Deep Community Detection
Pin-Yu Chen and Alfred O. Hero III, Fellow, IEEE

Abstract—A deep community in a graph is a connected component that can only be seen after removal of nodes or edges from the rest of the graph. This paper addresses the problem of detecting deep communities using a new centrality measure, called the local Fiedler vector centrality (LFVC). The LFVC is associated with the sensitivity of algebraic connectivity to node or edge removals. We prove that a greedy node/edge removal strategy, based on successive minimization of LFVC, is nearly as good as the optimal, but intractable, combinatorial batch removal strategy. Under a stochastic block model framework, we show that the greedy LFVC strategy can extract deep communities with probability one as the number of observations becomes large. We apply the greedy LFVC strategy to real-world social network datasets. Compared with conventional community detection methods we demonstrate improved ability to identify important communities and key members in the network.

Index Terms—local Fiedler vector centrality, random graphs, spectral graph theory, social networks, submodularity

I. INTRODUCTION

In social, biological and technological networks [1], [2], community detection aims to extract tightly connected subgraphs in networks. Recently, community detection has attracted a great deal of interest in network science [3], [4]. Community detection is often cast as graph partitioning. Many graph partitioning methods exist in the literature, including graph cuts [5], [6], probabilistic models [7], [8], and node/edge pruning strategies based on different criteria [11]. [9]. [10]. Many community detection methods are based on detecting noisy parts or edges with high centrality. Node and edge centralities are quantitative measures that are used to evaluate the level of importance and/or influence of a node or an edge in a network. Centralities can be based on combinatorial measures such as shortest paths or graph diffusion distances between every node pair [12], [13]. Centrality measures can also be based on spectral properties of the adjacency and graph Laplacian matrices associated with the graph [13]. Many of these measures require global topological information and therefore may not be computationally feasible for very large networks.

Nonparametric community detection methods, such as the edge betweenness method [11] and the modularity method [9], can be viewed as edge removal strategies that aim to maximize a centrality measure, e.g., modularity or betweenness measure. It is worth noting that these methods presume that each node in the graph is affiliated with a community. However, in some applications it is quite often that the graphs contain edges connecting to irrelevant “noisy” nodes that are not members of any community. In such cases, noisy nodes and edges mask the true communities in the graph. Detection of these masked communities is a difficult problem that we called “deep community detection”. The formal definition of a deep community is given in Sec. III. Due to the presence of noisy nodes and edges, deep communities elude detection when standard partitioning methods are applied. In this paper, a new partitioning strategy is applied to detect deep communities.

The new partitioning strategy is based on a novel spectral measure [14], [15] of centrality for discovering deep communities, called local Fiedler vector centrality (LFVC). LFVC is associated with the sensitivity of algebraic connectivity [16] when a subset of nodes or edges are removed from a graph. We show that LFVC relates to a monotonic submodular set function which ensures that greedy node or edge removals based on LFVC are nearly as effective as the optimal combinatorial batch removal strategy.

Our approach utilizes LFVC to iteratively remove noisy parts in the graph to reveal deep communities. Under a “signal plus noise” stochastic block model framework [7], [8], [17], we use random matrix theory to show that the greedy LFVC strategy can asymptotically identify the deep communities with probability one. As compared with the modularity method [9] and the L1 norm subgraph detection method [18], [19], we show that the proposed greedy LFVC approach has superior deep community detection performance. We illustrate the proposed deep community detection method on several real-world social networks. When our proposed greedy LFVC approach is applied to the network scientist coauthorship dataset [20], it reveals deep communities that are not identified by conventional community detection methods. When applied to social media, the Last.fm online music dataset, we show that LFVC has the best performance in detecting users with similar interest in artists.

The rest of this paper is organized as follows. Sec. II summarizes commonly used centrality measures, the definitions of community, and relevant spectral graph theory. Sec. III gives a definition of deep communities. The proposed local Fiedler vector centrality (LFVC) is defined in Sec. IV. In Sec. V we introduce the signal plus noise stochastic block model for a deep community and establish asymptotic community detection performance of the greedy LFVC strategy. We apply the greedy LFVC strategy to real-world social network datasets in Sec. VI. Finally, Sec. VII concludes the paper. Throughout the paper we use uppercase letters in boldface (e.g., A) to represent matrices, lowercase letters in boldface (e.g., a) to represent vectors, and uppercase letters in calligraphic face (e.g., $\mathcal{A}$) to represent sets. Subscripts on matrices and vectors
indicate elements (e.g., $A_{ij}$ is the element of $i$-th row and $j$-th column of matrix $A$, and $a_i$ is the $i$-th element of vector $a$). $(\cdot)^T$ denotes matrix and vector transpose.

II. CENTRALITIES, COMMUNITIES, AND SPECTRAL GRAPH THEORY

A. The graph Laplacian matrix and algebraic connectivity

Consider an undirected and unweighted graph $G = (V, E)$ without self loops or multiple edges. We denote by $V$ the node set, with $|V| = n$, and by $E$ the edge set, with $|E| = m$. The connectivity structure of $G$ is characterized by an $n$-by-$n$ binary symmetric adjacency matrix $A$, where $A_{ij} = 1$ if $(i, j) \in E$, otherwise $A_{ij} = 0$. Let $d_i = \sum_{j=1}^{n} A_{ij}$ denote the degree of node $i$. The degree matrix $D = \text{diag}(d)$ is a diagonal matrix with the degree vector $d = [d_1, d_2, \ldots, d_n]$ on its diagonal. The graph Laplacian matrix of $G$ is defined as $L = D - A$. Let $\lambda_1(L)$ denote the $i$-th smallest eigenvalue of $L$ and let $I = [1, \ldots, 1]^T$ denote the vector of ones. We have the representation $[14], [21]$

$$x^T L x = \frac{1}{2} \sum_{i \in V} \sum_{j \in V} A_{ij} (x_i - x_j)^2,$$

(1)

which is nonnegative, and $L I = (D - A) I = 0$. Therefore $\lambda_1(L) = 0$ and $L$ is a positive semidefinite (PSD) matrix.

The algebraic connectivity of $G$ is defined as the second smallest eigenvalue of $L$, i.e., $\lambda_2(L)$. $G$ is connected if and only if $\lambda_2(L) > 0$. Moreover, it is a well-known property [10] that for any non-complete graph,

$$\lambda_2(L) \leq \text{node connectivity} \leq \text{edge connectivity},$$

(2)

where node/edge connectivity is the least number of node/edge removals that disconnects the graph. $[23]$ is the main motivation for our proposed node/edge pruning approach. A graph with larger algebraic connectivity is more resilient to node and edge removals. In addition, let $d_{\text{min}}$ be the minimum degree of $G$, it is also well-known [14], [22] that $\lambda_2(L) \leq 1$ if and only if $d_{\text{min}} = 1$. That is, a graph with a leaf node (i.e., a node with a single edge) cannot have algebraic connectivity larger than 1. For any connected graph, we can represent the algebraic connectivity as

$$\lambda_2(L) = \min_{\|x\|_2 = 1, \ x \perp 1} x^T L x,$$

(3)

by the Courant-Fischer theorem [23] and the fact that the constant vector is the eigenvector associated with $\lambda_1(L) = 0$.

B. Some examples of centralities

Centrality measures can be classified into two categories, global and local measures. Global centrality measures require complete topological information for their computation, whereas local centrality measures only require local topological information from neighboring nodes. Some examples of node centralities are:

- **Betweenness** [24]: betweenness measures the fraction of shortest paths passing through a node relative to the total number of shortest paths in the network. Specifically, betweenness is a global measure defined as

$$\text{betweenness}(i) = \sum_{k \neq i} \sum_{j \neq i, j > k} \frac{\phi_{kj}(i)}{\phi_{kj}},$$

where $\phi_{kj}$ is the total number of shortest paths from $k$ to $j$ and $\phi_{kj}(i)$ is the number of such shortest paths passing through $i$. A similar notion is used to define the edge betweenness centrality [1].

