}$$

and the sub-optimization problem as

$$\begin{aligned}&\overline{WP}:\min WP(B_1,\cdots,B_k),\\ &\quad s.t.\sum_{i=1}^{k}\sqrt{|B_i|}=n,\\ &\quad WP(B_1,\cdots,B_k)=w\cdot\sum_{i=1}^{k}|B_i|-\sum_{i=1}^{k}|A_{B_i}|.\end{aligned} \tag{8}$$

3.2 Selection of the weight coefficient

The main role of the weight w is to balance the inserted and deleted edges for getting disjoint cliques from the network. From (5) we know that smaller w means larger punishment for deleted edges, and larger w means larger punishment for inserted edges. If the network is sparse, large size cliques hardly exist. So if we want to detect certain communities a smaller w is needed. On the contrary, if the network is dense, a larger w is proper. That is to say, the first factor we should consider is that we select the weight w to be proportional to the average degree of the network. Speaking in details, for a network

with n nodes, suppose its overall number of edges is M. Set $D_e = M/n$, which can further be normalized to $\overline{D}_e = 1 - 1/D_e$, that is $0 \leq \overline{D}_e < 1$. We mean that the weight $w \propto \overline{D}_e$.

On the other hand, clustering coefficient is another important topological feature of a network. For the network with n nodes, the local clustering coefficient for an individual node i with d_i neighbors and K_i edges between its neighbors is $C_i = 2K_i/(d_i(d_i-1))$. We know that $0 \leq C_i \leq 1$. It is equal to 1 for a node at the center of a fully interlinked cluster, and 0 for a node that is part of a loosely connected group (star-like subgraph). Define C as the average of C_i over all i: $C = (1/n)\sum_{i=1}^{n} C_i$, which is called the average clustering coefficient of the network, and it is thought as a measure of the network's potential modularity [39, 40]. Here the second factor we consider for the selection of w is that it must be proportional to C, i.e., $w \propto C$ $(0 \leq C \leq 1)$.

Summarize the above considerations, we take the following form for the weight w

$$w = \frac{1}{2}(\overline{D}_e)^2 \cdot C, \tag{9}$$

if $\overline{D}_e \geq 0.5$ and $C \geq 0.5$. When $\overline{D}_e < 0.5$ or $C < 0.5$, which means that the network is comparatively sparser, now to detect certain communities the weight w needs to be properly adjusted to be more smaller. In these cases we take

$$w = \frac{1}{2}(\overline{D}_e)^{1/\overline{D}_e} \cdot C \tag{10}$$

or

$$w = \frac{1}{2}(\overline{D}_e)^2 \cdot C^{0.5/C}, \tag{11}$$

respectively.

4 Experiments

To examine the usefulness of our weighted parsimony model (WPM) (5)–(8) with the weight (9), (10) or (11) for community detection, several real-world networks are investigated. As in Subsection 2.2, the simulated annealing algorithm for module identification is used [35].

4.1 The karate club network

Here the karate club network is again investigated for illustrating the effectiveness of the WPM. This network has $\overline{D}_e = 0.5641$ and $C = 0.5706$. Unlike the partition in Subsection 2.2 where too many small size groups are obtained (Figure 1), now we get three groups with 5, 12 and 17 nodes, respectively (Figure 2). Dashed curve with label 1 in Figure 2 represents the actual division of original club. Our result is very close to it except node 10. Actually, this node is equally linked with the two parts of the original division. This may lead to the computational difference. As a matter of fact, several overlapping community detection algorithms detect it shared between the two parts [41, 42, 43]. Moreover, WPM can detect another small community composed of nodes 5, 6, 7, 11, 17 which only connect with node 1 in the original club. All these indicate that the application of our weighted model WPM to the empirically observed network can not only uncover its real situation, but also detect more complex substructure.

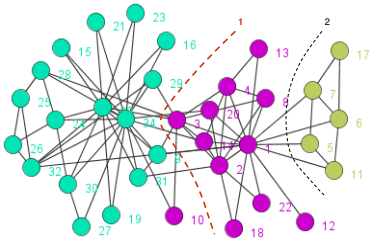

Figure 2: The community structure of the karate club network obtained by the weighted parsimony model WPM, three groups with three different colors.

