Nodal domain partition and the number of communities in networks*

Bian He\textsuperscript{1}, Lei Gu\textsuperscript{2,3} and Xiao-Dong Zhang\textsuperscript{1,4}

\textsuperscript{1} Department of Mathematics, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, People’s Republic of China
\textsuperscript{2} Intel–NTU Connected Context Computing Center, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan
\textsuperscript{3} School of Mathematical Sciences, Fudan University, Shanghai 200433, People’s Republic of China
E-mail: hbtxwd@yahoo.com.cn, gehirn724@gmail.com and xiaodong@sjtu.edu.cn

Received 10 October 2011
Accepted 23 January 2012
Published 9 February 2012

Online at stacks.iop.org/JSTAT/2012/P02012
doi:10.1088/1742-5468/2012/02/P02012

Abstract. It is difficult to detect and evaluate the number of communities in complex networks, especially when the situation involves an ambiguous boundary between the inner- and inter-community densities. In this paper, discrete nodal domain theory is used to provide a criterion to determine how many communities a network has and how to partition these communities by means of topological structure and geometric characterization. By capturing the signs of the Laplacian eigenvectors, we separate the network into several reasonable clusters. The method leads to a fast and effective algorithm with application to a variety of real network data sets.

Keywords: analysis of algorithms, clustering techniques

* This work is supported by the National Natural Science Foundation of China (no. 10971137), the National Basic Research Program (973) of China (no. 2006CB805900), and a grant of Science and Technology Commission of Shanghai Municipality (STCSM, no. 09XD1402500).

4 Author to whom any correspondence should be addressed.
1. Introduction

With the rapid development of computer science, real-world networks have been widely studied, including the World-Wide Web [2, 22], metabolic networks [29, 23], epidemiology [35, 42], and scientific collaboration and citation networks [37, 44]. Consequently, plenty of models, such as the WS model [13], the BA model [1] and the random configuration model [10], have been proposed to depict the topological features and dynamic behaviors of networks. Recently, a particular and useful network structure, called ‘communities or clusters’, has been given considerable attention. A common description of communities is division of nodes into groups with dense connection inside and sparse connection outside. Nodes in the same community share similar functional patterns in the dynamic of the whole network.

Practically, some researchers consider individual behaviors of nodes or edges that affect the surroundings [24, 19], while some others focus on the dynamics of ensembles of all the nodes in networks [3, 14]. These emphases clarify two possible directions in community detection: (1) detail exploitation—identifying nodes or edges whose absence has most influence on network dynamics, which mostly are the boundary parts; (2) global sights—testing for partition structures either close enough to the original network or distinct from the corresponding random pattern network.

It has been long accepted that centrality measures [45, 20, 21, 30] are used to depict the relative importance of the nodes in a network [4, 45, 25]. In 2002, a fast algorithm aimed at identifying each edge of a network as a betweenness measure gave rise to an explosive growth of activities in this field. Girvan and Newman [27] used the scale to quantify the roles of edges in information transmission following paths of minimal length.

Nodal domain partition and the number of communities in networks

Contents

1. Introduction 2
2. Spectral partition and nodal domain theory 3
   2.1. Spectral partition .................................. 3
   2.2. Nodal domains by eigenvectors ..................... 5
3. Partitioning by the weak nodal domain 7
4. Experiments with WNDP 9
   4.1. Dolphin graph .................................. 9
   4.2. The political book graph .......................... 10
   4.3. Capocci graph .................................. 10
   4.4. Computer-generated graphs .......................... 13
5. Conclusion 13
Acknowledgments 15
References 16

1. Introduction

With the rapid development of computer science, real-world networks have been widely studied, including the World-Wide Web [2, 22], metabolic networks [29, 23], epidemiology [35, 42], and scientific collaboration and citation networks [37, 44]. Consequently, plenty of models, such as the WS model [13], the BA model [1] and the random configuration model [10], have been proposed to depict the topological features and dynamic behaviors of networks. Recently, a particular and useful network structure, called ‘communities or clusters’, has been given considerable attention. A common description of communities is division of nodes into groups with dense connection inside and sparse connection outside. Nodes in the same community share similar functional patterns in the dynamic of the whole network.

Practically, some researchers consider individual behaviors of nodes or edges that affect the surroundings [24, 19], while some others focus on the dynamics of ensembles of all the nodes in networks [3, 14]. These emphases clarify two possible directions in community detection: (1) detail exploitation—identifying nodes or edges whose absence has most influence on network dynamics, which mostly are the boundary parts; (2) global sights—testing for partition structures either close enough to the original network or distinct from the corresponding random pattern network.