- **Closeness** [23]: closeness is a global measure of geodesic distance of a node to all other nodes. A node is said to have high closeness if the sum of its shortest path distances to other nodes is small. Let $r(i, j)$ denote the shortest path distance between node $i$ and node $j$ in a connected graph. Then we define closeness($i$) = $1/\sum_{j \in V, j \neq i} r(i, j)$.

- **Eigenvector centrality** (eigen centrality) [13]: eigenvector centrality is the $i$-th entry of the eigenvector associated with the largest eigenvalue of the adjacency matrix $A$. It is defined as $\text{eigen}(i) = \lambda_{\text{max}}^{-1} \sum_{j \in V} A_{ij} \xi_j$, where $\lambda_{\text{max}}$ is the largest eigenvalue of $A$ and $\xi$ is the eigenvector associated with $\lambda_{\text{max}}$. It is a global measure since eigenvalue decomposition on $A$ requires global knowledge of the graph topology.

- **Degree** ($d_i$): degree is the simplest local node centrality measure which accounts for the number of neighboring nodes.

- **Ego centrality** [26]: consider the $(d_i+1)$-by-$(d_i+1)$ local adjacency matrix of node $i$, denoted by $A(i)$, and let $I$ be an identity matrix. Ego centrality can be viewed as a local version of betweenness that computes the shortest paths between its neighboring nodes. Since $[A^2(i)]_{kj}$ is the number of two-hop walks between $k$ and $j$, and

$$[A^2(i) \circ (I - A(i))]_{kj}$$

is the total number of two-hop shortest paths between $k$ and $j$ for all $k \neq j$, where $\circ$ denotes entrywise matrix product. Ego centrality is defined as $\text{ego}(i) = \sum_k \sum_{j > k} 1/[A^2(i) \circ (I - A(i))]_{kj}$.

C. Community and modularity

Many possible definitions of communities exist in the literature [4], [10], [12]. One widely adopted definition is based on the relations between the number of internal and external connections of a subgraph $S \subset G$ [27]. For a subgraph $S \subset G$ with node set $V_S$, let $d_{\text{int}}^i(S) = \sum_{j \in S} A_{ij}$ denote the number of internal edges of node $i$ in $S$ and $d_{\text{ext}}^i(S) = \sum_{j \in V \setminus V_S} A_{ij}$ denote the number of external edges of node $i$ outside $S$. $S$ is said to be a community in the strong sense if $d_{\text{int}}^i(S) > d_{\text{ext}}^i(S)$ for all $i \in S$, and $S$ is said to be a community in the weak sense if $\sum_{i \in S} d_{\text{int}}^i(S) > \sum_{i \in S} d_{\text{ext}}^i(S)$.

Newman [9] defines a community by comparing the internal and external connections of a subgraph with that of a random graph having the same degree pattern, and he proposes a quantity called modularity to construct a graph partitioning of $G$ into communities. First consider partitioning a graph into two communities. Recalling that $m = |E|$ is the number of edges in the graph, define $B_{ij} = A_{ij} - \frac{d_i d_j}{2m}$. $B_{ij}$ can be interpreted as the number of excessive edges between $i$ and $j$ since $B_{ij}$ is the difference of actual edges minus the expected edges of the degree-equivalent random graph. Let $s$ be the membership vector such that $s_i = 1$ if $i$ is in community 1 and $s_i = -1$ if $i$ is in community 2. Modularity $Q$ is proportional
to the total number of excess edges in each community, i.e.,
\[ Q = \frac{1}{4m} \sum_{i \in V} \sum_{j \in V} B_{ij} (s_i s_j + 1) = \frac{1}{4m} s^T B s \] (4)
since \( \sum_{j \in V} \sum_{i \in V} B_{ij} = 0 \). Maximizing this quadratic form yields a partition of \( G \) into two communities. The associated membership vector \( s \) can be obtained by computing the largest eigenvector \( \mathbf{b}_{\text{max}} \) of \( B \) and extracting its polarity, i.e., \( s = \text{sgn}(\mathbf{b}_{\text{max}}) \).

To divide a network into more than two communities, Newman proposes a recursive partitioning. It is also verified in [29] that there is no performance difference between the modularity method [9], the statistical inference method [7], [30], and the normalized cut method [6]. However, the modularity method may fail to detect small communities even when community structures are apparent [31]–[33].

In [11], a node removal strategy based on targeting high-degree nodes is proposed to improve the performance of the modularity method. The authors of [11] argue that high-degree nodes incur more noisy connections than low-degree nodes, and it is experimentally demonstrated that removing high-degree nodes can better reveal the community structure.

D. The Fiedler vector

The Fiedler vector of a graph is the eigenvector associated with the second smallest eigenvalue \( \lambda_2(L) \) of the graph Laplacian matrix \( L \) [10]. The Fiedler vector has been widely used in graph partitioning, image segmentation, and data clustering [5], [6], [34]–[36]. Analogously to modularity partitioning, the Fiedler vector performs community detection by separating the nodes in the graph according to the signs of the corresponding Fiedler vector elements. Similarly, hierarchical community structure can be detected by recursive partitioning with the Fiedler vector.

In this paper, we use the Fiedler vector to define a new centrality measure. One advantage of using the Fiedler vector over other global centrality measures is that it can be computed in a distributed manner via local information exchange over the graph Laplacian matrix associated with the network [37].

III. DEEP COMMUNITY

A deep community is defined in terms of an additive signal (community) plus noise model. Let \( A_1, \ldots, A_g \) denote the \( n \times n \) mutually orthogonal binary adjacency matrices associated with \( g \) non-singleton connected components in a noiseless graph \( G_0 \) over \( n \) nodes. Assume the nodes have been permuted so that \( A_1, \ldots, A_g \) are block diagonal with block indices \( I_1, \ldots, I_g \). The observed graph \( G \) is a noise corrupted version of \( G_0 \) where random edges have been inserted between the connected components of \( G_0 \). More specifically, let \( A_{\text{mse}} \) be a random adjacency matrix with the property that \( A_{\text{mse}}(i, j) = 0 \), \( i, j \in I_k \), for \( k = 1, \ldots, g \) and where the rest of the elements of \( A_{\text{mse}} \) are Bernoulli i.i.d random variables. Then the adjacency matrix \( A \) of \( G \) satisfies the signal plus noise model
\[ A = \sum_{k=1}^{g} A_k + A_{\text{mse}}. \] (5)

The deep community detection problem is to recover connected components \( A_1, \ldots, A_g \) from the noise corrupted observations \( A \). The \( A_k \)'s are called deep communities in the sense that they are embedded in a graph with random interconnections between connected components. Models similar to (5) have been used for hypothesis testing on the existence of dense subgraphs embedded in random graphs [18], [19]. The null hypothesis is the noise only model (i.e., \( A_k = 0 \) \( \forall k \)). The alternative hypothesis is the signal plus noise model (5) with \( A_k \neq 0 \). The authors in [18], [19] propose to use the L1 norm of the eigenvectors of the modularity matrix \( B \) as test statistics. This statistic is compared with our proposed local Fiedler vector deep community detection method in Sec. IV.

We formulate deep community detection as an iterative node/edge removal procedure on the observed adjacency matrix \( A \) where centrality measures are used to determine the nodes/edges to be pruned. The proposed spectral centrality measure (LFVC) in Sec. IV is proposed for this purpose. For node removal based deep community detection, after some subset of nodes are removed from the graph, a discovered deep community is defined as a remaining non-singleton connected subgraph augmented by the removed nodes that were previously adjacent to the subgraph before their removal. For edge removal based deep community detection, a discovered deep community is defined as a remaining non-singleton connected subgraph.

More specifically, for the node removal procedure consider a non-singleton connected subgraph \( S \) with node set \( V_S \) after removal of a set of nodes \( R \) from \( G \). The discovered deep community is defined as the set \( V_S \cup \{ i \in R : A_{ij} = 1 \ \text{for some} \ j \in S \} \).

