Internet Mathematics, 11:333–351, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1542-7951 print/1944-9488 online
DOI: 10.1080/15427951.2014.968295

A LOCAL CLUSTERING ALGORITHM FOR CONNECTION GRAPHS

Fan Chung and Mark Kempton
University of California, San Diego, California, USA

Abstract We give a clustering algorithm for connection graphs, that is, weighted graphs in which each edge is associated with a d-dimensional rotation. The problem of interest is to identify subsets of small Cheeger ratio that have a high level of consistency, i.e., that have a small edge boundary and for which the rotations along any distinct paths joining two vertices are the same or within some small error factor. We use PageRank vectors as well as tools related to the Cheeger constant to give a clustering algorithm that runs in nearly linear time.

1. INTRODUCTION

In this work, we study connection graphs, which are generalizations of weighted graphs in which each edge is associated with both a positive scalar weight and a d-dimensional rotation matrix for some fixed positive integer d. The Laplacian of connection graphs are higher dimensional versions of the normalized Laplacian matrices, which are linear operators acting on the space of vector-valued functions (instead of the usual real-valued functions).

Connection graphs arise in applications involving high-dimensional datasets in which some data points are related by rotation matrices. Some early usage of connection graphs can be traced back to work in graph gauge theory for computing the vibrational spectra of molecules and examining spins associated with vibrations [9]. There have been more recent developments in related research on principal component analysis [13], cryo-electron microscopy [11, 15], angular synchronization of eigenvectors [10, 14], and vector diffusion maps [16]. In computer vision, there has been a great deal of work dealing with the many photos that are available on the web, upon which information networks of photos can be built. The edges of the associated connection graphs correspond to the rotations determined by the angles and positions of the cameras used [1]. Recently, related work has been done on a synchronization problem, for which the connection Laplacian acts on the space of functions that assign an orthogonal matrix to each vertex [4].

For high-dimensional datasets, a central problem is to uncover lower-dimensional structures in spite of possible errors or noises. An approach for reducing the effect of errors is to consider the notion of inconsistency, which quantifies the difference of accumulated rotations while traveling along distinct paths between two vertices. In many applications, it

Address corrrespondence to Mark Kempton, Department of Mathematics, University of California San Diego, La Jolla, CA 92093-0112, USA. E-mail: mkempton@ucsd.edu

is desirable to identify edges causing the inconsistencies, or to identify portions of the graph that have relatively small inconsistency. In [8], an algorithm is given, that utilizes a version of effective resistance from electrical network theory, which deletes edges of a connection graph in such a way that reduces inconsistencies. In this study, rather than deleting edges, our focus is on identifying subsets of a connection graph with small inconsistency. The notion of ϵ-consistency of a subset of the vertex set of a connection graph will be introduced, which quantifies the amount of inconsistency for the subset to within an error ϵ. This can be viewed as a generalization of the notion of consistency.

One of the major problems in computing is to design efficient clustering algorithms for finding a good cut in a graph. That is, it is desirable to identify a subset of the graph with a small edge boundary in comparison to the overall volume of the subset. Many clustering algorithms have been derived, including some with quantitative analysis (e.g., [2, 3]). As we are looking for ϵ-consistent subsets, it is natural that clustering and the Cheeger ratio should arise in examining local subsets of a graph. In this study, we will combine the clustering problem and the problem of identifying ϵ-consistent subsets. In particular, we will give an algorithm that uses PageRank vectors to identify a subset of a connection graph that has a small cut, given that there is a subset with a small cut that is ϵ-consistent.

The notion of PageRank was first introduced by [5] in 1998 for Google's web search algorithms. It has since proven useful in graph theory for quantifying relationships between vertices in a graph. Algorithms from [2] and [3] utilize PageRank vectors to locally identify good cuts in a graph. In [8], a vectorized version of PageRank is given for connection graphs. Here, we use these connection PageRank vectors in a manner similar to that of [3] to find good cuts under the assumption of an ϵ-consistent subset.

