Identifying and Characterizing Nodes Important to Community Structure Using the Spectrum of the Graph

Yang Wang, Zengru Di, Ying Fan*

Department of Systems Science, School of Management and Center for Complexity Research, Beijing Normal University, Beijing, People's Republic of China

Abstract

Background: Many complex systems can be represented as networks, and how a network breaks up into subnetworks or communities is of wide interest. However, the development of a method to detect nodes important to communities that is both fast and accurate is a very challenging and open problem.

Methodology/Principal Findings: In this manuscript, we introduce a new approach to characterize the node importance to communities. First, a centrality metric is proposed to measure the importance of network nodes to community structure using the spectrum of the adjacency matrix. We define the node importance to communities as the relative change in the eigenvalues of the network adjacency matrix upon their removal. Second, we also propose an index to distinguish two kinds of important nodes in communities, i.e., "community core" and "bridge".

Conclusions/Significance: Our indices are only relied on the spectrum of the graph matrix. They are applied in many artificial networks as well as many real-world networks. This new methodology gives us a basic approach to solve this challenging problem and provides a realistic result.

Citation: Wang Y, Di Z, Fan Y (2011) Identifying and Characterizing Nodes Important to Community Structure Using the Spectrum of the Graph. PLoS ONE 6(11): e27418. doi:10.1371/journal.pone.0027418

Editor: Peter Csermely, Semmelweis University, Hungary

Received June 25, 2011; **Accepted** October 16, 2011; **Published** November 14, 2011

Funding: This work is supported by the Natural Science Foundation of China (NSFC) under grants No. 70771011 and No. 60974084, the Program for New Century Excellent Talents in University of Ministry of Education of China (No. NCET-09-0228), and fundamental research funds for the Central Universities of Beijing Normal University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yfan@bnu.edu.cn

Introduction

Networks, despite their simplicity, represent the interaction structure among components in a wide range of real complex systems, from social relationships among individuals, to interactions of proteins in biological systems, to the interdependence of function calls in large software projects. The network concept has been developed as an important tool for analyzing the relationship of structure and function for many complex systems in the last decades[1–5]. Many real-world systems show the existence of structural modules that play significant and defined functional roles, such as friend groups in social networks, thematic clusters on the world wide web, functional groups in biochemical or neural networks [6]. Exploring network communities is important for the reasons listed below [7]: 1) communities reveal the network at a coarse level, 2) communities provide a new aspect for understanding dynamic processes occurring in the network and 3) communities uncover relationships among the nodes that, although they can typically be attributed to the function of the system, are not apparent when inspecting the graph as a whole. As a result, it is not surprising that recent years have witnessed an explosion of research on community structure in graphs, and a huge number of methods or techniques have been designed [6,8–17](see [9] as a review).

It is believed that community structure is important to the function of a system [18–20]. In many situations, it might be desirable to control the function of modular networks by adjusting the structure of communities. For example, in biological systems, one might like to identify the nodes that are key to communities and protect them or disrupt them, such as in the case of lung cancer [19]. In epidemic spreading, one would like to find the important nodes to understand the dynamic processes, which could yield an efficient method to immunize modular networks [20]. Such strategies would greatly benefit from a quantitative characterization of the node importance to community structure. Some important work related to this topic has been proposed. In 2006, Newman proposed a community-based metric called "Community Centrality" to measure node importance to communities [8]. His basic idea relies on the modularity function Q. Those vertices that contribute more to Q are more important for the communities than those vertices that contribute less. Kovacs et al. also proposed an influence function to measure the node importance to communities [21].

In fact, the important nodes can have distinct functions with respect to community structure. Some previous studies have also revealed such classifications. Guimera et al. have proposed a classification of the nodes based on their roles within communities, using their within-module degree and their participation coefficient [22]. They divided the hubs into three categories: provincial hubs, connector hubs and kinless hubs. Other approaches have also been suggested to discuss the connection between nodes and modularity in biological networks, by dividing hub nodes into two categories called "party hubs" and "date hubs" [23–25]. When removed from the network, party and date hubs have strikingly distinct effects on the overall topology of the network. Recently, Kovacs et al. proposed an interesting approach. They introduced an integrative method family to detect the key nodes, overlapping communities and "date" and "party" hubs [21]. In a very recent work, the authors mentioned that modular networks naturally allow the formation of clusters, and hubs connecting the modules would enhance the integration of the whole network, such as in the case of neuron networks [26]. As a result, it is intuitive that nodes that are important to communities can be divided into "community cores" and "bridges". However, using the previous methods such as participation coefficient and the influence function to distinguish these two kinds of vertices, the exact communities of the network must first be given [21,22]. In contrast, it is interesting to characterize node importance to communities without knowing the exact partition of the network.

It is understood that the adjacency matrix contains all the information of the network. Developing methods based only on the adjacency matrix of the network to detect important nodes to communities and then distinguish them as either "community core" or "bridge" is an interesting and important problem in network research. In this manuscript, based only on the adjacency matrix of the network, we try to access the fundamental questions: how to evaluate the node importance to communities and how to distinguish different kinds of important nodes? It is implied that in many cases the spectrum of the adjacency matrix gives an indication of the community structure in the network [27]. If the network has c strong communities, the c largest eigenvalues of the adjacency matrix are significantly larger than the magnitudes of all the other eigenvalues. These large eigenvalues are key quantities to the community structure. For this reason, we suggest a basic approach to solve the above open problem using the spectrum of the graph. We define the importance of nodes to communities as the relative change in the c largest eigenvalues of the network adjacency matrix upon their removal. Furthermore, using the eigenvectors of the graph Laplacian, we divide the important nodes into community cores and bridges. We apply our method to many networks, including artificial networks and real-world networks. This new methodology gives us a basic approach to solve this challenging problem and provides a realistic result.