Let \( L \) be the resulting \( n \times n \) graph Laplacian matrix after removing a subset of nodes or edges from the graph. The following theorem provides an upper bound on the number of deep communities in the remaining graph \( G \).

Theorem 1. For any node removal set \( R \) of \( G \) with \( |R| = q \), let \( r \) be the rank of the resulting graph Laplacian matrix \( L \) and let \( \|L\|_s = \sum_{i} \lambda_i(L) \) denote its nuclear norm. The number \( \epsilon \) of remaining non-singleton connected components in \( G \) has the upper bound
\[ \epsilon \leq n - q - r \]
\[ \leq n - q - \frac{\|L\|_s}{\lambda_n(L)} \]
\[ = n - q - \frac{2\tilde{m}}{\lambda_n(L)}, \] (6)
where \( \tilde{m} \) is the number of edges in \( G \). The first equality in (6) is attained if all connected components in \( G \) are non-singletons. The second equality in (6) is attained if all non-singleton connected components are complete subgraphs of the same size. Similarly, for any edge removal set of \( G \), let \( r \) be the rank of the resulting graph Laplacian matrix \( L \). The
number $\epsilon$ of remaining non-singleton connected components in $\tilde{G}$ has the upper bound $\epsilon \leq n - r \leq n - \frac{\|L\|_s}{\lambda_n(L)} = n - 2\tilde{m} / \lambda_n(L).

Proof: From (4) a graph is connected if and only if the algebraic connectivity is greater than zero. Furthermore, the smallest eigenvalue of the associated graph Laplacian matrix is always 0. Therefore $n - q - r$ is the number of connected components (including the singleton nodes) in $G$ \cite{14}. Since the definition of a complete graph excludes singleton nodes, the first inequality in (4) becomes equality if all connected components in $G$ are non-singleton.

Using a well-known matrix norm inequality \cite{23} that $\|M\|_* \leq r\|M\|_2$ for any square matrix $M$ of rank $r$, where $\|M\|_2 = \max_{\|x\|_2=1}\|Mx\|_2 = \lambda_n(M)$. We have

$$n - q - r \leq n - q - \frac{\|L\|_s}{\lambda_n(L)} = n - q - \frac{2\tilde{m}}{\lambda_n(L)},$$

where $\|L\|_s = \text{trace}(L) = 2\tilde{m}$ is the total degree of $\tilde{G}$.

Next we show that the second equality in (5) is attained if each non-singleton connected graph is a complete subgraph of the same size. Consider a graph consisting of $g$ disjoint complete subgraphs of $n' \geq 2$ nodes and $n'(n' - 1)/2$ edges. The largest eigenvalue of each subgraph is $n'$ and $\|L\|_* = g \cdot n'(n' - 1)$. The upper bound becomes $g \cdot n' - \frac{g(n'(n' - 1))}{n'} = g$, which is exactly the number of non-singleton connected components in $G$. These results can be directly applied to edge removals in $G$ by setting $q = 0$ since no nodes are removed.

The upper bound in Theorem 1 can be further relaxed by applying the inequality $\lambda_n(L) \leq 2d_{\text{max}}$ \cite{14}, where $d_{\text{max}}$ is the maximum degree of $\tilde{G}$. Other bounds on $\lambda_n(L)$ can be found in \cite{38}.

The next theorem shows that the largest non-singleton connected component size can be represented as a matrix one norm of a matrix whose column vectors are orthogonal and sparsest among all binary vectors that form a basis of the null space of $L$.

**Theorem 2.** Define the sparsity of a vector to be the number of zero entries in the vector. Let $\text{null}(L)$ denote the null space of $L$ and let $X$ denote the matrix whose columns are orthogonal and they form the sparsest basis of $\text{null}(L)$ among binary vectors. Let $\psi(G)$ denote the largest non-singleton connected component size of $\tilde{G}$. Then $\psi(G) = \|X\|_1 = \max_i \|x_i\|_1$, where $x_i$ is the $i$-th column vector of binary matrix $X$.

**Proof:** Let $r$ be the rank of $L$. We prove that there exists an $n \times (n - r)$ binary matrix $X = [x_1 \ x_2 \ldots \ x_{n-r}]$ whose columns $\{x_i\}_{i=r}^n$ satisfy: 1) $\|x_i\|_1$ is the size of the $i$-th connected component of $\tilde{G}$; 2) they are orthogonal; 3) they span $\text{null}(L)$. Assume $G$ consists of $g$ connected components. Then there exits a matrix permutation (node relabeling) such that

$$\tilde{L} = \begin{bmatrix} \tilde{L}_1 & 0 & 0 & 0 \\ 0 & \tilde{L}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \tilde{L}_g \end{bmatrix}.$$  \hfill (7)

Associated with the $i$-th block matrix $\tilde{L}_i$, we define $x_i$ as an $n \times 1$ binary vector $x_i$ in $\text{null}(L)$ having the form $x_i = [0 \ldots 0 \ 1 \ldots 0 \ldots 0]^T$, where the locations of the nonzero entries correspond to the indexes of the $i$-th block matrix. It is obvious that $\|x_i\|_1 = \sum_{i=1}^n |x_{ij}|$ equals the size of the $i$-th component and such $\{x_i\}_{i=r}^n$ are mutually orthogonal. Furthermore, there exists no other binary matrix which is sparser than $X$ with column span equal to $\text{null}(L)$. If there existed another binary matrix that were sparser than $X$, then it would contradict the fact that its column vectors have sums equal to the component sizes of $\tilde{G}$. Therefore the largest non-singleton connected component size of $G$ is $\psi(G) = \|X\|_1 = \max_i \|x_i\|_1$.

**IV. THE PROPOSED NODE AND EDGE CENTRALITY:**

**LOCAL FIEDLER VECTOR CENTRALITY (LFVC)**

In this section, we motivate the LFVC centrality measure through a sensitivity analysis of algebraic connectivity subject to node/edge removals. In particular, we show that node/edge removals result in low rank perturbations to the graph Laplacian matrix when $n > d_{\text{max}}$, where $d_{\text{max}}$ is the maximum degree. The node and edge LFVC are then defined to correspond to an upper bound on algebraic connectivity.

**A. Edge-LFVC**

Considering the graph $\tilde{G}(i,j) = (V, E \cup (i, j))$ by adding an edge $(i,j) \notin E$ to $G$, we have $L = L + \Delta L$ and $\Delta L = \Delta D - \Delta A$, where $\Delta D$ and $\Delta A$ are the augmented degree and adjacency matrices, respectively. Denote the resulting graph Laplacian matrix by $\tilde{L}(i,j)$. Let $e_i$ be a zero vector except that its $i$-th element is equal to 1. Then

$$\Delta D = \text{diag}(e_i) + \text{diag}(e_j) = e_i e_i^T + e_j e_j^T; \quad \Delta A = e_i e_j^T + e_j e_i^T,$$

and therefore

$$\tilde{L}(i,j) = L + (e_i - e_j)(e_i - e_j)^T.$$  \hfill (10)

Thus, the resulting graph Laplacian matrix $\tilde{L}(i,j)$ after adding an edge $(i,j)$ to $G$ is the original graph Laplacian matrix $L$ perturbed by a rank one matrix $(e_i - e_j)(e_i - e_j)^T$. Similarly, when an edge $(i,j) \in E$ is removed from $G$, we have $\tilde{L}(i,j) = L - (e_i - e_j)(e_i - e_j)^T$.