It has been long accepted that centrality measures [45, 20, 21, 30] are used to depict the relative importance of the nodes in a network [4, 45, 25]. In 2002, a fast algorithm aimed at identifying each edge of a network as a betweenness measure gave rise to an explosive growth of activities in this field. Girvan and Newman [27] used the scale to quantify the roles of edges in information transmission following paths of minimal length.
across a network. The removal of edges with high *betweenness* leads to exposure of the community structure.

Besides these ‘detail exploitation’ methods, researchers also make use of quality functions defined on whole network clustering. One of these most popular quality functions is *modularity* [38, 40, 41]. The basic thought was to compare the difference between the total actual fractions of edges inside groups and the expected fractions when edges were placed at random. An improved version [40] was also developed to measure the difference between the actual network and the *null model* which yielded networks that were not supposed to have natural community structures [36, 40].

Furthermore, a probabilistic framework embedded with ‘stochastic matrices’ provides another important quality function [16, 17]. By introducing a metric on the space of Markov chains $K$ which represent random walks on the network [17], simple stochastic structure is finally detected as the best approximation to the dynamics behavior of $K$. This simple structure contains the community information of the original network.

Despite the complexity of the NP-hard algorithm [33], improved approximation techniques calculate ‘optimized’ partition in partition problems [39, 40, 28, 16]. However, the question about the number of communities is still not easily answered [40, 17].

In this paper we introduce a method applying weak-nodal-domain partition (WNDP) to give the number of clusters by exploring the information in the graph Laplacian eigenvectors. In section 2, we begin with a brief review of the spectral partition method and classical nodal domain theorem. In section 3, the main method and algorithm will be presented. In section 4, the algorithm is applied to three real networks with fast and efficient results. In section 5, an exceptional case is demonstrated for the further understanding of the method.

### 2. Spectral partition and nodal domain theory

An exact mathematical definition of a *cluster* has not been made explicit so far, but common agreement is focused on the minimization of edges whose disappearance will separate the network into groups with no inter-connection. Our purpose is to establish a mathematical framework to grasp the essential properties of the network not only precisely but also effectively. In particular, quite a number of the clustering methods are involved with special matrices, for example an adjacency matrix, a graph Laplacian etc. All these matrices share the topological information of the network ostensibly or inconspicuously.

So far, surprising results in clustering have been gained by using eigenvectors of special matrices [11, 34]. Scientists are interested in the projection from properties of these matrices to the corresponding networks. In the following sections, the eigenvectors of a graph Laplacian $L$ will serve as a tool to detect and analyze the community structure of networks.

#### 2.1. Spectral partition

Let $G = (V, E)$ be an undirected graph with a labeled node set $V = \{1, 2, \ldots, n\}$. As an unweighted graph, the *adjacency matrix* $A = (A_{ij})$ is defined to be

$$A_{ij} = \begin{cases} 1, & \text{if } i \sim j, \\ 0, & \text{otherwise.} \end{cases}$$
Meanwhile, the graph Laplacian matrix is defined to be $\mathcal{L} = D - A$. Here, $D = \text{Diag}(d_1, \ldots, d_n)$ is the diagonal degree matrix where $d_i = \sum_{j=1}^{n} A_{ij}$. Another important concept is the cut size:

$$\text{Cut} = \frac{1}{2} \sum_{i,j \text{ in different groups}} A_{ij}. \quad (1)$$

Noting that the factor $\frac{1}{2}$ is a compensation for the double count as $A_{ij} = A_{ji}$. It is the number of edges connecting different communities. A traditional way is to minimize the cut size under all the possible partition choices.

Given a partition of $V$ into $k$ sets $A_1, A_2, \ldots, A_k$, we rewrite equation (1) in a universal form:

$$\text{Cut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_{l=1}^{k} \text{Cut}_l, \quad (2)$$

where $\text{Cut}_l = \sum_{i \in A_l, j \in \overline{A}_l} A_{ij}$ and $\overline{A}_l$ is the complement of $A_l$ in $V$.