Methods

Centrality Metric Based on the Spectrum of the Adjacency Matrix

We consider a binary network $G=(V,E)$ with n nodes. The adjacency matrix A is the matrix with elements $A_{ij}=1$ if there is an edge joining vertices i and j, otherwise 0. We denote each eigenvalue of A by λ and the corresponding eigenvector by v, such that $Av=\lambda v$. The eigenvectors are orthogonal and normalized. The eigenvalues are ordered by decreasing magnitude: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. It is easy to show that A is symmetric and the eigenvalues of A are real. Consider the case of networks that have c communities. It is implied that when these communities are disconnected, each one has its own largest eigenvalue. With proper labeling of the nodes, the matrix A will have a block matrix structure with $c \times c$ blocks. Blocks on the diagonal correspond to the adjacency matrices of the individual communities, while the off-diagonal blocks correspond to the edges between communities; in other words, we can consider them as a perturbation. Therefore, A can be written as

$$A = A_0 + \delta A, \tag{1}$$

where A_0 is a matrix whose diagonal block elements are the diagonal block elements of A and whose off-diagonal block elements are zeros, while δA is a matrix with zeros on its diagonal blocks and with the off-diagonal blocks of A as its off-diagonal block elements. Chauhan et al. have proved that if the perturbation strength is small, the largest eigenvalues of disconnected communities are perturbed more weakly than the perturbation applied [27]. The spectrum of the adjacency matrix of a network gives a clear indication of the number of communities in the network. If the network has c strong communities, the c largest eigenvalues are well separated from others. These eigenvalues are key quantities to the community structure.

For this reason, we define the importance of node k to communities as the relative change in the c largest eigenvalues of the network adjacency matrix upon its removal:

$$P_k = -\sum_{i=1}^{c} \frac{\Delta\lambda_i}{\lambda_i}, \tag{2}$$

where c is the number of communities. To avoid the computational cost, we use perturbation theory to provide approximations of P_k in terms of the corresponding eigenvector v. Let us denote the matrix before the removal of the node by A and the matrix after the removal by $A+\Delta A$; the eigenvalue of this matrix is $\lambda+\Delta\lambda$, and the corresponding eigenvector is $v+\Delta v$. For large matrices, it is reasonable to assume that the removal of a node has a small effect on the whole matrix and the spectral properties of the network, so that ΔA and $\Delta\lambda$ are small. We obtain

$$(A+\Delta A)(v+\Delta v) = (\lambda+\Delta\lambda)(v+\Delta v). \tag{3}$$

The effect on the adjacency matrix A of removing node k is given by $(\Delta A)_{ij} = -A_{ij}(\delta_{ik}+\delta_{jk})$. We cannot assume that the Δv is small because $\varDelta v_k = -v_k$, so we set $\Delta v = \delta v - v_k k$ where δv is small and is the unit vector for the k component. Left multiplying (3) by v^T and neglecting second order terms $v^T \Delta A \delta v$ and $v^T \Delta\lambda \delta v$, we obtain

$$\Delta\lambda = \frac{v^T \Delta A v - v^T v_k \Delta A \hat{e}_k}{v^T v - v_k^2}. \tag{4}$$

For a large network ($n \gg 1$), we know that $v^T v \gg v_k^2$; therefore, we can write

$$\Delta\lambda \approx \frac{v^T \Delta A v - v^T v_k \Delta A \hat{e}_k}{v^T v} \tag{5}$$

Because $(\Delta A)_{ij} = -A_{ij}(\delta_{ik}+\delta_{jk})$, we obtain

$$v^T \Delta A v = -2\lambda v_k^2, v^T v_k \Delta A \hat{e}_k = -\lambda v_k^2. \tag{6}$$

Finally, the importance of node k to the community structure is obtained by

$$P_k = -\sum_{i=1}^{c} \frac{\varDelta\lambda_i}{\lambda_i} \approx \sum_{i=1}^{c} \frac{v_{ik}^2}{v_i^T v_i}, \tag{7}$$

where c is the number of communities, v_{ik} is the kth element of v_i and P_k lies in the interval [0,1]. If P_k is large, node k is important to the community structure; otherwise, k is on the periphery of the community.

If a network which has n nodes and c communities, it indicates that $\sum_{k=1}^{n} P_k = c$. In order to let the sum of the index scales to 1, we define the new index as $I_k = P_k / c$ that obeys $\sum_{k=1}^{n} I_k = 1$. Then we consider an ER random network with n nodes as a null model, the network is homogeneous and there expects no important nodes to communities. So the index of each node in the null model would be $1/n$. Thus $1/n$ could be a criterion to evaluate the significance of the nodes. If index I of a node is large than $1/n$ we consider it as important nodes.

Using this metric I, we can quantify the node importance to the community structure. If the node is important to the community structure, when we remove it from the network, the relative changes of the c largest eigenvalues are large; otherwise, the changes are small. Before applying I, the value of c needs to be determined. The determination of the number of communities is important in community analysis and still open for researchers. Generally speaking, every algorithm for detecting communities should have a method to give the best number of the partition. So there are already some suggestions to determine the number of communities [9]. Using the spectrum of the graph is also an easy way to detect the optimal number of the communities [27,28]. If c is given, our method can characterize the node importance to communities without knowing the exact partition of the network.

Distinguish Two Kinds of Important Nodes

As mentioned above, there are two kinds of nodes that are important to communities. One is the "community core", and the other is the "bridge" between communities. Each will affect communities deeply upon its removal. When we remove the "community core", the community structure in the network will become fuzzy, while the community structure will become clear when we remove the "bridge". See Fig. 1 for an example. Vertices 1 and 8 are the "community cores", and they organize their respective communities. Meanwhile, node 15 is the "bridge" between the two communities. The "community core" is the leader in the community, and it can organize the function of each community. In contrast, the "bridge" connects the modules and can enhance the integration of the whole network. It is believed that a combination of both segregation and integration, such as in neural systems, is crucial [26]. It is clear that effectively disconnected and fully non-synchronous regions cannot allow collective or integrative action of the elements. Similarly, a fully synchronized regime does not allow separated or segregated performance of the elements. Therefore, both situations are biologically unrealistic, as can be seen from the existence of related conditions, such as epileptic seizures (collective phenomena) and Parkinson's disease (segregated phenomena) [29]. For this reason, both the "community core" and the "bridge" are important to communities, but they play different roles. The metric I we proposed before can determine the nodes that are important to communities, but now a method to distinguish these two kinds of important nodes is needed.

In agreement with earlier findings [21,23–25], we assumed that bridge nodes should have more inter-modular positions than community cores. The existence of bridge nodes often leads to some inter-modular edges. Given a graph, the simplest and most direct way to construct a partition of the graph is to solve the mincut problem (minimize the number of edges between communities R) [30]. In practice, however, this method often does not lead to satisfactory partitions. The problem is that, in many cases, the solution of mincut simply separates one individual vertex from the rest of the graph. Of course, this is not what we want to achieve in clustering, as clusters should be reasonably large groups of points. Due to this shortcoming in the mincut problem, one common objective function to encode the desired information is RatioCut [31]:

$$RatioCut(C_1,\cdots C_c) \doteq \sum_{i=1}^{c} \frac{R(C_i,\bar{C}_i)}{|C_i|}, \qquad (8)$$

where $|C_i|$ is the size of community C_i. If the sizes of the communities are almost the same, the RatioCut problem reduces to the mincut problem.