Considering removing an edge $(i,j) \in E$ from $G$ resulting in $\tilde{L}(i,j)$ above. Let $y$ denote the Fiedler vector of $L$, computing $y^T \tilde{L}(i,j)y$ gives an upper bound on $\lambda_2(L(i,j))$ as

$$\lambda_2(\tilde{L}(i,j)) \leq y^T \tilde{L}(i,j)y = y^T (L - (e_i - e_j)(e_i - e_j)^T)y = \lambda_2(L) - (y_i - y_j)^2$$

following the definition of $\lambda_2(L) = \min_{\|x\|_2=1, x \perp 1} x^T L x$ in \cite{3}. It is worth mentioning that for any connected graph $G$ there exists at least one edge removal such that the inequality $\lambda_2(\tilde{L}(i,j)) < \lambda_2(L)$ holds, otherwise $y_i = y_j$ for all $i,j \in V$ and this violates the constraints that $|y_i| = 1$ and $\sum_{i=1}^n y_i = 0$. Consequently, there exists at least one edge removal that leads to a decrease in algebraic connectivity.
Similarly, when we remove a subset of edges $E_R \subset E$ from $G$, where $|E_R| = h$. We obtain an upper bound

$$\lambda_2(\tilde{L}(E_R)) \leq \lambda_2(L) - \sum_{(i,j) \in E_R} (y_i - y_j)^2. \quad (12)$$

Correspondingly, we define the local Fiedler vector edge centrality as

$$\text{edge-LFVC}(i,j) = (y_i - y_j)^2. \quad (13)$$

The top $h$ edge removals which lead to the largest decrease on the right hand side of (12) are the $h$ edges with the highest edge-LFVC.

B. Node-LFVC

When a node $i \in V$ is removed from $G$, all the edges attached to $i$ will also be removed from $G$. Similar to (11), the resulting graph Laplacian matrix $\tilde{L}(i)$ can be regarded as a rank $d_i$ matrix perturbation to $L$. Since $L - \tilde{L}(i) = \sum_{j \in N_i} (e_i - e_j)(e_i - e_j)^T$, where $N_i$ is the set of neighboring nodes of node $i$, we obtain an upper bound

$$\lambda_2(\tilde{L}(i)) \leq y^T \tilde{L}(i)y$$

$$= y^T (L + \tilde{L}(i) - L)y$$

$$= \lambda_2(L) - \sum_{j \in N_i} (y_i - y_j)^2. \quad (14)$$

Similar to edge removal, for any connected graph, there exists at least one node removal that leads to a decrease in algebraic connectivity.

If a subset of nodes $R \subset V$ are removed from $G$, where $|R| = q$, then

$$L - \tilde{L}(R) = \sum_{i \in R} \sum_{j \in N_i} (e_i - e_j)(e_i - e_j)^T$$

$$- \frac{1}{2} \sum_{i \in R} \sum_{j \in R} A_{ij}(e_i - e_j)(e_i - e_j)^T,$$

where the last term accounts for the edges that are attached to the removed nodes at both ends. Consequently, similar to (11), we obtain an upper bound for multiple node removals

$$\lambda_2(\tilde{L}(R)) \leq \lambda_2(L) - \sum_{i \in R, j \in N_i} (y_i - y_j)^2 \quad (16)$$

$$+ \frac{1}{2} \sum_{i \in R} \sum_{j \in R} A_{ij}(y_i - y_j)^2.$$

We define the local Fiedler vector node centrality as

$$\text{node-LFVC}(i) = \sum_{j \in N_i} (y_i - y_j)^2, \quad (17)$$

which is the sum of the square terms of the Fiedler vector elementwise differences between node $i$ and its neighboring nodes, and it is also the sum of edge-LFVC of $i$'s neighboring nodes. From (14) and (16), node-LFVC is associated with the upper bound on the resulting algebraic connectivity for node removal when $|R| = 1$. A node with higher centrality implies that it plays a more important role in the network connectivity structure.

C. Monotonic submodularity and greedy removals

Fixing $|R| = q$, consider the problem of finding the optimal node removal set $R_{\text{opt}}$ that maximizes the decrease in the upper bound on algebraic connectivity in (16). The computational complexity of this batch removal problem is of combinatorial order $\binom{n}{q}$. Here we show that the greedy LFVC removal procedure, shown in Algorithm 1 and whose computation is only linear in $n$, has bounded performance loss relative to the combinatorial algorithm in terms of achieving, within a multiplicative constant $1 - 1/e$, an upper bound on algebraic connectivity, where $e$ is Euler’s constant. Let

$$f(R) = \sum_{i \in R} \sum_{j \in N_i} (y_i - y_j)^2 - \frac{1}{2} \sum_{i \in R} \sum_{j \in R} A_{ij}(y_i - y_j)^2$$

and recall from (16) that $\lambda_2(\tilde{L}(R)) \leq \lambda_2(L) - f(R)$. Note that when $|R| = 1$, $f(R)$ reduces to node-LFVC as $A_{ii} = 0$. The following lemma provides the cornerstone to Theorem 3.

**Lemma 1.** The function $f(R)$ in (18) is equal to

$$f(R) = \frac{1}{2} \sum_{i \in R} \sum_{j \in V} (y_i - y_j)^2 - \frac{1}{2} \sum_{i \in R} \sum_{j \in R} A_{ij}(y_i - y_j)^2.$$

Furthermore, $f(R) \geq 0$ and $f(\emptyset) = 0$, where $\emptyset$ is the empty set.

**Proof:** By the relation

$$\sum_{i \in R} \sum_{j \in V} A_{ij} = \sum_{i \in R} \sum_{j \in N_i} (y_i - y_j)^2 \quad (19)$$

and $V = \{V/R\} \cup \{R\}$, we have

$$f(R) = \sum_{i \in R} \sum_{j \in N_i} (y_i - y_j)^2 - \frac{1}{2} \sum_{i \in R} \sum_{j \in V} A_{ij}(y_i - y_j)^2$$

$$+ \frac{1}{2} \sum_{i \in R} \sum_{j \in V/R} A_{ij}(y_i - y_j)^2$$

$$= \sum_{i \in R} \sum_{j \in N_i} (y_i - y_j)^2 - \frac{1}{2} \sum_{i \in R} \sum_{j \in V/R} A_{ij}(y_i - y_j)^2 \geq 0. \quad (20)$$

$f(\emptyset) = 0$ follows directly from the definition of $f(R)$ in (18).

The following theorem establishes monotonic submodularity of $f(R)$. As will be seen below (see (27)), this implies that greedy node removal based on LFVC is almost as effective as the combinatorially complex batch algorithm that searches over all possible removal sets $R$.  

**Theorem 3.** $f(R)$ is a monotonic submodular set function.

**Proof:** We first prove the monotonic property. Consider two node removal sets $R_1 \subset R_2 \subset V$. Then using Lemma 1
and the fact that $R_1/R_2 = \emptyset$,

$$f(R_2) - f(R_1) = \sum_{i \in R_2} \sum_{j \in N_i} (y_i - y_j)^2 - \sum_{i \in R_1} \sum_{j \in R_2} A_{ij}(y_i - y_j)^2$$

$$= \sum_{i \in R_2/R_1} \sum_{j \in N_i} A_{ij}(y_i - y_j)^2 - \sum_{i \in R_1/j \in R_2} A_{ij}(y_i - y_j)^2$$

$$\geq 0. \quad (21)$$

Therefore $f(R)$ is a monotonic increasing set function (i.e., $f(R_2) \geq f(R_1)$ for all $R_1 \subset R_2 \subset V$).

Furthermore, $f(R)$ is a submodular set function \([39, 40]\) since for any node $v \in V, v \notin R_2, R_1 \subset R_2 \subset V$, we have from \([13]\) that

$$f(R_1 \cup v) - f(R_1) = \sum_{j \in N_v} (y_v - y_j)^2 - \sum_{j \in R_1} A_{ij}(y_v - y_j)^2$$

$$\geq \sum_{j \in N_v} (y_v - y_j)^2 - \sum_{j \in R_2} A_{ij}(y_v - y_j)^2$$

$$= f(R_2 \cup v) - f(R_2). \quad (22)$$

This diminishing returns property of $f(R)$ implies that it is a submodular set function.