We then set an $n \times k$ matrix $S = (S_{ij})$ to indicate the positions of nodes in communities:

$$S_{ij} = \begin{cases} 1, & \text{if } i \in A_j, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Note that the columns $S_j = (S_{ij})^T$ of $S$ are mutually orthogonal, and the matrix satisfies normalization $\text{Tr}(S^T S) = n$. Thus,

$$\text{Cut}_l = \sum_{i \in A_l, j \in \overline{A}_l} A_{ij} = \sum_{i \in A_l} \sum_{j \in \overline{A}_l} A_{ij} - \sum_{i \in A_l} \sum_{j \in A_l} A_{ij}. \quad (4)$$

Now put equation (4) into a matrix form

$$\text{Cut}_l = \sum_i D_i S_{il} S_{il} - \sum_{ij} A_{ij} S_{il} S_{jl}$$

$$= S_l^T \mathcal{L} S_l.$$

Hence,

$$\text{Cut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_l S_l^T \mathcal{L} S_l$$

$$= \frac{1}{2} \text{Tr}(S^T \mathcal{L} S). \quad (5)$$

So the problem of minimizing $\text{Cut}(A_1, \ldots, A_k)$ is rewritten as

$$\frac{1}{2} \min_{A_1, \ldots, A_k, \text{Tr}(S^T S) = n} \text{Tr}(S^T \mathcal{L} S) \quad \text{for all } S \text{ in equation (3).} \quad (6)$$

Note that the optimization problem is based on the matrix $S$ whose entries are only allowed to take special values 0 or 1, which indicates a complicated calculation to solve the issue. In order to find a possible solution, we make a relaxation on the discrete condition and allow $S_{ij}$ to be arbitrary values in $\mathbb{R}$. Hence, a general form is

$$\min_{A_1, \ldots, A_k, S^T S = I_{k \times k}} \text{Tr}(S^T \mathcal{L} S). \quad (7)$$

doi:10.1088/1742-5468/2012/02/P02012
This is the standard form of a trace minimization problem, and the Rayleigh–Ritz theorem tells us the best minimization can be achieved by \( S \) containing the first \( k \) eigenvectors of \( \mathcal{L} \) as its columns. Let us make use of the Laplacian eigenvectors to rewrite equation (7).

Since \( \mathcal{L} \) is a positive semi-defined symmetric matrix, its eigenvalues are all real and non-negative. Thus, respectively, let \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \cdots \leq \lambda_n \) be defined as the eigenvalues of \( \mathcal{L} \), as well as their corresponding normalized eigenvectors \( f_1, \ldots, f_n \). For each row of the Laplacian matrix \( \mathcal{L} \), we have

\[
\sum_{j=1}^{n} L_{ij} = D_{ii} - \sum_{j=1}^{n} A_{ij} = 0.
\]

This implies that \((1, \ldots, 1)^T\) is the eigenvector of \( \mathcal{L} \) corresponding to eigenvalue 0, so \( 0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \cdots \leq \lambda_n \).

Thus, \( \mathcal{L} = F D F^T \), where \( F = (f_1 | \cdots | f_n) \) is the eigenvector matrix and \( D \) is a diagonal matrix with \( D_{ii} = \lambda_i \).

\[
\text{Tr}(S^T \mathcal{L} S) = \sum_{l=1}^{k} S_l^T \mathcal{L} S_l = \sum_{j=2}^{n} \lambda_j \sum_{l=1}^{k} (f_j^T S_l)^2.
\]

By setting \( w_j = \sum_{l=1}^{k} (f_j^T S_l)^2 \), we have

\[
\text{Tr}(S^T \mathcal{L} S) = \sum_{j=2}^{n} \lambda_j w_j.
\]

Note that the optimization has been split into pieces according to the eigenvalues \( \{\lambda_j\} \) as shown in equation (8) with the column-based operation on the eigenvector matrix \( F \). Thus, our course is clear: to find a separation placing as much as possible of the weight \( w_j \) on the side of small eigenvalues with as little as possible on the large ones. In terms of equation (9), the properties of eigenvectors have a great influence in the processing, which leads to a further exploration of the eigenspaces to help approximate the optimization.

2.2. Nodal domains by eigenvectors

Recall the basic characterization of the eigenvalues and eigenvectors

\( \mathcal{L} f_j = \lambda_j f_j \).

By multiplying both sides by vector \( f_j^T \), we have

\( f_j^T \mathcal{L} f_j = \lambda_j f_j^T f_j = \lambda_j \),

where \( f_j \) is normalized and \( \lambda_j \) is the eigenvalue in terms of the Rayleigh quotient of \( \mathcal{L} \). Actually, for \( j \in \{2, \ldots, n\} \), eigenvalues can be characterized as

\[
\lambda_j = \inf_{f \perp W_{j-1}} f^T \mathcal{L} f,
\]

\text{doi:10.1088/1742-5468/2012/02/P02012}
where $W_{j-1}$ is the subspace spanned by the eigenvectors of the smallest $j-1$ eigenvalue [9]. Simple calculation shows that for arbitrary $f \in \mathbb{R}^n$

$$f^T \mathcal{L} f = \sum_{u \sim v} (f(u) - f(v))^2,$$

which can be substituted into equation (10) to give

$$\lambda_j = \inf_{f \perp W_{j-1}} \sum_{u \sim v} (f(u) - f(v))^2,$$

where $\lambda_j$ is achieved with $f$ being the exact $j$th corresponding eigenvector.