The Condition of $c=2$. If the network is divided into only two communities ($c=2$), we define an index vector s with N elements:

$$s_i = \begin{cases} \sqrt{|\bar{C}|/|C|} & \text{if vertex } i \in C, \\ -\sqrt{|C|/|\bar{C}|} & \text{if vertex } i \in \bar{C}. \end{cases} \qquad (9)$$

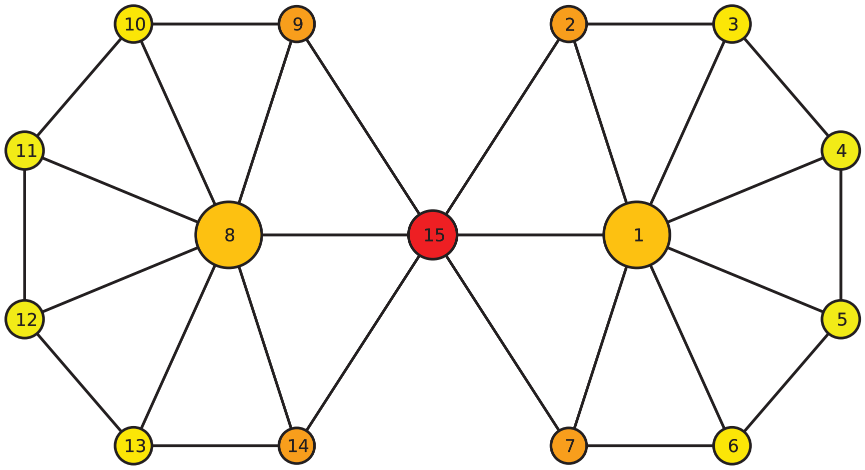

Figure 1. Sketch of a network composed of 15 nodes. The diameter of one vertex is proportional to the centrality metric I. Moreover, the color of one vertex is related to the index w-score. Red vertices behave like "overlapping" nodes or "bridges" between communities, and yellow vertices often lie inside their own communities.
doi:10.1371/journal.pone.0027418.g001

Then the RatioCut function is obtained as follows [28]:

$$RatioCut(C,\bar{C}) = \frac{1}{|V|}\mathbf{s}^T \boldsymbol{L}\mathbf{s}, \tag{10}$$

where $|V|$ is the number of vertices in the network and $\boldsymbol{L}$ is the graph Laplacian. $\boldsymbol{L}$ is defined as $L_{ij} = -A_{ij}$ for $i \neq j$ and $L_{ii} = k_i$, where k_i is the degree of node i. We also have two constraints on $\boldsymbol{s}$: $\sum_{i=1}^{n} s_i = 0$ and $\sum_{i=1}^{n} s_i^2 = n$. Here the partition problem is equal to the problem

$$\min \mathbf{s}^T \boldsymbol{L}\mathbf{s};\ \text{subject to} \sum_{i=1}^{n} s_i = 0, \sum_{i=1}^{n} s_i^2 = n. \tag{11}$$

If the components of the vector $\boldsymbol{s}$ are allowed to take arbitrary values, it can be seen immediately that the solution of this problem is given by the vector s that is the eigenvector corresponding to the second-smallest eigenvalue of $\boldsymbol{L}$, denoted by $\boldsymbol{u}_2$. So we can approximate a minimizer of RatioCut by the second eigenvector of $\boldsymbol{L}$. Unfortunately, the components of s are only allowed to take two particular values.

Thus, the simplest solution is achieved by assigning vertices to one of the groups according to the sign of the eigenvector $\boldsymbol{u}_2$. In other words, we assign vertices as follows: if $\boldsymbol{u}_2^i > 0$, we assign vertex i to community C; otherwise, we assign it to $\bar{C}$. Assignation priority begins with the most positive and the most negative; the node with the most positive magnitude is first to be assigned to C, then the second and so on, while the node with the most negative magnitude is similarly the first to be assigned to $\bar{C}$. If a node's corresponding element is close to zero, it may have nearly equal membership in both communities, and we can assign it to both communities. In conclusion, if the network is divided into only two communities, we can use this method to characterize which are the "community cores" and which are the "bridge" between communities. If node i is a "community core", $|\boldsymbol{u}_2^i|$ is relatively large; otherwise, $|\boldsymbol{u}_2^i|$ is near zero.

The Condition of $c > 2$. Consider the division of a network into c nonoverlapping communities, where c is the number of communities. We define an $n \times c$-index matrix S with one column for each community, $\boldsymbol{S} = (\boldsymbol{s}_1|\boldsymbol{s}_2|\cdots|\boldsymbol{s}_c)$, by

$$s_{i,j} = \begin{cases} 1/\sqrt{|C_j|} & \text{if} \quad \text{vertex} \quad i \in C_j, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Following the previous section, we obtain

$$RatioCut = Tr(\boldsymbol{S}^T \boldsymbol{L}\boldsymbol{S}), \tag{13}$$

where Tr is the trace of a matrix and $\boldsymbol{S}^T$ is the transpose matrix of $\boldsymbol{S}$. $\boldsymbol{L}$ is a semi-positive and symmetric matrix. We can write $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is the eigenvector of $\boldsymbol{L}$, $\boldsymbol{U} = (\boldsymbol{u}_1|\boldsymbol{u}_2|\cdots|\boldsymbol{u}_n)$ and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues $D_{ii} = \beta_i$. We therefore obtain

$$RatioCut = \sum_{j=1}^{n}\sum_{k=1}^{c} \beta_j (u_j^T s_k)^2. \tag{14}$$

It can also be written as

$$RatioCut = \sum_{k=1}^{c}\sum_{j=1}^{n} \beta_j [\sum_{i=1}^{n} U_{ij} S_{ik}]^2. \tag{15}$$

Now we define the vertex vector of i as r_i, and let

$$[r_i]_j = U_{ij}. \tag{16}$$

If the network has almost equal-sized communities, then equation (15) can be written as

$$RatioCut \approx \frac{\sum_{k=1}^{c}\sum_{j=1}^{n} \beta_j [\sum_{i \in G_k} [r_i]_j]^2}{|C|}, \tag{17}$$

where G_k is the set of vertices belonging to community k and $|C|$ is the community size.