1.1. A Summary of the Results

The results in this article can be summarized as follows:

- We define the notion of ϵ-consistency and establish several inequalities relating ϵ-consistency with the smallest eigenvalue of the connection Laplacian and the Cheeger ratio of subsets of a connection graph.
- We define connection PageRank vectors and establish several inequalities relating the sharp drops in the connection PageRank vectors to the Cheeger ratio and the ϵ-consistency of the subsets.
- We give an algorithm that outputs a subset of the vertices (if one exists) that is a good cut and that intersects an ϵ-consistent subset in large way. The runtime of the algorithm is $O(d^2 x \frac{\log^2 m}{\phi^2})$, where m is the number of edges, d is the dimension of the rotations, ϕ is the target Cheeger ratio, and x is the target volume.

The remainder of this article is organized as follows: In Section 2, we give some of the basic definitions of a connection graph, the connection Laplacian, and the notion of consistency, as well as some useful facts on consistency from [8]. In Section 3 we introduce the notion of ϵ-consistency, which generalizes the notion of consistency, and gives some results relating ϵ-consistency of a connection graph to the spectrum of the normalized connection Laplacian. In Section 4 we examine subsets of a connection graph that are ϵ-consistent, and relate the spectrum of the normalized Laplacian to the Cheeger ratio of such subsets. In Section 5, we utilize connection PageRank vectors in the study of ϵ-consistent subsets, and present a local partition algorithm for a connection graph, completed with complexity analysis.

2. PRELIMINARIES

2.1. The Normalized Connection Laplacian

Suppose $G = (V, E, w)$ is an undirected graph with vertex set V, edge set E, and edge weights $w_{uv} = w_{vu} > 0$ for edges (u, v) in E. Let $\mathcal{F}(V, \mathbb{R})$ denote the space of all functions $f : V \to \mathbb{R}$. The usual adjacency matrix A, combinatorial Laplacian matrix L, and normalized Laplacian $\mathcal{L}$, are all operators on the space $\mathcal{F}(V, \mathbb{R})$. (See, for example, [6] for definitions of A, L, and $\mathcal{L}$.) For undefined terminology, the reader is referred to [8].

Now suppose each oriented edge (u, v) is also associated with a rotation matrix $O_{uv} \in \mathsf{SO}(d)$ satisfying $O_{uv}O_{vu} = I_{d\times d}$. Here $\mathsf{SO}(d)$ denotes the special orthogonal group of dimension d, namely, the group of all $d \times d$ matrices S satisfying $S^{-1} = S^T$ and $\det(S) = 1$. Let O denote the set of rotations associated with all oriented edges in G. The *connection graph*, denoted by $\mathbb{G} = (V, E, O, w)$, has G as the *underlying graph.* The *connection adjacency matrix* $\mathbb{A}$ of $\mathbb{G}$ is defined by:

$$\mathbb{A}(u, v) = \begin{cases} w_{uv}O_{uv} & \text{if } (u, v) \in E \\ 0_{d\times d} & \text{if } (u, v) \notin E, \end{cases}$$

where $0_{d\times d}$ is the zero matrix of size $d \times d$. We view $\mathbb{A}$ as a block matrix in which each block is either a $d \times d$ rotation matrix O_{uv} multiplied by a scalar weight w_{uv}, or a $d \times d$ zero matrix. The matrix $\mathbb{A}$ is an operator on the space $\mathcal{F}(V, \mathbb{R}^d) = \{f : V \to \mathbb{R}^d\}$. The matrix $\mathbb{A}$ is symmetric as $O_{uv}^T = O_{vu}$ and $w_{uv} = w_{vu}$.

The *connection Laplacian* $\mathbb{L}$ of a graph $\mathbb{G}$ is defined by

$$\mathbb{L} = \mathbb{D} - \mathbb{A},$$

where $\mathbb{D}$ is the diagonal matrix defined by the diagonal blocks $\mathbb{D}(u, u) = d_u I_{d\times d}$ for $u \in V$. Here, d_u is the weighted degree of u in G, i.e., $d_u = \sum_{(u,v)\in E} w_{uv}$. The connection Laplacian is an operator on $\mathcal{F}(V, \mathbb{R}^d)$, where its action on a function $f : V \to \mathbb{R}^d$ is given by

$$\mathbb{L}f(v) = \sum_{u\sim v} w_{uv}\left(f(v) - f(u)O_{uv}\right).$$

(The elements of $\mathcal{F}(V, \mathbb{R}^d)$ are sometimes viewed as row vectors so that $f(u)O_{uv}$ is the product of matrix multiplication of $f(u)$ and O_{uv}.)