It is easy to see that under the condition $f \perp W_{j-1}$, the eigenvector $f_j$ is the weight function that minimizes the total weight difference between pairs of adjacent nodes through the whole network. Such a property indicates a structure in which nodes are more likely to appear in the same group if their corresponding elements in $f_j$ are close [33]. However, these values are widely distributed such that there is no such criterion to depict the boundary between groups.

Let us look back at equation (9). Matrix $S$, which represents the partition, should be chosen to make $w_j$ relatively large while $j$ is small. For a single $f_j$, the corresponding $w_j$ reaches a maximum when nodes are grouped as follows:

$$x \in S_1, \quad \text{if } f_j(x) > 0,$$

$$x \in S_2, \quad \text{if } f_j(x) < 0.$$  

(12)

Note that the set $\{x | f_j(x) = 0\}$ is not mentioned as it makes no contribution to the maximization of $w_j$. However, the network structure is complicated, so nodes in the same group $S_1$ or $S_2$ might not be connected at all (see figure 1 for instance).

Hence, instead of clustering nodes as in equations (12), we add a natural constraint that each group should be a connected subgraph referring to the graph’s geometric structure. The square operation in equation (8) suggests that two operations on subgraphs are preferred: (a) nodes within the same subgroup share the same sign in $f_j$; (b) each subgroup contains as many nodes as possible. This consideration leads us to focus on partitions deduced according to the signs of each component of the eigenvectors.

An interesting result, named the Courant–Hilbert nodal theorem [7], gives a detailed description of the domains cut by zeros of each eigenfunction.

Given the self-adjoint second order differential equation $L[u] + \lambda \rho u = 0 \ (\rho > 0)$ for a domain $G$ with arbitrary homogeneous boundary conditions, if its eigenfunctions
are ordered according to increasing eigenvalues, then the zeros of the \( n \)th eigenfunction \( u_n \) divide the domain into no more than \( n \) subdomains. No assumptions are made about the number of independent variables (from Courant and Hilbert Methods in Mathematical Physics).

This inspires us to take into account the following discrete nodal domain theorem on graph \( G = (V, E) \). A positive (negative) strong nodal domain of a function \( f \) on \( V(G) \) is a maximal connected induced subgraph of \( G \) on nodes \( v \in V \) with \( f(v) > 0 \) \((f(v) < 0)\). Meanwhile, a positive (negative) weak nodal domain of a function \( f \) on \( V(G) \) is a maximal connected induced subgraph with \( f(v) \geq 0 \) \((f(v) \leq 0)\) that contains at least one node with nonzero value. The relationship between eigenvectors and nodal domains was introduced and studied by Davies, Gladwell, Leydold, Stadler and others [6,12]. One of those important results is the following:

**Theorem 1.** [6] Let \( M \) be a symmetric matrix with non-negative diagonal entries and \( M_{uv} < 0 \) as \( u \sim v \). If \( \lambda_1 \leq \cdots < \lambda_k = \lambda_{k+1} \cdots = \lambda_{k+r-1} < \lambda_{k+r} \leq \cdots \leq \lambda_n \) are eigenvalues of \( M \) and \( f_k \) is the corresponding eigenvector of \( \lambda_k \), then the number of strong nodal domains of \( f_k \) is no more than \( k + r - 1 \) and the number of weak nodal domains of \( f_k \) is no more than \( k \).

This theorem defines the matrix \( M \) in a general expression, in which the Laplacian of the graph is included. It shows a natural structure framework in which we achieve a maximization of \( w_j \) satisfying the connectivity information. An interesting viewpoint occurs when there are zero components in the eigenvectors. Nodes representing zero elements are shared between communities defined in the weak nodal domains and communities in this overlapping structure are the basic functional parts in the dynamic system of the network.