Minimizing the RatioCut can be equated with the task of choosing the nonnegative quantities so as to place as much of the weight as possible in the terms corresponding to the low eigenvalues and as little as possible in the terms corresponding to the high eigenvalues. This equates to the following maximization problem:

$$Max \sum_{k=1}^{c}\sum_{j=1}^{p} \beta_j [\sum_{i \in G_k} [r_i]_j]^2, \tag{18}$$

where p is a parameter. We could choose $p = c$ if the community structure was clear. To this end, we propose an easy way to distinguish two kinds of important nodes using the theory of the graph Laplacian. If the community structure is quite clear, we focus on the *vertex vector magnitude* $|r_i|$ in the first p terms, denoted by the b:

$$b_i = \sum_{j=1}^{p} [r_i]_j^2. \tag{19}$$

If the index b of a given vertex is nearly zero, it indicates that the presence of that node results in a large RatioCut. Thus it is considered as a "bridge" node. Moreover, it also need to state the criterion of the index b. The same as P_k in Eq. (7), for a network with n nodes and c communities, it indicates that $\sum_{k=1}^{n} b_k = c$. We can also define the new index as $w_k = b_k/c$ and then $\sum_{k=1}^{n} w_k = 1$. Then we consider an ER random network with n nodes as a null model, the network is homogeneous and there expects no "bridge" nodes to communities. So the index of each node in the null model would be $1/n$. Thus $1/n$ could also be a criterion to evaluate the "bridgeness" of the nodes. If the w-score of a given vertex is smaller than $1/n$, we believe that this vertex has nearly equal membership in more than one community, and it is likely to be the "bridge" of these communities. This discrimination process equates to the "fuzzy" division of the network into communities. In many cases, this type of fuzzy division could result in a more accurate picture of real-world networks.

Our method requires less computational cost than other methods. Since most of the real-world network is sparse, combining the Lanczos and QL algorithms, we expect to be able to find all eigenvalues and eigenvectors of a sparse symmetric matrix in time $O(mn)$, where m and n is the number of edges and nodes, respectively [32]. On the other hand, the method proposed in Ref. [8] is slower than ours since the modularity matrix is not sparse. So from this point of view, our method has the advantage compared with the method proposed in Ref. [8]. On the other hand, the method proposed by Ref. [21] has runtime complexity $O(n(n+m))$ and $O(m(n+m))$.

Results

Now we test the validity of our indices I and w-score introduced before in various artificial networks and real-world networks.

Artificial Networks

First, we consider a sketch composed of 15 nodes (see Fig. 1) formed by two communities. It is intuitive that vertices 1, 8 and 15 are important to the community structure in this sketch. Vertices 1 and 8 are the so-called "community cores", and they organize both the communities. Vertex 15 is the "bridge" between communities, and it connects these two communities. As we discussed before, removing vertex 1 or 8 will make the community structure fuzzy, and removing vertex 15 will make it clear.

Here we use the index H proposed by Hu et al.[14] to measure the significance of communities:

$$H=\frac{n}{\bar{k}\sum_{j=c+1}^{n}\frac{1}{|\overline{\beta}-\beta_j|}}, \qquad (20)$$

where β is the eigenvalue of the graph Laplacian, $\overline{\beta}$ is the average value of β_2 through β_c, $\bar{k}$ is the average degree of the network and n is the number of vertices in the network. In networks with strong communities (many links are within communities with very sparse connections outside), H is always large. Here we focus on the change of H due to the removal of vertices, denoted by ΔH. We also use the centrality metric proposed by Newman [8], which we denote here by M. The results are shown in Tab. 1. Through ΔH, it is implied that vertices 1 and 8 are more important than other vertices because the magnitude of ΔH is relatively larger than others. Moreover, their removal makes the communities fuzzy, while vertex 15 acts like a "bridge" between the communities, and its removal makes the communities clear. We can see that our centrality metric performs quite well; it can identify not only the "community cores", but also the "bridge" between communities. M can also identify the "community cores", but it has some problems. One issue is that its values tend to span a rather small dynamic range from largest to smallest. Moreover, in some cases (such as this sketch), M cannot recognize important vertices among communities. In calculating the index H, we need to go through every vertex in the network, incurring significant computational cost. In contrast, our method provides a more efficient way, requiring less computational cost, and yields the correct answer.

Here we use the classical **GN benchmark** presented by Girvens and Newman to test the measurements [12]. Each network has $N=128$ nodes that are divided into four communities ($c=4$) with 32 nodes each. Edges between two nodes are introduced with different probabilities, which depend on whether the two nodes belong to the same community or not. Each node has $<k_{in}>$ links on average with its fellows in the same community and $<k_{out}>$ links with the other communities, and we impose $<k_{in}>+<k_{out}>=16$. The communities become fuzzier and thus more difficult to identify as k_{out} increases. Because the GN benchmark is a homogenous network, there should not be any nodes that are important to the community structure. To check whether our conjecture is correct or not, we let $<k_{in}>=12$ so that the community structure is quite clear and average the result for the GN benchmark over 100 configurations of networks. From the result, all the nodes' index I lie in the interval [0.007,0.008]. The mean value of I is 0.0078, and the standard deviation is 0.0008. It can be concluded that, in the GN benchmark, there are no nodes that are important to the community structure.

Table 1. Centrality metrics of the example sketched in Fig. 1.

Vertex Label	I	M	ΔH	w-score
1	0.16	0.758	-0.145	0.0623
8	0.16	0.758	-0.145	0.0623
15	0.086	0.69	0.116	0.0333
2,7,9,14	0.045	0.704	0.04	0.0529
3,6,10,13	0.05	0.7535	-0.021	0.0739
4,5,11,12	0.052	0.7327	-0.054	0.0837

doi:10.1371/journal.pone.0027418.t001

We may also test the method on the more challenging **LFR benchmark** presented by Lancichinetti et al.[33]. In the LFR benchmark, the degree distribution obeys a power-law distribution $p(k)\propto k^{-\alpha}$, and the sizes of the communities are also taken from a power-law distribution with an exponent γ. Moreover, each node shares a fraction $1-\mu$ of its links with other nodes of its own community and a fraction μ with others in the rest of the network. The community structure can be adjusted by the mixing parameter μ. Without loss of generality, we let $\alpha=2.5,\gamma=1.0,\mu=0.25$ and the size of the network $N=1000$. Our numerical results in the LFR benchmark are shown in Fig. 2. In this case, there is no "bridge" between communities because $\mu=0.25$. We may also calculate the w-score, of which the mean value is 0.001 and the standard deviation is 2.5×10^{-4}. which indicates that there is no obvious "bridge" nodes in LFR benchmark. Moreover, the centrality metric is positively correlated with node degree ($r^2=0.907$), but some vertices have quite high centrality while having relatively low degree, and thus the correlation index is not very high. Moreover, we have varied the