Based on Theorem 3, we propose a greedy node-LFVC based node removal algorithm for deep community detection as summarized in Algorithm 1. The following theorem shows that this greedy algorithm has bounded performance loss compared with the optimal combinatorial batch removal strategy.

**Theorem 4.** Fix the target number of nodes to be removed as $|R| = q$. Let $R_{opt}$ be the optimal node removal set that maximizes $f(R)$ and let $R_k$ be the greedy node removal set at the $k$th stage of Algorithm 1 where $|R_k| = k$. Then

$$f(R_{opt}) - f(R_q) \leq \frac{1}{e} f(R_{opt}).$$

Furthermore,

$$\lambda_2(\mathbf{L}(R_q)) \leq \lambda_2(\mathbf{L}) - (1 - e^{-1}) f(R_{opt}). \quad (23)$$

**Proof:** By submodularity of $f(R)$ in Theorem 3 there exists a $v \in R_{opt}/R_k$ \([40]\) such that

$$f(R_k \cup v) - f(R_k) \geq \frac{1}{q} (f(R_{opt}) - f(R_k)). \quad (24)$$

After algebraic manipulation, we have

$$f(R_{opt}) - f(R_{k+1}) \leq \left(1 - \frac{1}{q}\right) (f(R_{opt}) - f(R_k)) \quad (25)$$

and therefore

$$f(R_{opt}) - f(R_q) \leq \left(1 - \frac{1}{q}\right)^q f(R_{opt}) \leq \frac{1}{e} f(R_{opt}). \quad (26)$$

Applying this result to \([16]\), we have

$$\lambda_2(\mathbf{L}(R_q)) \leq \lambda_2(\mathbf{L}) - (1 - e^{-1}) f(R_{opt}). \quad (27)$$

The submodularity of the function $f$ implies that after $q$ greedy iterations the performance loss is within a factor $1/e$ of optimal batch removal \([39]\). In other words, when removing $R_q$ from $G$, the algebraic connectivity is guaranteed to decrease by at least $(1 - e^{-1}) f(R_{opt})$ of its original value. Consequently, identifying the top $q$ nodes affecting algebraic connectivity can be regarded as a monotonic submodular set function maximization problem, and the greedy algorithm can be applied iteratively to remove the node with the highest node-LFVC. Similarly, we can use edge-LFVC to detect deep communities by successively remove the edge with the highest edge-LFVC from the graph, and it is easy to show that the term $\sum_{(i,j) \in E}(y_i - y_j)^2$ in \([12]\) is a monotonic submodular set function of the edge removal set $E_R$.

**V. DEEP COMMUNITY DETECTION IN STOCHASTIC BLOCK MODEL**

To demonstrate the effectiveness of using the proposed LFVC for deep community detection, we compare its detection performance to that of other methods for a synthetic network generated by a stochastic block model \([7, 8, 17]\). Consider a deep community of size $n_{in}$ embedded in a network of size $n, n_{in} < n$, and let $n_{out} = n - n_{in}$ denote the rest of the graph size. The average number of edges between the members...
of the deep community is denoted by $c_{in}$, and the average number of edges between the members that are not in the deep community is denoted by $c_{out}$. We assume a restricted stochastic block model characterized by the following group connected probability matrix

$$P = \begin{pmatrix}
\text{deep} & \text{others} \\
\text{others} & \begin{pmatrix} p_{in} & p_{out} \\
p_{out} & p_{out} \end{pmatrix}
\end{pmatrix},$$

(28)

where $p_{in} = \frac{c_{in}}{n} = \frac{\ell_{in}}{n}$ and $p_{out} = \frac{c_{out}}{n} = \frac{\ell_{out}}{n}$ are the edge connection probabilities within and outside the deep community, respectively. The planted clique problem [41] is a special case of the stochastic block model in (28) when $p_{in} = 1$. The detection performance of the modularity method has been analyzed in [42] for the planted clique problem.

The adjacency matrix generated by the stochastic block model in (28) is a random binary matrix with partitioned structure

$$A = \begin{bmatrix} A_{in} & C \\ C^{T} & A_{out} \end{bmatrix},$$

(29)

where $A_{in}$ and $A_{out}$ are the adjacency matrices of a Erdos-Renyi graph with edge connection probabilities $p_{in}$ and $p_{out}$, respectively. $C$ is an $n_{in}$-by-$n_{out}$ binary matrix with its entry being 1 with probability $p_{out}$. $A$ reduces to the special case of (3) when $g = 1$. $A_{in} = \begin{bmatrix} A_{in} & 0 \\ 0 & 0 \end{bmatrix}$ represents within-community connectivity structure and $A_{out} = \begin{bmatrix} 0 & C \\ C^{T} & A_{out} \end{bmatrix}$ represents the noisy part outside the deep community.

In the following paragraphs we use random matrix theory and concentration inequalities to show that asymptotically the nodes/edges in the noisy part are more likely to have high LFVC. Therefore the removal strategy based on LFVC will with high probability detect the noisy part to reveal the deep community.

Let $1_{in}$ be the all-one vector of length $n_{in}$ and $1_{out}$ be the all-one vector of length $n_{out}$, and let $D_{in} = \text{diag}(C^{1}_{out})$ and $D_{out} = \text{diag}(C^{T}1_{in})$. Following (29), the corresponding graph Laplacian matrix can be represented as

$$L = \begin{bmatrix} L_{in} + D_{in} & -C \\ -C^{T} & L_{out} + D_{out} \end{bmatrix},$$

(30)

where $L_{in}$ and $L_{out}$ are the graph Laplacian matrices of the two Erdos-Renyi graphs.

**Theorem 5.** Let $n = n_{in} + n_{out}$ and let $y = [y_{in} \ y_{out}]^{T}$ be the Fiedler vector of the graph Laplacian matrix $L$ in (29). Consider the stochastic block model in (28). For a fixed $p_{in}$ there exists an asymptotic threshold $p_{out}^{*}$ such that almost surely,

$$\begin{cases}
\sqrt{n_{in}} y_{in} \rightarrow \pm 1_{in} & \sqrt{n_{out}} y_{out} \rightarrow \mp 1_{out}, \quad \text{if } p_{out} \leq p_{out}^{*} \\
1^{T}_{in} y_{in} \rightarrow 0 & 1^{T}_{out} y_{out} \rightarrow 0, \quad \text{if } p_{out} > p_{out}^{*}
\end{cases}$$

as $n_{in} \rightarrow \infty$, $n_{out} \rightarrow \infty$ and $\frac{\ell_{in}}{n_{in}} \rightarrow c > 0$.

**Proof:** The proof can be found in Appendix A.

Consequently, when $p_{out} \leq p_{out}^{*}$ the elements of the Fiedler vector tend to have opposite signs within and outside the deep community. Recalling the edge and node LFVC in (13) and (17), Theorem 5 implies that the nodes/edges with high LFVC are more likely to be present outside the deep community. On the other hand, when $p_{out} > p_{out}^{*}$ both $y_{in}$ and $y_{out}$ have alternating signs. Therefore the nodes/edges with high LFVC are more likely to be present within the deep community, resulting in incorrect node/edge removals. One can interpret $c_{in} = n_{in} \cdot p_{in}$ as the signal strength and $c_{out} = n_{out} \cdot p_{out}$ as the noise level. Based on Theorem 5 for a fixed $c_{out}$ there exists an asymptotic signal strength threshold $c_{in}^{*}$ such that $y_{in}$ and $y_{out}$ become constant vectors with opposite signs when $c_{in} \geq c_{in}^{*}$. These results are consistent with the planted clique detection analysis in [42] that a clique is detectable if the ratio of its within-clique connections to its outside-clique connections is above a certain threshold.