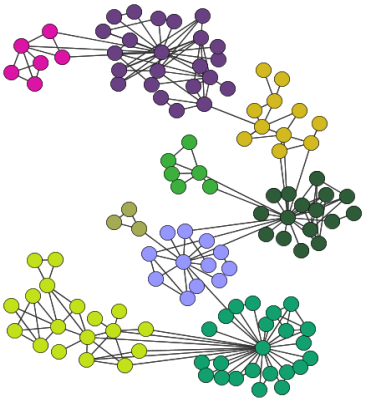

Figure 3: The community structure of scientific collaboration network obtained by the weighted parsimony model WPM.

4.2 The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [10] is another widely used test example for methods of detecting communities in complex networks [10, 38]. This network consists of 118 nodes (scientists) and 200 edges. It is a sparse network with $\overline{D}_e = 0.4100$ and $C = 0.6119$. It has higher average clustering coefficient C than the karate club network because it has more small size cliques.

With the weighted parsimony model WPM (5)–(8) and the weight (10) we can get 9 groups. Figure 3 shows the detected community structure which is visually very reasonable. This also indicates the advantage of the weighted model WPM to the original parsimony model where 64 groups are obtained for this scientific collaboration network (Subsection 2.2).

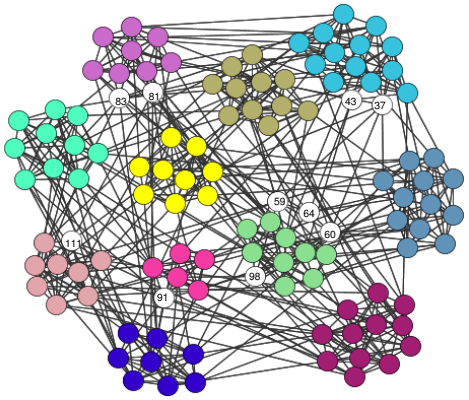

Figure 4: The community structure of football network obtained by the weighted parsimony model WPM.

4.3 The football team network

The third real network we have tested is the college football network of the United States, 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 links represent 613 games played in the course of the year. The teams are divided into 12 conferences of 8-12 teams each and, generally, games are more frequent between members of the same conference than between teams of different conferences. The natural community structure in the network makes it a commonly used benchmark for community-detecting algorithm testing [10, 38].

For this network $\overline{D}_e = 0.8124$, which means it is a very dense network (the average degree of every node is $2D_e = 10.6609$). But it has comparatively smaller average clustering coefficient $C = 0.4032$ because for such a network it must need more edges to form large size (for example, 10 or 11) cliques. Using our weighted parsimony model WPM (5)–(8) with the weight (11) eleven communities are detected (Figure 4). All of them correspond almost exactly to the original conferences except 10 nodes. Three nodes including 60, 64, 98 are classified in the other two groups for their weak link with their original group. Nodes 59 and 111 are respectively classified into inconsistent groups for their more links with current groups than their original ones just as the papers [10, 15, 30] have obtained. Because there are few edges among all five members of the 12th conference, these five nodes are distributed to other groups because they have more links with those groups. The community structure found by our model seems to suggest a more precise organization than original conferences.

4.4 The journal index network

The journal index network constructed by Rosvall and Bergstrom [45] consists of 40 journals as nodes from four different fields: physics, chemistry, biology and ecology, and 189 links connecting nodes if at least one article from one journal cites an article in the other journal during 2004. 10 journals with the highest impact factor in the four different

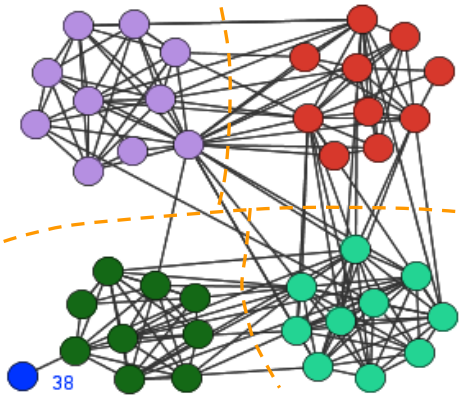

Figure 5: The community structure of journal index network obtained by the weighted parsimony model WPM.

fields were selected.