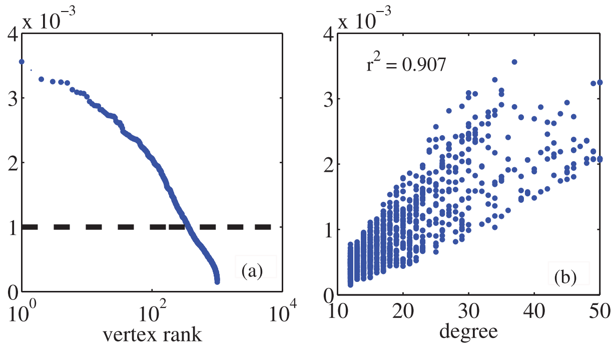

Figure 2. The distribution of index I and the correlation between I and node degree k in LFR benchmark.(a) The Zipf plot of the nodes' centrality to communities. The dash line indicates the threshold $1/n$. (b) The centrality metric we propose is correlated with node degree. The parameters in the LFR benchmark are as follows: $\alpha=2.5,\gamma=1.0,\mu=0.25$ and the size of the network $N=1000$.
doi:10.1371/journal.pone.0027418.g002

parameter μ in the LFR benchmark and given the changes of indices with the change of μ. In the related calculations, we used the predetermined number of communities as the c in the metrics. Because if $\mu > 0.5$ the whole network becomes fuzzy and how to determine the community number c is a tough problem. We consider the largest degree nodes in both the biggest and the smallest communities and the results are obtained by averaging over 20 independent realizations. From the result in Fig. 3, it is implied that with the network become fuzzy, the index I of the largest degree nodes in both the biggest and the smallest communities tend to become bigger while the index w-score becomes smaller.

Real-world Networks

We apply our method to some real-world networks, such as the Zachary club network [34], the word association network [35], the scientific collaboration network [36], and the C. elegans neural network [37].

First, we consider a famous example of a social network, the **Zachary's karate club network**. This network represents the pattern of friendships among members of a karate club at a North American university. It contains 34 vertices, and the links between vertices are the friendships between people. The nodes labeled as 1 and 34 correspond to the club instructor and the administrator, respectively. They had a conflict which resulted in the breakup of the club. Most other nodes have a relationship with node 1, node 34, or both. In this network, $c=2$. The numerical results are shown in Fig. 4 and Fig. 5. In Fig. 4(a), we can see that nodes 1 and 34 are the most important nodes in the communities. Our method to distinguish important nodes are shown in Fig. 4(b). Node 3 is considered as a "bridge" node between communities and displays a smaller value of w-score. Moreover, we compared the "bridge" nodes with overlapping nodes found by the method suggested in Ref. [38]. We found that the two results are usually consistent with each other. That means the bridges are usually overlapping nodes, such as node 3. However, there are some differences. For instance, our method considers vertex 14 as a bridge node while in Ref. [38] the authors doesn't consider it as an overlapping node. However, vertex 14 has the degree 5 and it links both communities so considering it as a bridge node is also acceptable. From what we discussed before, bridge nodes are more likely to be overlapping nodes. Furthermore, we compare our method with Newman's. This result is also shown in Fig. 4(a), and the two metrics are normalized by

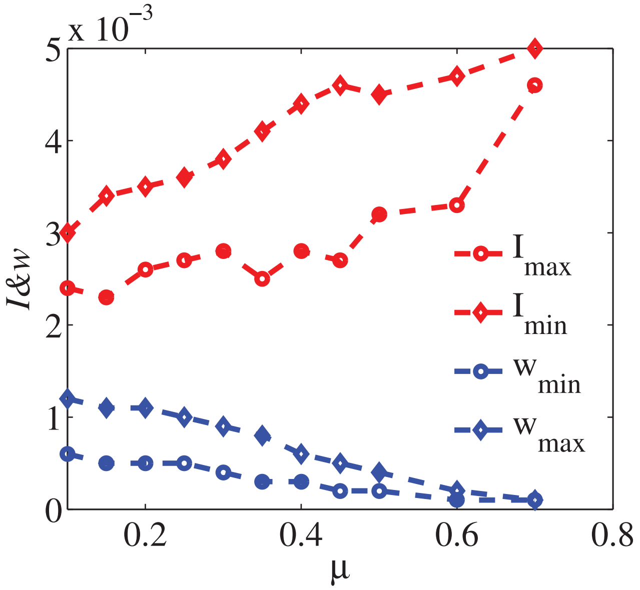

Figure 3. The indices I and w-score as a function of the parameter μ in LFR benchmark. The parameters in the LFR benchmark are as follows: $\alpha=2.5, \gamma=1.0$ and the size of the network $N=1000$. The results are obtained by averaging over 20 independent realizations.
doi:10.1371/journal.pone.0027418.g003

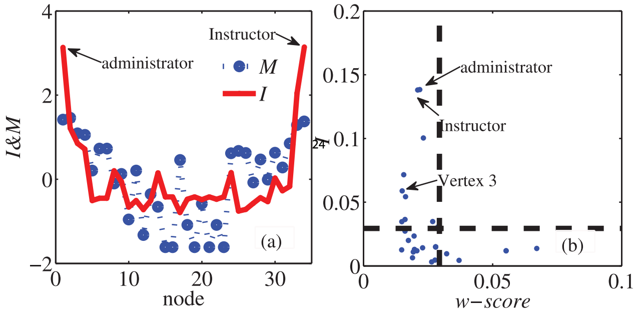

Figure 4. The usage of our method in Zachary's karate club network. It is shown that our method works quite well. Nodes 1 and 34 are the instructor and the administrator, respectively. In Fig. 4(a), we can see that these two nodes are more important to the community structure than other nodes. We also compare our method with Newman's and find that the two methods exhibit some differences. In Fig. 4(b), it is implied that Node 3 is likely to be a "bridge" node since it displays a rather low w-score.
doi:10.1371/journal.pone.0027418.g004

$$x_{nor} = \frac{x - <x>}{\sigma_x}, \qquad (21)$$

where $<x>$ is the average value of each index and σ_x is the standard deviation of each index. It is implied that these two methods have some differences. In our method, nodes 1 and 34 are absolutely more important than other nodes, while in Newman's method, nodes 2 and 33 are also quite important, even more than node 1. In this network, the modularity function Q reaches its maximum value when the network is divided into 4 communities; this fact may be the cause of the differences between the results of these two methods. The visualization of the karate network with our two measurements is sketched in Fig. 5. The diameter of each vertex is proportional to the centrality metric I. A large diameter indicates an important vertex. Additionally, the color of each vertex is related to the index w-score. Red vertices behave like "overlapping" nodes or "bridges" between communities, and yellow vertices often lie inside their own communities.