The proposed deep community detection method in Algorithm 1 is implemented by sequentially removing the node (edge) with the highest LFVC until the graph is decomposed into several subgraphs. For comparison, the modularity method is implemented by dividing the network into two communities. The L1 norm subgraph detection method [18], [19] is also implemented. For the null model (i.e., absence of community structure) of the L1 norm subgraph detection method, we generate 500 Erdos-Renyi random graphs with edge connection probability $p_{out}$ and compute the mean and standard deviation of the L1 norm of each eigenvector associated with the modularity matrix. Let $\text{mean}(i)$ and $\text{std}(i)$ denote the mean and standard deviation of the L1 norm of $i$-th eigenvector in the null model and let $\ell_{i}$ denote the L1 norm of the $i$-th eigenvector associated with the modularity matrix $B$. The test statistic is

$$t = \min_{i=1,...,n} \frac{\ell_{i} - \text{mean}(i)}{\text{std}(i)}.$$  

(31)

The presence of a dense subgraph is declared if $|t| \geq 2$, which corresponds to 5% false alarm probability [18], [19]. Let $i^{*} = \arg \text{min}_{i=1,...,n} \frac{\ell_{i} - \text{mean}(i)}{\text{std}(i)}$. The deep community is identified by selecting the $n_{in}$ entries of the $i^{*}$-th eigenvector having the largest magnitude.

Consider detecting a deep community of size 40 embedded in a network of size 200 generated by the stochastic block model (28). As shown in Fig. 1 deep community detection based on LFVC is capable of removing the noisy part while retaining the deep community component. On the other hand, the modularity method is overly influenced by the noisy part resulting in incorrect community detection. In particular it incorrectly identifies members of the deep community as different groups, especially when $\frac{c_{out}}{c_{in}}$ is small. The inaccuracy of the modularity method is due to the fact that in the low $\frac{c_{out}}{c_{in}}$ regime, the network is overly similar to the corresponding degree-equivalent random graph model used to define the modularity metric. That is, modularity does not provide enough evidence that a community is present. Nonetheless, a substantial improvement of the modularity method is observed after using LVFC to denoise the graph (curve labeled "deep modularity" in Fig. 1).

The fraction of removed nodes via LFVC is shown in Fig. 2. In this setting, the fraction of removed nodes via node-LFVC is slightly less than that via edge-LFVC. Observe that
Fig. 1. Fraction of correctly identified deep community as embedded in simulated stochastic block model using (28). $n_{in} = 40$, $n_{out} = 200$, and $c_{out} = 2$. The results are averaged over 500 trials. The proposed deep community detection algorithm based on LFVC is capable of removing the noisy part and uncovering the community structure. The performance of the modularity method can be improved by removing nodes/edges with high LFVC (curve labeled deep-modularity).

Fig. 2. Fraction of removed nodes via node and edge LFVC for simulated stochastic block model using (28). $n_{in} = 40$, $n_{out} = 200$, and $c_{out} = 2$. The results are averaged over 500 trials. Comparing to Fig. 1, when $c_{in}/c_{out} \geq 2.5$, the proposed deep community detection algorithm based on LFVC can remove some noisy part from the observed adjacency matrix while retaining the deep community structure. The results validate Theorem 5 that the Fiedler vector tends to have opposite signs within and outside the deep community when $c_{in}$ is above the threshold $c_{in}^* = 5$.

The fraction of removed nodes is stable when $c_{in}/c_{out} \geq 2.5$. The results validate Theorem 5 that the Fiedler vector tends to have opposite signs within and outside the deep community when $c_{in}$ is above the threshold $c_{in}^* = 5$. Consequently, the results show that removal of nodes/edges with high LFVC helps to reveal the true community structure and improve community detection performance.

Fig. 3. Dolphin social network [43] with $n = 62$ nodes and $m = 159$ edges. (a) The modularity method. (b) Edge-LFVC community detection with $h = 6$ edge removals. (c) Node-LFVC community detection with $q = 4$ node removals. Using node-LFVC, we are able to identify the four dolphins that interact with two groups as marked in gray circles.

VI. DEEP COMMUNITY DETECTION ON REAL-WORLD SOCIAL NETWORK DATASETS

In this section, we use the proposed node and edge centrality measures to perform deep community detection on several datasets collected from real-world social networks. The community structures are revealed by using LFVC as summarized in Algorithm 1. We define $h$ the number of edge removals, $q$ the number of node removals and $g$ the number of deep communities. The results are compared with the modularity method and other node centralities introduced in Sec. II-B. For data visualization, vertex shapes and colors represent different communities, and edges attached to the removed nodes are retained in the figures in comparison with other methods. Nodes with cross labels (black X labels) are singleton survivors that do not belong to any deep communities using Algorithm 1.

A. Dolphin social network

It is shown in [43] that there are tight social structures in dolphin populations. Most dolphins interact with other dolphins of the same group and only a few dolphins can interact with dolphins from different groups. This observation indicates the presence of overlapping community structures, where a node is assigned with multiple memberships if it can communicate with different groups.

We compare the results of separating 62 dolphins into two communities as proposed in [43]. For this dataset, community detections based on modularity, edge-LFVC and node-LFVC have high concordance on the community structures as shown in Fig. 3. To partition the graph into two communities, we need to remove 6 edges based on edge-LFVC or remove 4 nodes based on node-LFVC. The four dolphins that are able to communicate between these two communities are further identified by node-LFVC.
B. Zachary’s karate club

Zachary’s karate club [44] is a widely used example for social network analysis, which contains interactions among 34 karate students. Based on the student activities, Zachary determines the ground-truth community structure for $g = 2$, which coincides with the result of the modularity method in Fig. 4 (a). However, the visualization indicates that there are some deep communities embedded in these two communities, such as the five-node community in the upper left corner. Indeed, the modularity will keep increasing if we further divide communities into 3 and 4 small communities as shown in Fig. 5 (b) and (c), respectively.

As shown in Fig. 5 (a), using edge-LFVC, the five-node community in the left upper corner is revealed when we partition the graph into two connected subgraphs. In Fig. 5 (b), three communities are revealed and the only node with a single acquaintance is excluded from any deep community. Excluding this node makes the community structure more tightly connected compared with Fig. 5 (b). For $g = 4$, the community structure in Fig. 5 (c) much resembles Fig. 4 (c) except that we exclude the node having a single acquaintance.

Using node-LFVC, we are able to extract important communities and key members as shown in Fig. 6. For $g = 2$, only one node removal is required to partition the graph into two connected subgraphs, which implies that this node is common to the two communities. For $g = 3$, two deep communities (green circle and blue triangle) are discovered in the largest community (the blue triangle community in Fig. 6 (a)), where these two deep communities have dense internal connections compared with the external connections to other members in the largest community. These discovered deep communities are important communities embedded in the network since they play an important role in connecting the singleton survivors indicated by black X labels. Similar observations hold for $g = 4$ in Fig. 6 (c).

C. Coauthorship among network scientists

We next examine the coauthorship network studied by Newman [20]. Nodes represent network scientists and edges represent the existence of coauthorship. Multiple memberships are expected to occur in this dataset since a network scientist may collaborate with other network scientists across different regions all the while having many collaborations with his/her colleagues and students at the same institution. As a result, one would expect node-LFVC to be advantageous for detecting deep communities.

As shown in Fig. 7, the first node with the highest node-LFVC is Yamir Moreno, who is a network scientist in Spain but has many collaborators outside Spain. The local (two-hop) coauthorship network of Yamir Moreno is shown in Fig. 7. The red square community represents the network scientists in Spain and Europe, whereas the blue triangle community...
represents the rest of the network scientists.

After removing Yamir Moreno from the network, the node with the highest node-LFVC in the remaining largest community is Mark Newman, who is associated with 5 community memberships and 3 singleton survivors as shown in Fig. 8. Each community can be related to certain relationship such as colleagues, students and research institutions. Notably, Lusseau is detected as a singleton survivor in the deep community detection process in Fig. 8. This can be explained by the fact that although Lusseau has coauthorship with Newman, his research area is primarily in zoology and he has no interactions with other network scientists in the dataset since other network scientists are mainly specialists in physics. Also note that the modularity method (gray dashed box) fails to detect these deep communities and it detects 25 out of 28 network scientists in Fig. 8 as one big community.