Recall that for any orientation of edges of the underlying graph G on n vertices and m edges, the combinatorial Laplacian L can be written as $L = B^T W B$, where W is an $m \times m$ diagonal matrix with $W_{e,e} = w_e$, and B is the edge-vertex incident matrix of size $m \times n$ such that $B(e, v) = 1$ if v is e's head; $B(e, v) = -1$ if v is e's tail; and $B(e, v) = 0$ otherwise. A useful observation for the connection Laplacian is the fact that it can be written in a similar form. Let $\mathbb{B}$ be the $md \times nd$ block matrix given by

$$\mathbb{B}(e, v) = \begin{cases} O_{uv} & v \text{ is } e\text{'s head,} \\ -I_{d\times d} & v \text{ is } e\text{'s tail,} \\ 0_{d\times d} & \text{otherwise.} \end{cases}$$

Let the block matrix $\mathbb{W}$ denote the diagonal block matrix given by $\mathbb{W}(e, e) = w_e I_{d\times d}$, where $\mathbb{W}$ is actually of size $md \times md$. Then, it can be verified by direct computation that, given an orientation of the edges, the connection Laplacian also can alternatively be defined as

$$\mathbb{L} = \mathbb{B}^T \mathbb{W} \mathbb{B}.$$

We define the *normalized connection Laplacian* $\hat{\mathcal{L}}$ to be the operator on $\mathcal{F}(V, \mathbb{R}^d)$ given by

$$\hat{\mathcal{L}} = \mathbb{D}^{-1/2} \mathbb{L} \mathbb{D}^{-1/2} = I_{nd \times nd} - \mathbb{D}^{-1/2} \mathbb{A} \mathbb{D}^{-1/2}.$$

We remark that $\mathbb{L}$ and $\hat{\mathcal{L}}$ are symmetric, positive semidefinite matrices. Using the Courant–Fischer theorem (see, for example, [12]), we can investigate the eigenvalues of $\hat{\mathcal{L}}$ by examining the *Rayleigh quotient*

$$\mathcal{R}(g) = \frac{g \hat{\mathcal{L}} g^T}{g g^T},$$

where $g : V \to \mathbb{R}^d$ is thought of as a $1 \times nd$ row vector. Defining $f = g \mathbb{D}^{-1/2}$, we see that

$$\mathcal{R}(g) = \frac{f \mathbb{L} f^T}{f \mathbb{D} f^T} = \frac{\sum_{(u,v) \in E} w_{uv} \, \| f(u) O_{uv} - f(v) \|_2^2}{\sum_{v \in V} d_v \, \| f(v) \|_2^2}.$$

It is not hard to see that $\mathcal{R}(f) \leq 2$. In particular, letting $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{nd}$ denote the eigenvalues of $\hat{\mathcal{L}}$, we see that $\lambda_k \leq 2$ for all k.

2.2. Consistency

For a connection graph $\mathbb{G} = (V, E, O, w)$, we say that $\mathbb{G}$ is *consistent* if

$$\inf_{\substack{f : V \to \mathbb{R}^d \\ \|f\|_2 = 1}} \sum_{(u,v) \in E} w_{uv} \| f(u) O_{uv} - f(v) \|_2^2 = 0.$$

An equivalent definition for consistency is that there exists a function $f : V \to \mathbb{R}^d$ assigning a vector $f(u) \in \mathbb{R}^d$ to each vertex $u \in V$ such that for all edges $uv \in E$, $f(v) = f(u) O_{uv}$. Therefore, for any two vertices u, v in a consistent graph, any two distinct paths starting and ending at u and v, $P_1 = (u = u_1, u_2, \ldots, u_k = v)$ and $P_2 = (u = v_1, v_2, \ldots, v_l = v)$, then the product of rotations along either path is the same. That is,

$$\prod_{i=1}^{k-1} O_{u_i u_{i+1}} = \prod_{j=1}^{l-1} O_{v_j v_{j+1}}.$$

For any cycle $C = (v_1, v_2, \ldots, v_k, v_{k+1} = v_1)$ of the underlying graph, the product of rotations along the cycle C is the identity, i.e., $\prod_{i=1}^{k} O_{v_i v_{i+1}} = I_{d \times d}$.