Second, we analyze the **word association network** starting from the word "Bright". This network was built on the University of South Florida Free Association Norms [35]. An edge between words A and B indicates that some people associate the word B to the word A. The graph displays four communities, corresponding to the categories *Intelligence, Astronomy, Light, Colors*. The word *Bright* is related to all of them by construction. We applied our method to this network, and the results are shown in Fig. 6. From the results, we can observe that our method considers *Bright, Sun, Smart, Moon* as important nodes to the community structure. It may be inferred from the result that *Moon* and *Smart* are the "community cores", while *Bright* and *Sun* are the "bridges" between communities. Indeed, our metric yields the correct answer. For example, *Smart* is the core of the community *Intelligence*, while *Moon* is the core of the community *Astronomy*. Meanwhile, the w-score of node *Bright* is

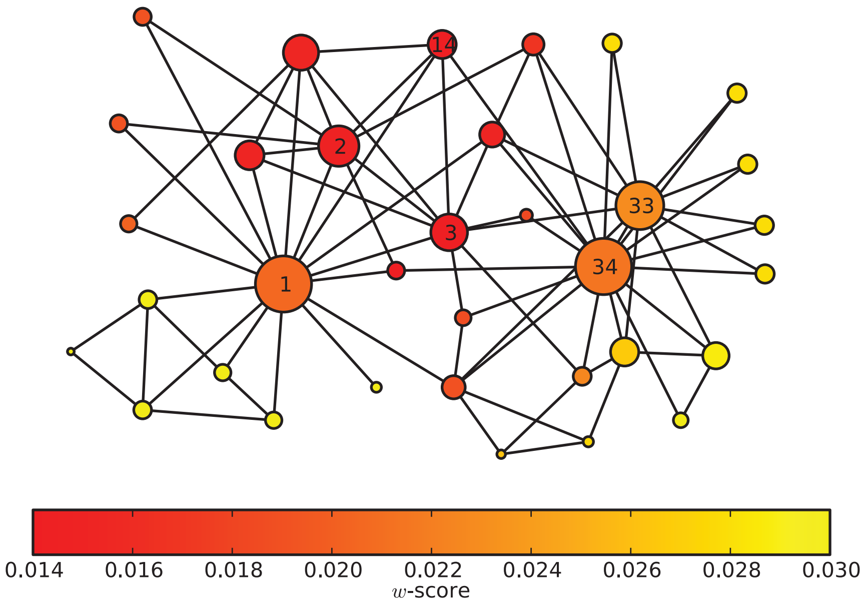

Figure 5. Sketch of the Zachary's karate club network, which is composed of 34 vertices. Vertex diameters indicate the community centrality I. The color of each vertex is proportional to the index w-score.
doi:10.1371/journal.pone.0027418.g005

0.006, which is close to zero. We would therefore conclude that it is a "bridge" between communities, and *Bright* is in fact the "bridge" among these four communities, as the network was originally derived from it. Moreover, we have investigated the effect of node removal on the indices Q and H and the results show that the removal of "community core" makes the network fuzzy while the community structure becomes clear when the "bridge" is removed.

We may also apply our method to social networks, such as the **scientist collaboration network** [36], and neural networks, such as the **C. elegans neural network** [37]. We analyzed the largest connected component of each network. The scientist collaboration network represents scientists whose research centers on the properties of networks of one kind or another. There are 379 vertices, representing scientists who are divided into 12 communities. Edges are placed between scientists who have published at least one paper together. The neural network of C. elegans contains 302 neurons and 2,359 links. This network is divided into 3 communities, with each node representing a neuron and each link representing a synaptic connection between neurons. Here we consider the C. elegans neural network to be undirected. The results are shown in Fig. 7.

In the scientist collaboration network, our centrality metric I identifies "group leaders", such as M. Newman, S. Boccaletti, and A. Barabasi. Their w-scores are not very large because they often have some collaboration between scientists outside their own communities. We can also find so-called "community cores" based on our method, such as R. Sole, and "bridge" vertices among

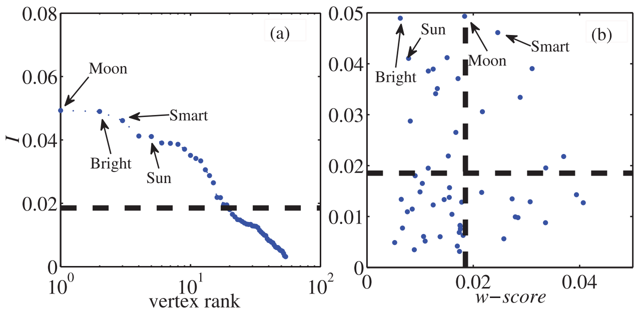

Figure 6. Index I and w-score for the nodes of the word association network. The node importance versus vertex rank is shown in (a). In (b), we distinguish "community cores" and "bridges" using the index w-score.
doi:10.1371/journal.pone.0027418.g006

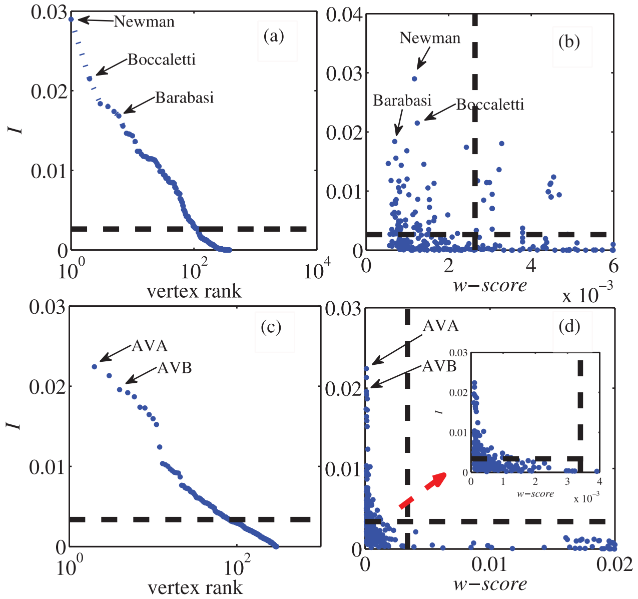

Figure 7. The usage of our method in scientist collaboration network and C. elegans neural network. The centrality metric I and w-score for the scientist collaboration network (a,b). The centrality metric I and w-score are also calculated in the C. elegans neural network (c,d).
doi:10.1371/journal.pone.0027418.g007

Table 2. Centrality metrics I and w-score in a complete weighted network.