### 3. Partitioning by the weak nodal domain

The eigenvector affording the second smallest eigenvalue of the Laplacian of the network is called the Fiedler vector [18]. It has been applied in graph bi-partitioning [43] and spectral clustering [5]. Generally speaking, these methods provide partitions that attract rather a large weight on the smallest nonzero eigenvalue in equation (9) to make \( \text{Cut}(A_1, \ldots, A_k) \) small. However, bad results were found for many graphs in [26]. Taking the ‘cockroach graph’ [26], for example, the second eigenvector offers a division cutting horizontally through the ladder while, obviously, the ideal cut is the long dotted line (see figure 2(a)).
The discrete nodal domain theorem indicates that the Fiedler vector decides weak
nodal domains, no more than two, but for the cockroach graph the ideal cut by visual
observation separates the graph into three parts. Naturally, we are interested in the
behavior of the third eigenvector. Calculation shows that the third eigenvector provides
the exact same separation as the ideal cut does (see figure 2(b)).

This result leads us back to equation (9). Despite the weight expression \( w_j = \sum_{l=1}^{k} (f_j^T S_l)^2 \), eigenvalues also play an important role in the minimization of the formula. Actually the cockroach graph possesses two eigenvalues \( \lambda_2 = 0.0057 \) and \( \lambda_3 = 0.0062 \) that are relatively close, compared with \( \lambda_4 = 0.0246 \) and other eigenvalues, which implies that heavy weight on \( \lambda_3 \) is also a possible optimized choice. Hence, we can apply eigenvectors
other than the Fiedler vector to the partitioning, and this is the so-called WNDP.

In most situations eigenvectors corresponding to smaller eigenvalues provide the
WNDP with a smaller cut size, which makes multi-community structure less easy to
determine. To avoid this difficulty, we take the famous quality function modularity \( Q \) [40] as a criterion:

\[
Q = \text{Tr}(S^T (A - P) S),
\]

where \( P = (P_{ij}) \) is the random configuration correspondence of \( A \) satisfying \( P_{ij} = d_id_j/2|E| \) [40].

This quantity function depicts the difference between how many edges are actually
within communities and how many edges are to be expected within communities. Interestingly, we have

\[
Q = \text{Tr}(S^T ((D - P) - (D - A)) S) = \text{Tr}(S^T (D - P) S) - \text{Tr}(S^T (D - A) S),
\]

where the first half represents the expected edge number outside communities and the
other half is the actual edge number of such a set. Note that the second half is exactly
the cut size of the network mentioned before. We are interested in the difference between
WNDP and the exact optimization result of modularity \( Q \).

Considering a WNDP, \( Q \) will be improved in three ways:

(i) two communities merge together: volume to be \( V_i = \sum_{h \in A_i} d_h \) and \( V_j = \sum_{h \in A_j} d_h \),
number of connecting edges to be \( c_{ij} \); this alteration only happens when

\[
\frac{V_i \cdot V_j}{2|E|} < c_{ij};
\]

(ii) one community splits into two: with the same definition as above; this alteration only
happens when

\[
\frac{V_i \cdot V_j}{2|E|} > c_{ij};
\]

(iii) move a single node from one community to another: a node \( v \) with \( n_i \) connections to
\( A_i \) and \( n_j \) connections to \( A_j \), it belongs to \( A_i \) and gives a higher \( Q \) when

\[
\frac{V_i \cdot n_i}{2|E|} - \frac{V_j \cdot n_j}{2|E|} < n_i - n_j.
\]

\[\text{doi:10.1088/1742-5468/2012/02/P02012}\]
The first two operations will not usually be needed if the community structure is relatively clear (see examples below), thus slight alterations on nodes lead to a good approximation.

This algorithm is practically processing in the form of a matrix calculation. Specifically, given the function $f$, the complexity of finding the weak nodal domains is $O(|V|)$.

Remember that there could be zero components in the eigenvectors. Define a node $v$ to be a zero node if $f(v) = 0$. Similarly a zero component is a maximal connected subgraph of zero nodes. We preprocess the graph in following steps:

(i) Contract all zero components into single nodes which inherit the connections of the components. Here, multiple edges are allowed. This leads to graph $G_1$.

(ii) Split each zero node in $G_1$, say $v$, into two connected individual nodes $v^+$ and $v^-$, in which $v^+$ inherits all neighbors of $v$ with positive values in $f$ and $v^-$ inherits the corresponding negative parts. (Note that either all neighbors of $v$ are zero nodes or $v$ has all neighbors of different signs.) This leads to graph $G_2$.

We call the new graph $G_2$ after the two steps of a weak domain graph. It is the graph our algorithm will be applied to. Note that the most complex part of the method is to calculate the eigenvectors of the sparse Laplacian matrix, which is known to be polynomial of order $O(n^2)$, and one can apply the shifted power method to get the $k$th eigenvector which is of complexity $O(n/\log(\lambda_k - \lambda_{k-1}))$.