D. Last.fm online music system

Last.fm is an online music system which allows users to tag their favorite songs and artists and make friends with other users. We use the friendship dataset collected in [45] for deep community detection based on node-LFVC and the other centralities introduced in Sec. II-B. Two quantities, the...
normalized largest community size and the number of discovered communities with respect to node removals, are used to evaluate the performance of community detection when different node centralities are applied. These two quantities reflect the effectiveness of graph partitioning. The number of removed nodes is the number of stages for performing deep community detection and removing more nodes reveals more deep communities and key members in the network.

As shown in Fig. 9(a), the normalized largest community size decays linearly with respect to the number of node removals. Among all node centralities, node-LFVC has the steepest decaying rate. Furthermore, using node-LFVC discovers more deep communities, as shown in Fig. 9(b) during the first 50 node removals. The only node centrality that is comparable to node-LFVC is betweenness centrality.

To validate the effectiveness of deep community detection, we use the user-artists dataset in [45] to compute the listening similarity in each discovered community. The dataset contains 17632 artists and records the number of times each user has listened to an artist. Let \( w_i \) be a 17632-by-1 vector with its \( j \)-th entry being the number of times of the \( i \)-th user has listened to the \( j \)-th artist. The residual community similarity (RCS) is defined as the sum of cosine similarity between each user in the same community excluding the nodes that have been removed and the singleton survivors. The residual community similarity of a deep community \( C_k \) is defined as

\[
\text{RCS}(C_k) = \sum_{i \in C_k, i \neq R} \sum_{j \in C_k, j > i, j \neq R} \frac{w_i^T w_j}{\|w_i\|_2 \|w_j\|_2}.
\] (32)

The residual sum of community similarity (RSCS) is defined as the sum of RCS of each discovered community. That is,

\[
\text{RSCS} = \sum_{k=1}^g \text{RCS}(C_k).
\] (33)

As shown in Fig. 10, the residual sum of community similarity based on node-LFVC is larger than that for other centralities. This suggests that node removals based on node-LFVC can best detect friendship communities that share common interest in artists. Note that although betweenness may detect more communities in Fig. 9(b), Fig. 10 shows that the residual sum of community listening similarity based on betweenness is smaller than that based on node-LFVC, which indicate that node-LFVC reveals more accurate community structure than betweenness. The residual sum of community similarity decreases with respect to the number of discovered communities due to the fact that the removed nodes and singleton survivors are excluded for similarity computation.

**VII. CONCLUSION**

Based on bounds on the sensitivity of algebraic connectivity to node or edge removals, we proposed a centrality measure called local Fiedler vector centrality (LFVC) for deep community detection. We proved that LFVC relates to a monotonic submodular set function such that greedy node removals based on LFVC can be applied to identify the most vulnerable nodes or edges with bounded performance loss compared to the optimal combinatorial batch removal strategy. Asymptotic results on the Fiedler vector are provided to show that LFVC can successfully remove the noisy part while retaining the deep community structure in networks generated by the stochastic block model. In comparison to the modularity method [9] and the L1 norm subgraph detection method [19], we show that LFVC can achieve better community detection performance in correctly identifying the embedded deep communities.

The proposed method provides better resolution for discovering important communities and key members in the real-world social network datasets studied here. In particular, for the Last.fm online music system dataset, LFVC is shown to significantly outperform other centralities for deep community detection in terms of the residual sum of community listening similarity. This new measure can likely offer new insights on community structure in other social, biological and technological networks.

**APPENDIX A**

**PROOF OF THEOREM 6**

Let \( x = [x_{in} \space x_{out}]^T \). By (3) we have \( \lambda_2(L) = \min_{k \neq 1} x^T L x \) subject to \( x_{in}^T x_{in} + x_{out}^T x_{out} = 1 \) and \( x_{in}^T 1_{in} + x_{out}^T 1_{out} = 0 \). Using Lagrange multipliers \( \mu, \nu \) and (30), the Fiedler vector \( y = [y_{in} \space y_{out}]^T \) of \( L \) is a minimizer of the function over \( x \):

\[
\Gamma = x_{in}^T [(L_{in} + D_{in}) y_{in} + x_{out}^T (L_{out} + D_{out}) y_{out} - 2x_{in}^T C x_{in} - \mu(x_{in}^T y_{in} + x_{out}^T y_{out} - 1) - \nu(x_{in}^T 1_{in} + x_{out}^T 1_{out})].
\] (34)

Differentiating (34) with respect to \( x_{in} \) and \( x_{out} \) respectively, and substituting \( y \) to the equations,

\[
2(L_{in} + D_{in}) y_{in} - 2C y_{in} - 2\mu y_{in} - \nu 1_{in} = 0;
\] (35)

\[
2(L_{out} + D_{out}) y_{out} - 2C y_{out} - 2\mu y_{out} - \nu 1_{out} = 0.
\] (36)

Multiplying \( 1_{in}^T \) to (35) and \( 1_{out}^T \) to (36) from the left hand side, we have

\[
21_{in}^T D_{in} y_{in} - 21_{in}^T C x_{in} - 2\mu 1_{in}^T y_{in} - \nu 1_{in} = 0;
\] (37)

\[
21_{out}^T D_{out} y_{out} - 21_{out}^T C x_{out} - 2\mu 1_{out}^T y_{out} - \nu 1_{out} = 0.
\] (38)

Since \( 1_{in}^T D_{in} = 1_{out}^T C^T \) and \( 1_{in}^T C = 1_{out}^T D_{out} \), summing (37) and (38) we obtain for the Lagrange multiplier \( \nu \):

\[
\nu = -\frac{2\mu}{n} (y_{in}^T 1_{in} + y_{out}^T 1_{out}) = 0
\] (39)
due the fact that \( y \perp 1 \). Applying \( \nu = 0 \) and left multiplying (35) by \( y_{in}^T \) and (36) by \( y_{out}^T \), we have

\[
y_{in}^T (L_{in} + D_{in}) y_{in} - y_{in}^T C y_{out} - \mu y_{in}^T y_{in} = 0;
\] (40)

\[
y_{out}^T (L_{out} + D_{out}) y_{out} - y_{out}^T C y_{in} - \mu y_{out}^T y_{out} = 0.
\] (41)

Since \( y \) is the Fiedler vector, summing (40) and (41) together we obtain

\[
\mu = \lambda_2(L).
\] (42)

Let \( \mathcal{C} = p_{out} 1_{in} 1_{out}^T \), where its entry is the mean of an i.i.d Bernoulli random variable of an entry in \( C \). Let \( \sigma_i(M) \) denote the \( i \)-th largest singular value of \( M \) and write \( C = \mathcal{C} + \mathcal{C} - \mathcal{C} =: \mathcal{C} + \Delta \). Since \( \Delta_{ij} = 1 - p_{out} \) with probability \( p_{out} \) and \( \Delta_{ij} = -p_{out} \) with probability \( 1 - p_{out} \), by Latala’s theorem [46], \( \mathbb{E} \sigma_i(\Delta / \sqrt{n_{in} n_{out}}) \to 0 \) as \( n_{in}, n_{out} \to \infty \). Therefore by
Talagrand’s concentration inequality \[47\], the singular values of \( C/\sqrt{n_{in}n_{out}} \) converge to \( p_{out} \); i.e., \( \sigma_1(C/\sqrt{n_{in}n_{out}}) \xrightarrow{a.s.} p_{out} \), and \( \sigma_i(C/\sqrt{n_{in}n_{out}}) \xrightarrow{a.s.} 0 \) for \( i \geq 2 \) when \( n_{in} \to \infty \) and \( n_{out} \to \infty \). Note that \( n_{in} \) and \( n_{out} \) have to grow with a constant rate so that the entries in \( C \) are bounded when \( n \to \infty \). Furthermore, as proved in \[48\], the left/right singular vectors of \( C \) and \( \overline{C} \) close to each other in the sense that the squared inner product of their left/right singular vectors converges to 1 almost surely when \( n_{in}n_{out} \to \infty \). Consequently, we have \( \frac{1}{n_{in}} D_{in} 1_{in} = \frac{1}{n_{out}} D_{out} 1_{in} = \frac{1}{n_{out}} \sqrt{n_{out}} \to \frac{1}{n_{in}} \sqrt{n_{in}} \), \( n_{in} \) and \( n_{out} \) are not constant vectors. Therefore \[52\] implies \( y_{in} \) and \( y_{out} \) are both constant vectors. Finally, by the constraints \( y_{in}^T 1_{in} + y_{out}^T 1_{out} = 0 \) and \( 1_{in}^T y_{in} + 1_{out}^T y_{out} = 1 \), we have when \( p_{out} \leq p_{out}^* \) almost surely,

\[
\sqrt{n_{in}} y_{in} \xrightarrow{a.s.} \pm 1_{in}, \quad \sqrt{n_{out}} y_{out} \xrightarrow{a.s.} \pm 1_{out},
\]