Vertex Label	I	w-score
4	0.15	0.0955
9	0.15	0.0955
11	0.067	0.0455
others	0.079	0.0955

doi:10.1371/journal.pone.0027418.t002

some communities, such as B. Kahng. As we know, the C. elegans neural networks are composed of sensory neurons, interneurons and motor neurons. The neurons with high centrality metrics often have the most important functions, and all of them are interneurons, such as AVA, AVB, AVD, and AVE. These classes, which synapse onto motor neurons in the ventral cord, are among the most prominent neurons in the whole nervous system. They generally have larger-diameter processes than other neurons and have many synaptic connections [37,39]. As a result, they have larger I than other vertices, while the typical w-score in these classes is quite small. In the C. elegans neural network, most of the important nodes are likely to be "bridge" nodes since the connection between communities is more necessary and frequent due to some special functions.

Applications in Weighted networks

Our method can be generalized to weighted networks because the adjacency matrix in an undirected weighted network is real and symmetric. Thus, in weighted networks, the importance of a node and its role in communities are also characterized by its I and w-score. Let us first consider an artificial weighted network. We use similarity weight in this weighted network. A higher weight means a closer relationship between vertices. At first, 10 nodes form a complete network and are divided into two communities with 5 nodes each. We assign vertices 4 and 9 as the core of each community, each of which has links with weight 2 connecting to vertices within its community and weight 0.2 connecting to outside vertices. All other intra-connections have weight 1, and all other inter-connections have weight 0.2. Then we introduce vertex 11 as the bridge between the two communities. It connects to all 10 nodes with weight 1. The index I and w-score for each node are given in Tab. 2. The results indicate that vertices 4, 9 and 11 are more important than the other vertices, while vertex 11 is a "bridge" between these two communities. Our method works quite well in this small artificial weighted network.

As an example of a real-world weighted network, we investigate the collaboration network among scientists working at the Santa Fe Institute (the SFI network). Here we consider it as a weighted, undirected network. Collaboration events between the scientists can be repeated again and again, and a higher frequency of collaboration usually indicates a closer relationship. Furthermore, weights can be assigned to the scientists' collaboration quite naturally: an article with n authors corresponds to a collaboration act of weight $\frac{1}{n-1}$ between every pair of its authors [40]. The results for the SFI collaboration network are sketched in Fig. 8. Vertex diameters indicate the community centrality I. The color of each vertex is proportional to the index w-score. Red vertices behave like "overlapping" nodes or "bridges" between communities, and yellow vertices often lie inside their own communities. We do not know the specific names; however, we observe that the positions of the large vertices are just like the "group leaders". Vertices 2, 12 and 24 are so-called "community cores" in communities because their w-scores are quite large. In fact, they are the group leaders in the fields of Mathematical Ecology, Statistical Physics and Structure of RNA, respectively. However, vertices 1, 9 and 11 are the "bridges" between communities, and they have relative small w-scores. Interestingly, the result in the weighted network is different from the one in the corresponding unweighted network. It can be concluded that the edge weight may affect the result. For example, vertex 9 and vertex 11 collaborate quite often; this makes both of them quite important in a weighted network, while in an unweighted network, neither of them is very important to the community structure.

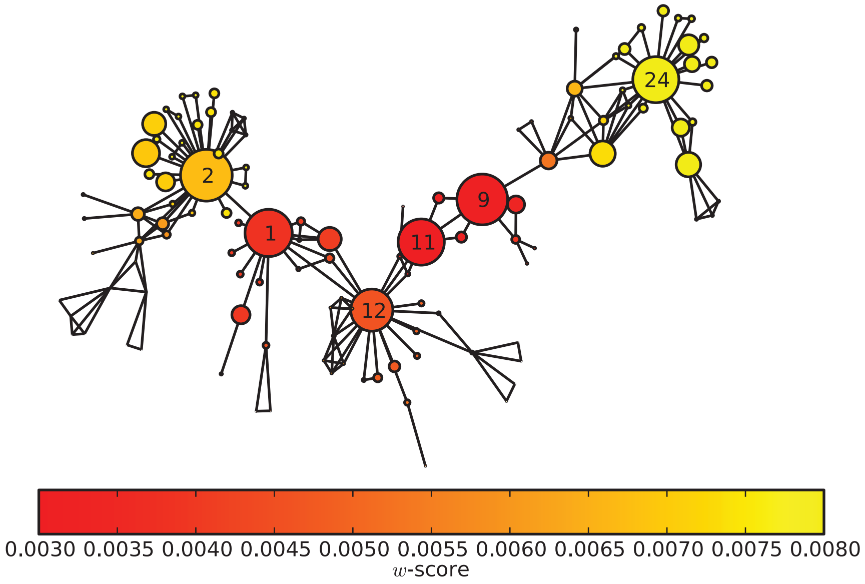

Figure 8. Sketch of the SFI scientific collaboration network as a weighted, undirected network. It has 118 scientists. Vertex diameters indicate the community centrality I. The color of each vertex is proportional to the index w-score.
doi:10.1371/journal.pone.0027418.g008

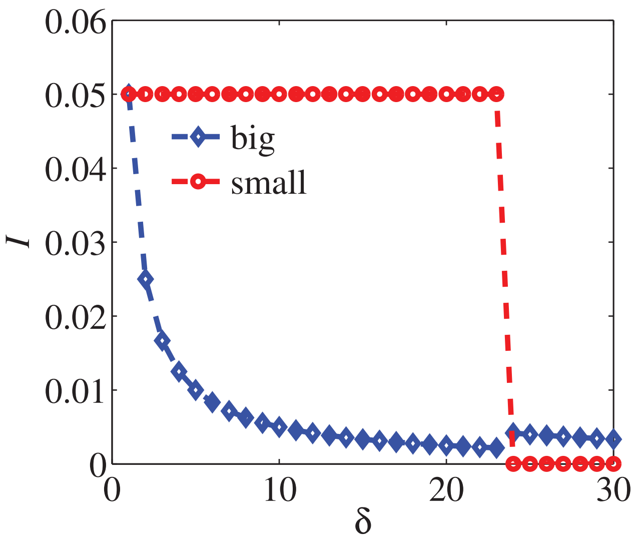

Figure 9. To test the limitation of our method. Considering a network composed with two communities but these two communities are not connected with each other and the size of the smaller one is always $N_{small}=10$. The figure shows the index I as a function of δ where the probability of connecting an edge between two nodes in each community $p=0.9$.
doi:10.1371/journal.pone.0027418.g009

Discussion

In this paper, we characterize the node importance to community structure using the spectrum of the graph. The eigenspectrum of the adjacency matrix gives a clear indication of the number of "dominant" communities in a network [27]. We give a centrality metric based on the spectrum of the adjacency matrix of the graph, and it can identify the nodes important to the community structure in many cases. In addition, we propose an index to distinguish the two kinds of important nodes that we term "community cores" and "bridges" using the spectrum of the graph Laplacian. We demonstrate a variety of applications of our method to both artificial and real-world networks representing social and neural networks. Our method works well in many cases without knowing the exact community structure, although the number of communities should be known.