The situation with zero components could be complicated, for that modularity is defined on a structure with no overlapping. In this paper, a limitation of the definition of zero value to be of order $O(10^{-17})$ will avoid overlapping structures. The ambiguous boundaries will be discussed in future work.

4. Experiments with WNDP

The whole comparison framework defined above allows us to test the best partition as well as the number of clusters. Or more precisely, we will find out what information inside the eigenvectors will be used to get a partition. Here, three different networks are used with our algorithm to test the WNDP method.

4.1. Dolphin graph

The study of the dolphin graph is based on a certain group of bottlenose dolphins living in Doubtful Sound, New Zealand. Lusseau [32] established this representative animal social network from 7 years of field studies of the dolphins, in which two dolphins are tied if there is an observation of statistically significant frequent association between them (see figure 3(b) for the detailed connections). Lusseau discovered an interesting phenomenon after years of observation: the whole group of dolphins split into two small subgroups following the departure of one key member named ‘SN89’. This observed division is represented with colors in the figure.

We process the graph with our algorithm, and the outcome indicates $f_2$ possessing reasonable WNDP (see figure 3(a)). The solid line in figure 3 depicts the division of the calculation result. Only ‘SN89’ does not fit the outcome. The background of this division...
4.2. The political book graph

The political book graph was assembled and studied by Krebs [31]. The database was taken from the online bookseller Amazon.com, in which 105 books were considered. Edges between books represent frequent co-purchasing by the same buyers. Krebs collected this information to study the relation between books describing different political views in 2004 around the US election. Naturally, this special graph inherits bilateral structure arising from the two main political parties in the United States, Democratic and Republican, respectively.

It is not surprising that WNDP by $f_2$ of the political book graph reaches its optimum modularity. Figure 4(b) is the exact partition, in which each side has its particular groups of authors and readers. An intuitive survey of the original graph shows our prediction still has two liberal books of presented in the conservative group.

4.3. Capocci graph

The Capocci graph is a simple graph (as shown in figure 6 up-left) generated by Capocci for the application of eigenvectors to identify communities [8]. The experiment shows that the second eigenvector of the right stochastic matrix, which is $D^{-1}A$, indicates three plateaus corresponding to the three evident components of the graph.

Figure 5 depicts the numerical result of our algorithm that the largest modularity is offered by $f_3$. Detailed partitioning is represented in figure 6(a), in which we also demonstrate respectively the WNDPs of eigenvectors corresponding to the first three
Figure 4. The political books graph [31]. (a) Modularities of WNDPs by different eigenvectors of Laplacians on the political books graph. WNDP by $f_2$ is selected by the algorithm. (b) Books with a dark color are liberal, the gray ones are centrist or unaligned, and the rests without colors are conservative. The gray books act as buffers between the ones with left-wing and right-wing points of view. A solid line divides the graph following WNDP of $f_2$. Only two confused nodes representing books purchased frequently by both sides are positioned incorrectly.
Figure 5. Modularities of WNDPs by different eigenvectors of Laplacians on the Capocci graph, among which $f_3$ suggests the best choice.

Figure 6. The Capocci graph [8]. (a) Top left: the original graph. Top right: partition by $f_2$. Bottom left: partition by $f_3$. Bottom right: partition by $f_4$. Generally speaking, eigenvectors other than these four above would possess more complicated nodal domain structures. (b) The eigenvalues of the Laplacian on the political book graph. Note that the first three eigenvalues are rather smaller than the rest.

nonzero eigenvalues. Obviously, WNDP on the bottom-left by the chosen $f_3$ gives a partition matching our visual observation.

Compared with the dolphin graph and the political book graph, the best WNDP of the Capocci graph is the one according to $f_3$ rather than $f_2$. This small alteration reminds us of the important roles that eigenvalues $\{\lambda_j\}$ have played in equation (9). To illustrate this idea, we separate small eigenvalues with relatively large ones

$$\sum_{j=2}^{n} \lambda_j w_j = \sum_{j=2}^{s} \lambda_j w_j + \sum_{j=s+1}^{n} \lambda_j w_j,$$

where $\{\lambda_i\}_{i=1}^{s}$ represent relatively small eigenvalues while $\{\lambda_i\}_{i=s+1}^{n}$ are the large ones. Our
experiments show that eigenvectors corresponding to \( \{ \lambda_i \}_{i=2}^s \) hold some useful information about the community structure of the graph.