For ease of notation, given a cycle $C = (v_1, v_2, ..., v_k, v_{k+1} = v_1)$, define $O_C = \prod_{i=1}^{k} O_{v_i v_{i+1}}$, and for a path joining distinct vertices u and v, $P_{uv} = (u = v_1, v_2, ..., v_k = v)$, define $O_{P_{uv}} = \prod_{i=1}^{k-1} O_{v_i v_{i+1}}$. Therefore, consistency can be characterized by saying $O_C = I_{d \times d}$ for any cycle C, or given any two vertices u and v of $\mathbb{G}$, then $O_{P_{uv}} = O_{P'_{uv}}$ for any two paths P_{uv}, P'_{uv} connecting u and v.

In [8], a spectral characterization of consistency for a connection graph is given in terms of the eigenvalues of the connection Laplacian $\mathbb{L}$. We note that an easy modification of the argument in [8] yields the similar statements for the normalized connection Laplacian. Namely, let $\hat{\mathcal{L}}$ be the normalized connection Laplacian of the connection graph $\mathbb{G}$, let $\mathcal{L}$ be the normalized Laplacian of the underlying graph G. For a connected connection graph $\mathbb{G}$, the following statements are equivalent:

(i) $\mathbb{G}$ is consistent.

(ii) The normalized connection Laplacian $\hat{\mathcal{L}}$ of $\mathbb{G}$ has eigenvalue 0.

(iii) The eigenvalues of $\hat{\mathcal{L}}$ are the n eigenvalues of $\mathcal{L}$, each of multiplicity d.

(iv) For each vertex u in G, we can find $O_u \in \mathsf{SO}(d)$ such that for any edge (u, v) with rotation O_{uv}, we have $O_{uv} = O_u^{-1} O_v$.

2.3. The Cheeger Ratio

Given a subset of the vertex set $S \subset V$ we define $E(S, \bar{S})$ to be the set of all edges having one endpoint in S and the other endpoint outside of S. We define the volume of S, denoted vol(S), by $\text{vol}(S) = \sum_{v \in S} d_v$. We define the *Cheeger ratio* of S, denoted $h_G(S)$, by

$$h_G(S) = \frac{|E(S, \bar{S})|}{\text{vol}(S)}.$$

The *Cheeger constant* (sometimes called the *conductance*) of a graph G is

$$h_G = \min \left\{ h(S) : S \subset V,\ \text{vol}(S) \leq \frac{1}{2} \text{vol}(G) \right\}.$$

Determining the Cheeger constant of a graph can be thought of as a discrete version of the classical isoperimetric problem from geometry. One of the classic results in spectral graph theory (see, for example, [6]) is the *Cheeger Inequality*, which relates the Cheeger constant of a graph to the eigenvalues of its normalized Laplacian. Given a graph G with normalized Laplacian $\mathcal{L}$ with eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, the Cheeger inequality states that

$$\frac{h_G^2}{2} \leq \lambda_2 \leq 2h_G.$$

We will be giving results analogous to the Cheeger inequality for ϵ-consistent connection graphs, and the Cheeger ratio will play a critical role in our algorithm and its analysis in Section 5.