If the network have very heterogeneous cluster sizes the limitation is likely to occur. There are two results for the limitation that are both related with the properties of the adjacency matrix. One is that we cannot find the real community structure when communities are very different in size. In Ref. [27], the authors have proved that if $N_{small}^2 < N_{large}$ where N is the size of the communities, the method cannot detect the small communities. The other problem is that when communities are very different in size, even we know the real communities by other methods, the index I may not show the real importance of the node in small communities because the index I is also based on the spectrum of the adjacency matrix. Considering a network composed with two isolated communities. The size of the smaller one is always 10 and we define $\delta = N_{large}/N_{small}$. Let each community be an ER random network with the probability of connecting $p=0.9$. The numerical result in Fig. 9 shows the similar limitation of the index I. It cannot identify the important nodes in the small communities when the communities are in very different size.

Our method can also be used in weighted networks. From our result in the SFI network, it can be inferred that edge weight may affect the result. Furthermore, it may generalize to directed networks because the Perron-Frobenius eigenvalues are often real and positive [41]. We have yet to treat the case of directed networks. The identification of such key nodes is important and could potentially be used to identify the organizer of the community in social networks, to develop an immunization strategy in an epidemic process, to identify key nodes in biological networks and so on. We hope our results may be helpful to future research.

Acknowledgments

The authors thank Di Huan, An Zeng, and Hongzhi You for help in the analysis of C. elegans neural network.

Author Contributions

Conceived and designed the experiments: YF ZD. Performed the experiments: YW. Analyzed the data: YW. Wrote the paper: YW ZD.

References

1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74: 47–97.
2. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45: 467–256.
3. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.
4. Watts DJ, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
5. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Physics Reports 424: 175–308.
6. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821–7826.
7. Lancichinetti A, Kivela M, Saramaki J, Fortunato S (2010) Characterizing the community structure of complex networks. PloS ONE 5: e11976.
8. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74: 036104.
9. Fortunato S (2010) Community detection in graphs. Physics Reports 486: 75–174.
10. Wu F, Huberman BA (2004) Finding communities in linear time: A physics approach. Eur Phys J B 38: 331–338.
11. Gfeller D, Ghappelier JC, De Los Rios P (2005) Finding instabilities in the community structure of complex networks. Phys Rev E 72: 056135.
12. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582.
13. Duch J, Arenas A (2005) Community detection in complex networks using extremal optimization. Phys Rev E 72: 027104.
14. Hu Y, Ding Y, Fan Y, Di Z (2010) How to measure significance of community structure in complex networks. ArXiv:1002.2007v1.
15. Hu Y, Nie Y, Yang H, Cheng J, Fan Y, et al. (2010) Measuring the significance of community structure in complex networks. Phys Rev E 82: 066106.
16. Karrer B, Levina E, Newman MEJ (2008) Robustness of community structure in networks. Rhys Rev E 77: 046119.
17. Lancichinetti A, Radicchi F, Ramasco JJ (2010) Statistical significance of communities in networks. Phys Rev E 81: 046110.
18. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100: 12123–12128.
19. Sun L, Li M, Jiang L, Tan L (2007) Comparative analysis of the gene co-regulatory network of normal and cancerous lung. Physica A 384: 739–746.
20. Liu Z, Hu B (2005) Epidemic spreading in community networks. Europhys Lett 72: 315.
21. Kovacs IA, Palotai R, Szalay MS, Csermely P (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS ONE 5: e12528.
22. Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433: 895–900.
23. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) Evidence for dynamically organized modularity in the yeast proteincprotein interaction network. Nature 430: 88–93.
24. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol 5: e154.

25. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2006) Stratus not altocumulus: A new view of the yeast protein interaction network. PLoS Biol 4: e317.
26. Zhao M, Zhou C, Chen Y, Hu B, Wang B (2010) Complexity versus modularity and heterogeneity in oscillatory networks: Combining segregation and integration in neural systems. Phys Rev E 82 82: 046225.
27. Chauhan S, Girvan M, Ott E (2009) Spectral properties of networks with community structure. Phys Rev E 80: 056114.
28. Luxburg UV (2007) A tutorial on spectral clustering. Statistics and Computing 17: 395–416.
29. Stam CJ (2005) Nonlinear dynamical analysis of eeg and meg: Review of an emerging field. Clin Neurophysiol 116: 2266–2301.
30. Fiedler M (1973) Algebraic connectivity of graphs. Czech Math J 23: 298–305.
31. Hagen L, Kahng A (1992) New spectral methods for ratio cut partitioning and clustering. IEEE Trans Computer-Aided Design 11: 1074–1085.
32. Newman MEJ (2010) Networks: An Introduction. Oxford UK: Oxford University Press.
33. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78: 046110.
34. Zachary WW (1977) An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33: 452–473.
35. Nelson DL, McEvoy CL, Schreiber TA (1998) The university of south florida word association, rhyme, and word fragment norms.
36. URL http://www-personal.umich.edu/mejn/netdata/.
37. White JG, Southgate E, Thomson JN, Brenner S (1986) The structure of the nervous system of the nematode caenorhabditis elegans. Philos Trans R Soc London B 314: 1–340.
38. Li D, Leyva I, Almendral JA, Sendi~na Nadal I, Buldú JM, et al. (2008) Synchronization interfaces and overlapping communities in complex networks. Phys Rev Lett 101: 168701.
39. Tsalik EL, Hobert OL (2003) Functional mapping of neurons that control locomotory behavior in caenorhabditis elegans. Neurobiol J 56: 178–197.
40. Ramasco JJ, Morris SA (2003) Social inertia in collaboration networks. Phys Rev E 73: 016122.
41. MacCluer CR (2000) The many proofs and applications of perron's theorem. SIAM Rev 42: 487–498.