This is a dense network with $\overline{D}_e = 0.7884$, at the same time it has high clustering coefficient $C = 0.7103$. This means on average the nodes in this network have large degrees and there are several cliques within it. The exception exists for node 38 (the journal *Conservation Biology*) whose degree is only 1 (Figure 5). The orange dashed curves in Figure 5 indicate the original four fields. With our model we can detect essentially the same communities as the actual partition, only with the singly connected node 38 split off as another one. Nevertheless, any postprocessing can easily put this node into its original community.

4.5 The dolphin network

The dolphin social network reported by Lusseau et al. [44] and recently studied by Rosvall and Bergstrom [45] is also used here. This network consists of 62 nodes and 159 edges. The dashed curve in Figure 6 displays the division along which the actual dolphin groups were observed to split [44].

This network is also a little dense ($\overline{D}_e = 0.6101$), but it has very low average clustering coefficient ($C = 0.2590$). Using our model two groups are obtained. From Figure 6 we see that the partition is almost completely consistent with the actual division except the node 40, which is equally linked to the two parts. In fact, the recent overlapping community detection algorithm detects it shared between the two groups [43]. Moreover, reminded of the results in [45], where the authors illustrated the partitions of the same dolphin network using four methods, i.e., their cluster-based compression, the edge-betweenness algorithm [10], the spectral analysis approximation [13], and maximizing the modularity Q [11]. Each of them split the network into two parts. The first two methods get the same result as ours, but the third mislaid three nodes, and the fourth (i.e., maximizing Q) mislaid eight nodes.

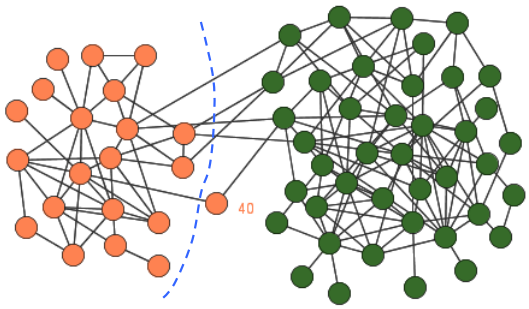

Figure 6: The community structure of dolphin network obtained by the weighted parsimony model WPM.

5 Discussion and conclusion

In this paper we propose a weighted parsimony criterion for community detection in complex networks. The main idea is that the community structure as the dense modules can be thought as isolated cliques by deleting and(/or) inserting minimal outer and(/or) inner connections. A positive parameter w as a weight coefficient is introduced to balance the inserted and deleted edges to get more reasonable community structure. The selection of w depends only on the topological structure of the network itself, i.e., the average degree $\overline{D}_e$ as well as the average clustering coefficient C. We use the weighted parsimony model to a series of real-world networks and satisfactory results are obtained. As a matter of fact, these networks belong to a wide category: some are with high $\overline{D}_e$ and high C (the karate club network and the journal index network), some are with high $\overline{D}_e$ and low C (the football team network and the dolphin network), and some are with low $\overline{D}_e$ and high C (the scientific collaboration network). This indicates that the weighted model can detect reasonable communities for a large kind of real-world networks.

Nevertheless, we couldn't find the real-world network with low $\overline{D}_e$ and low C, i.e., $0 < \overline{D}_e < 0.5$ and $0 < C < 0.5$, so we have no opportunity to evaluate the validity of the model for such network. The heuristic idea is that even in this situation the weight w is not appropriate to be too small to detect suitable communities. Perhaps (10) and (11) can be used for $\overline{D}_e \leq C < 0.5$ and $C \leq \overline{D}_e < 0.5$, respectively. Specially, the network with $\overline{D}_e \leq 0$ or $C = 0$ is rare in the real world, and the weighted parsimony model for this kind of network needs further study.

References

[1] S. H. Strogatz, Nature 410, 268-276 (2001).

[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. 74, 47-97 (2002).

[3] M. E. J. Newman, SIAM Rev. 45, 167-256 (2003).

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. 29(4), 251-262 (1999).

[5] J. Scott, Social Network Analysis: A Handbook, 2nd ed. (Sage Publications, London, 2000).

[6] M. E. J. Newman and J. Park, Phys. Rev. E 68, 036122 (2003).