(54)

By the PSD property of graph Laplacian matrix, \( y_{in}^T L_{in} y_{in} > 0 \) and \( y_{out}^T L_{out} y_{out} > 0 \) if and only if \( y_{in} \) and \( y_{out} \) are not constant vectors.

**References**

[1] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proc. National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[2] A. Bertrand and M. Moonen, “Seeing the bigger picture: How nodes can learn their place within a complex ad hoc network topology,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 71–82, 2013.

[3] A. Lancichinetti and S. Fortunato, “Community detection algorithms: A comparative analysis,” *Phys. Rev. E*, vol. 80, p. 056117, Nov 2009.

[4] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[5] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[6] U. Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[7] T. A. Snijders and R. Nowicki, “Estimation and prediction for stochastic blockmodels for graphs with latent block structure,” *Journal of Classification*, vol. 14, no. 1, pp. 75–100, 1997.

[8] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Phys. Rev. E*, vol. 83, p. 016107, Jan 2011.

[9] M. E. J. Newman, “Modularity and community structure in networks,” *Proc. National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[10] M. Coscia, F. Giannotti, and D. Pedreschi, “A classification for community discovery methods in complex networks,” *Statistical Analysis and Data Mining*, vol. 4, no. 5, 2011.

[11] H. Wen, E. A. Leicht, and R. M. D’Souza, “Improving community detection in networks by targeted node removal,” *Phys. Rev. E*, vol. 83, p. 016114, Jan 2011.

[12] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[13] M. Newman, *Networks: An Introduction*. Oxford University Press, Inc., 2010.

[14] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[15] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[16] M. Fiedler, “Algebraic connectivity of graphs,” *Czechoslovak Mathematical Journal*, vol. 23, no. 98, pp. 298–305, 1973.

[17] P. W. Holland, K. B. Laskey, and S. Leinhardt, “Stochastic blockmodels: First steps,” *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[18] B. Miller, N. Bliss, and P. Wolfe, “Toward signal processing theory for graphs and non-Euclidean data,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010, pp. 5414–5417.

[19] B. Miller, N. Bliss, and P. J. Wolfe, “Subgraph detection using eigenvector L1 norms,” in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 1633–1641.

[20] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E*, vol. 74, p. 036104, Sep 2006.

[21] R. Merris, “Laplacian matrices of graphs: a survey,” *Linear Algebra and its Applications*, vol. 197-198, pp. 143–176, 1994.

[22] N. M. de Abreu, “Old and new results on algebraic connectivity of graphs,” *Linear Algebra and its Applications*, vol. 423, pp. 53–73, 2007.

[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[24] Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, vol. 40, pp. 35–41, 1977.

[25] S. Sabidussi, “The centrality index of a graph,” *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.
[26] M. Everett and S. P. Borgatti, “Ego network betweenness,” Social Networks, vol. 27, no. 1, pp. 31–38, 2005.

[27] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, “Defining and identifying communities in networks,” Proc. National Academy of Sciences, vol. 101, no. 9, pp. 2658–2663, 2004.

[28] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Phys. Rev. E, vol. 69, p. 026113, Feb 2004.

[29] M. E. J. Newman, “Spectral methods for community detection and graph partitioning,” Phys. Rev. E, vol. 88, p. 042822, Oct 2013.

[30] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed membership stochastic blockmodels,” J. Mach. Learn. Res., vol. 9, pp. 1981–2014, June 2008.

[31] S. Fortunato and M. Barthelemy, “Resolution limit in community detection,” Proc. National Academy of Sciences, vol. 104, no. 1, pp. 36–41, 2007.

[32] B. H. Good, Y.-A. de Montjoye, and A. Clauset, “Performance of modularity maximization in practical contexts,” Phys. Rev. E, vol. 81, p. 046106, Apr 2010.

[33] R. R. Nadakuditi and M. E. J. Newman, “Graph spectra and the detectability of community structure in networks,” Phys. Rev. Lett., vol. 108, p. 188701, May 2012.

[34] A. Pothen, H. D. Simon, and K.-P. Liou, “Partitioning sparse matrices with eigenvectors of graphs,” SIAM J. Matrix Anal. Appl., vol. 11, no. 3, pp. 430–452, May 1990.

[35] D. A. Spielman and S.-H. Teng, “Spectral partitioning works: Planar graphs and finite element meshes,” Linear Algebra and its Applications, vol. 421, no. 2-3, pp. 284–305, 2007.

[36] S. E. Schaeffer, “Graph clustering,” Computer Science Review, vol. 1, no. 1, pp. 27–64, 2007.

[37] A. Bertrand and M. Moonen, “Distributed computation of the Fiedler vector with application to topology inference in ad hoc networks,” Signal Processing, vol. 93, no. 5, pp. 1106–1117, 2013.

[38] J.-M. Guo, “A new upper bound for the laplacian spectral radius of graphs,” Linear Algebra and its Applications, vol. 400, pp. 61–66, 2005.

[39] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions I,” Mathematical Programming, vol. 14, pp. 265–294, 1978.

[40] S. Fujishige, Submodular Functions and Optimization. Annals of Discrete Math., North Holland, 1990.

[41] N. Alon, M. Krivelevich, and B. Sudakov, “Finding a large hidden clique in a random graph,” Random Structures and Algorithms, vol. 31, no. 4, pp. 5–20, 2007.

[42] R. Nadakuditi, “On hard limits of eigen-analysis based planted clique detection,” in IEEE Statistical Signal Processing Workshop (SSP), Aug 2012, pp. 129–132.

[43] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson, “The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations,” Behavioral Ecology and Sociobiology, vol. 54, no. 4, pp. 396–405, 2003.

[44] W. Kruithof, “An information flow model for conflict and fission in small groups,” Journal of Anthropological Research, vol. 33, no. 4, pp. 452–473, 1977.

[45] I. Cantador, P. Brusilovsky, and T. Kuflik, “2nd workshop on information heterogeneity and fusion in recommender systems (HetRec),” in Proc. ACM conference on Recommender systems, 2011.

[46] R. Lalala, “Some estimates of norms of random matrices,” Proc. Am. Math. Soc., vol. 133, no. 5, pp. 1273–1282, 2005.

[47] M. Talagrand, “Concentration of measure and isoperimetric inequalities in product spaces,” Publications Mathmatiques de l’Institut des Hautes tudes Scientifiques, vol. 81, no. 1, pp. 73–205, 1995.

[48] F. Benaych-Georges and R. R. Nadakuditi, “The singular values and vectors of low rank perturbations of large rectangular random matrices,” Journal of Multivariate Analysis, vol. 111, no. 0, pp. 120–135, 2012.

[49] F. Radicchi and A. Arenas, “Abrupt transition in the structural formation of interconnected networks,” Nature Physics, vol. 9, no. 11, pp. 717–720, Nov. 2013.