4.4. Computer-generated graphs

Taking a rough look at the community structure, we contract each cluster of the WNDP into one single node inheriting connections of the cluster. Despite multiple edges, no circles exist in this simple structure which, in brief, is a tree. This interesting phenomenon comes from the fact that only two signs ‘+’ and ‘−’ are used to identify different nodes. We generate a graph artificially following the method used by Girvan and Newman [27], also called the ad hoc network, in which a whole graph of 128 nodes is divided into four communities of 32 nodes each. More precisely, in our special case, edges between pairs of nodes in the same community are placed with a possibility of 0.4 while ones in different communities share a possibility of 0.1. The randomness of the edges indicates multiple choices; however, we are only interested in the known community structure of these graphs. The graph in figure 7 is one of these special examples following the Girvan and Newman method on which our algorithm is applied. As each community is connected to all the other three communities, four signs are required to separate these four communities, which suggests our method is incomplete.

Calculation shows that WNDPs by eigenvectors according to the first three nonzero eigenvalues divide the graph into two parts (see figure 7). In other words, each eigenvector reveals part of the community structure. By combining this partial information we may recover knowledge of the whole structure. This leads us to a generalized definition of the nodal domains: a strong nodal domain of functions \( \{ f_1, \ldots, f_i \} \) on \( V(G) \) is a maximal connected induced subgraph of \( G \) on nodes \( v \in V \) which have corresponding vectors \( (f_1(v), \ldots, f_i(v)) \) belonging to the same quadrant of the \( i \)-dimensional Euclidean space. Note that \( i \) is the exact number of eigenvectors \( \{ f_1, \ldots, f_i \} \) which we process for the combined information. For the definition of weak nodal domain, the maximal connected induced networks are the ones whose corresponding vectors \( (f_1(v), \ldots, f_i(v)) \) belong to the same quadrant and boundary.

A convincing result emerges by applying our algorithm on these generalized WNDPs (see figure 7). A single eigenvector does separate the whole network into reasonable parts, but not subtly enough. Structure functioning on a relatively small scale will be demonstrated by an appropriate blend of certain eigenvectors.

5. Conclusion

The experiments above confirm that the nodal domains of the Laplacian eigenvectors contain useful information about the community structure of the graph. Still, exceptions arise when we have relatively large difference between volumes of different communities. We set up a graph with the following three steps (see figure 8):

- Use the ER model [15] to build a graph, named \( G_1 \), of 200 nodes whose average degree is 40 (nodes shaped in solid squares).
- Use the ER model to build another graph, named \( G_2 \), of 20 nodes whose average degree is 4 (nodes shaped in solid circles).
- Build an extra single node; first connect it with a random node in \( G_2 \), then connect it with five random nodes in \( G_1 \) (nodes shaped in triangles).
Calculation suggests that WNDP by the second eigenvector possesses the largest modularity. The solid line in figure 8 shows the exact division. But direct observation tells us the triangle has more connection with $G_1$ than that with $G_2$.

A survey of the algorithm has told us that the process of calculating the eigenvectors is equal to evaluating each point with an average of its neighbor’s eigenvector value, except under a certain rate which is the corresponding eigenvalue. Thus, the eigenvector $f_2$ of the Laplacian on the graph would evaluate nodes in $G_1$ with positive values and negative ones for $G_2$. Because $G_2$ has much fewer nodes than $G_1$, $f_2$ on $G_2$ possesses relatively small negative values, which indicates that on average the triangle possesses a negative
value as nodes in $G_2$ do. Further work should be focused on the positional adjustment of the boundary nodes like the triangle.

In section 3, we already find a solution for this situation by checking the boundary nodes for improvement on modularity $Q$. But recalling the two operations (13) and (14) in modifying $Q$, we note that if the volume of the two communities is small or unbalanced (with a wide gap), these two should not be separated for the inequalities (13) and (14) are not satisfied. That is why the modularity has a well-known resolution limit, that makes clusters smaller than a given size undetectable. A resolution coefficient $\lambda$ can make up for this limitation as the inequality goes as

$$\frac{V_i \cdot V_j}{2\lambda|E|} < c_{ij}, \quad \lambda \in (0, 1]$$

where the nodes are localized by considering the influence of their ranged neighbors.

To conclude, in this paper we introduce a method that applies weak nodal domains according to eigenvectors of the Laplacian of the graph. We calculate the modularity [39, 40] to decide which WNDP behaves best, that is, suggests the optimal number of communities in networks. We also test our algorithm in real-world models. As the examples show, for arbitrary graphs, WNDP works quite well in deciding the number of clusters in graphs. Future work will be focused on the mechanism of the WNDP algorithm, as well as finding other good parameters to choose the best eigenvectors.