[7] E. Almaas, B. Kov*á*cs, T. Vicsek, Z. N. Oltvai, and A.-L. Barab*á*si, Nature 427, 839-843 (2004).

[8] F. Rao and A. Caflisch, J. Mol. Biol. 342, 299-306 (2004).

[9] J. A. Dunne, R. J. Williams, and N. D. Martinez, Proc. Natl. Acad. Sci. USA 99, 12917-12922 (2002).

[10] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[11] M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).

[12] M. E. J. Newman, Eur. Phys. J. B 38, 321-330 (2004).

[13] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).

[14] M. E. J. Newman, Proc. Natl Acad. Sci. USA 103, 8577-8582 (2006).

[15] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang, and L. Chen, Phys. Rev. E 77, 036109 (2008).

[16] J. Zhang, S. Zhang, and X.-S. Zhang, Lecture Notes in Operations Research, 9, Optimization and Systems Biology, 290-299 (2008).

[17] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. 93, 218701 (2004).

[18] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).

[19] P. Ronhovde and Z. Nussinov, arXiv:0803.2548 [physics.soc-ph] (2008).

[20] G. Palla, I. Der*é*nyi, I. Farkas, and T. Vicsek, Nature 435, 814 (2005).

[21] J. M. Kumpula, M. Kivel*ä*, K. Kaski, and J. Saram*ä*ki, arXiv:0805.1449 [physics.soc-ph] (2008).

[22] R. Hildebrand, arXiv:0806.1896 [physics.data-an] (2008).

[23] A. Clauset, M. E. J. Newman, and C. Moore, in Proceedings of the 23rd International Conference on Machine Learning (Association of Computing Machinery, New York, 2006).

[24] B. Karrer, E. Levina, and M. E. J. Newman, Phys. Rev. E 77, 046119 (2008).

[25] W.Y.C.Chen, A.W.M.Dress, and W.Q.Yu, Mathematics in Computer Science 1, 441-457 (2008).

[26] W.Y.C.Chen, A.W.M.Dress, and W.Q.Yu, IET Syst. Biol. 1, 286-291 (2007).

[27] A.Clauset, C.Moore, and M.E.J.Newman, Nature 453, 98-101 (2008).

[28] W. E, T. Li, and E. Vanden-Eijnden, Proc. Natl. Acad. Sci. USA 105, 7907-7912 (2008).

[29] M.Rosvall and C.T.Bergstrom, Proc. Natl. Acad. Sci. USA 105, 1118-1123 (2008).

[30] J. Zhang, S. Zhang, and X.-S. Zhang, Physica A 387, 1675-1682 (2008).

[31] S. Fortunatoa and C. Castellano, arXiv:0712.2716 [physics.soc-ph] (2007).

[32] M. A. Porter, J.-P. Onnela, and P. J. Mucha, arXiv:0902.3788 [physics.soc-ph] (2009).

[33] M. Gr*ö*tschel and Y. Wakabayashi, Mathematical Programming 45, 59-96 (1989).

[34] M. Gr*ö*tschel and Y. Wakabayashi, Mathematical Programming 47, 367-387 (1990).

[35] R. Guimer*à* and L. A. N. Amaral, Nature 433, 895-900 (2005).

[36] W. W. Zachary, J. Anthropol.Res. 33, 452-473 (1977).

[37] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl Acad. Sci. USA 101, 2658-2663 (2004).

[38] F. Wu and B. A. Huberman, Eur. Phys. J. B 38, 331-338 (2004).

[39] D. J. Watts and S. Strogatz, Nature 393, 440-442 (1998).

[40] E. Ravasz,A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barab*á*si, Science 297, 1551-1555 (2002).

[41] S. Zhang, R.S. Wang, and X.-S. Zhang, Physica A 374, 483-490 (2007).

[42] S. Zhang, R.S. Wang, and X.-S. Zhang, Phys. Rev. E 76, 046103 (2007).

[43] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. 11, 033015 (2009).

[44] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson, Behav. Ecol. Sociobiol. 54, 396-405 (2003).

[45] M. Rosvall and C.T. Bergstrom, Proc. Natl. Acad. Sci. USA 104, 7327-7331 (2007).