**Acknowledgments**

The authors wish to thank the referees for their valuable comments and suggestions.
References

[1] Barabási A L and Albert R, 1999 Science 286 509
[2] Albert R, Jeong H and Barabási A L, 1999 Nature 401 130
[3] de Menezes M A and Barabási A L, 2004 Phys. Rev. Lett. 93 068701
[4] Bonacich P, 1987 Am. J. Sociol. 92 1170
[5] Belkin M and Niyogi P, 2002 Advances in Neural Information Processing Systems 14 (NIPS 2001) (Cambridge, MA: MIT Press) pp 585–91
[6] Biyikoglu T, Leydold J and Stadler P F, 2007 Springer Lecture Notes in Mathematics (Berlin: Springer)
[7] Courant R and Hilbert D, 1989 Methods of Mathematical Physics vol 1 (New York: Wiley)
[8] Capocci A, Servedio V D P, Caldarelli G and Colaiori F, 2004 Proc. 3rd Workshop on Algorithms and Models for the Web Graph (Springer Lecture Notes in Computer Science vol 3243) ed S Leonardi (Berlin: Springer)
[9] Chung F, 1997 Spectral Graph Theory (Providence, RI: American Mathematical Society)
[10] Chung F and Lu L, 2002 Ann. Comb. 6 125
[11] Cvetković D M, Rowlinson P and Simić S, 1997 Encyclopedia of Mathematics and its Applications vol 66 (Cambridge: Cambridge University Press)
[12] Dekel Y, Lee J R and Linial N, 2007 Proc. APPROX-RANDOM’2007 pp 436–48
[13] Watts D J and Strogatz S H, 1998 Nature 393 440
[14] de Menezes M A and Barabási A L, 2004 Phys. Rev. E 73 066132
[15] Erdös P and Rényi A, 1959 Publ. Math. (Debrecen) 6 290
[16] Li T, Liu J and E W, 2009 Phys. Rev. E 80 026106
[17] E W, Li T and Vanden-Eijnden E, 2008 Proc. Nat. Acad. Sci. 105 7907
[18] Fiedler M, 1973 Czech. Math. J. 23 298
[19] Freeman L C, 1977 Sociometry 40 1
[20] Freeman L C, 1979 Soc. Netw. 1 215
[21] Fortunato S, Latora V and Marchiori M, 2004 Phys. Rev. E 70 056104
[22] Faloutsos M, Faloutsos P and Faloutsos C, 1999 Comput. Commun. Rev. 29 251
[23] Fell D A and Wagner A, 2000 Nature Biotechnol. 18 1121
[24] Guimerà R and Amaral L A N, 2005 Nature 433 895
[25] Granovetter M, 1973 Am. J. Social. 78 1360
[26] Guattery S and Miller G, 1998 SIAM J. Matrix Anal. Appl. 19 701
[27] Girvan M and Newman M E J, 2002 Proc. Nat. Acad. Sci. 99 7821
[28] Jiang J Q, Dress A W M and Yanga G, 2009 Appl. Math. Lett. 22 1479
[29] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A L, 2000 Nature 407 651
[30] Stephenson K A and Zelen M, 1989 Soc. Netw. 11 1
[31] Rebs V, 2004 unpublished, see www.orgnet.com/
[32] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M, 2003 Behav. Ecol. Sociobiol. 54 396
[33] von Luxburg U, 2007 Stat. Comput. 17 395
[34] Merri R, 1998 Linear Algebra Appl. 278 221
[35] Moore C and Newman M E J, 2000 Phys. Rev. E 61 5678
[36] Molloy M and Reed B, 1995 Random Struct. Algorithms 6 161
[37] Newman M E J, 2001 Proc. Nat. Acad. Sci. 98 404
[38] Newman M E J and Girvan M, 2004 Phys. Rev. E 69 026113
[39] Newman M E J, 2004 Eur. Phys. J. B 38 321
[40] Newman M E J, 2006 Phys. Rev. E 74 036104
[41] Newman M E J, 2006 Proc. Nat. Acad. Sci. 103 8577
[42] Pastor-Satorras R and Vespignani A, 2001 Phys. Rev. Lett. 86 3200
[43] Pothen A, Simon H and Liou K P, 1990 SIAM J. Matrix Anal. Appl. 11 430
[44] Redner S, 1998 Eur. Phys. J. B 4 131
[45] Sabidussi G, 1966 Psychometrika 31 581

doi:10.1088/1742-5468/2012/02/P02012