3. ϵ-CONSISTENCY

We say a connection graph $\mathbb{G}$ is *ϵ-consistent* if, for every simple cycle $C = (v_1, v_2, ..., v_k, v_{k+1} = v_1)$ of the underlying graph G, we have $\|O_C - I_{d\times d}\|_2 \leq \epsilon$ where $O_C = \prod_{i=1}^{k} O_{v_i v_{i+1}}$. That is, the product of rotations along any cycle is within ϵ of the identity in the 2-norm. An equivalent formulation is as follows. Given vertices u and v, and two distinct paths from u to v, $P_1 = (v_1 = u, v_2, ..., v_k = v)$ and $P_2 = (u_1 = u, u_2, ..., u_l = v)$, define $O_{P_1} = \prod_{i=1}^{k-1} O_{v_i v_{i+1}}$ and $O_{P_2} = \prod_{i=1}^{l-1} O_{u_i u_{i+1}}$. Then $\mathbb{G}$ is ϵ-consistent if and only if $\left\| O_{P_1} - O_{P_2} \right\|_2 \leq \epsilon$. This follows from the observation that $O_C = O_{P_1} O_{P_2}^{-1} = O_{P_1} O_{P_2}^T$ and the fact that the 2-norm of a rotation matrix is 1. For ease of notation, we will simply use $\| \cdot \|$ to denote the ℓ_2 norm $\| \cdot \|_2$.

We observe that the triangle inequality implies that any connection graph is 2-consistent, and that a consistent connection graph is 0-consistent. We generalize the first part of the above mentioned result from [8] with the following theorem, which bounds the d smallest eigenvalues of the normalized connection Laplacian for an ϵ-consistent connection graph.

Theorem 3.1. *Let $\mathbb{G}$ be an ϵ-consistent connection graph whose underlying graph is connected. Let $\hat{\mathcal{L}}$ be the normalized connection Laplacian and let $0 \le \lambda_1 \le \cdots \le \lambda_{nd}$ be the eigenvalues of $\hat{\mathcal{L}}$. Then, for $i = 1, ..., d$,*

$$\lambda_i \le \frac{\epsilon^2}{2}.$$

Proof. We will define a function $f : V \to \mathbb{R}^d$ whose Rayleigh quotient will bound the smallest eigenvalue. For a fixed vertex $z \in V$, we assign $f(z) = x$, where x is a unit vector in $\mathbb{R}^d$. Fix a spanning tree T of G, and define f to be consistent with T. That is, for any vertex v of G assign $f(v)$ as follows. Let $P_{zv} = (z = v_1 v_2 \cdots v_k = v)$ be the path from z to v in T. Then let $f(v) = f(z)O_{P_{zv}}$. Notice that $\|f(v)\| = 1$ for all $v \in V$. We will examine the Rayleigh quotient of this function. Notice that for uv an edge of T, we have

$$\|f(u)O_{uv} - f(v)\| = \|f(v) - f(v)\| = 0$$

by construction. For any other edge uv of G, consider the cycle obtained by taking the path $P_{vu} = (v = v_1 v_2 ... v_k = u)$ in T, and adding in the edge uv. Then, by construction of f and the ϵ-consistency condition, we have

$$\begin{aligned}\|f(u)O_{uv} - f(v)\| &= \left\|f(v)O_{P_{vu}}O_{uv} - f(v)\right\| \\ &= \left\|f(v)\left(\textstyle\prod_{i=1}^{k-1} O_{v_i v_{i+1}} O_{v_k v_1} - I\right)\right\| \\ &\le \epsilon \, \|f(v)\| = \epsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}\lambda_1 \le \mathcal{R}(f) &= \frac{\sum_{(u,v)\in E} w_{uv}||f(u)O_{uv} - f(v)||^2}{\sum_v d_v||f(v)||^2} \\ &\le \frac{\sum_{(u,v)\in E} w_{uv}\epsilon^2}{\sum_v d_v} = \frac{\epsilon^2}{2}.\end{aligned}$$

The initial choice of the unit vector $x \in \mathbb{R}^d$ in the construction of f was arbitrary. We thus have d orthogonal choices for the initial assignment of x, which leads to d orthogonal functions satisfying this inequality. Therefore, by the Courant–Fischer theorem, $\lambda_1, \cdots, \lambda_d$ all satisfy this bound. □

The following result concerns the second block of d eigenvalues of $\hat{\mathcal{L}}$ for an ϵ-consistent connection graph and gives an analog to the upper bound in the Cheeger inequality.

Theorem 3.2. *Let $\hat{\mathcal{L}}$ be the normalized connection Laplacian of the ϵ-consistent connection graph $\mathbb{G}$, with eigenvalues $\lambda_1 \le \cdots \le \lambda_{nd}$, and let h_G denote the Cheeger constant of* *the underlying graph. Then for $i = d + 1, ..., 2d$,*

$$\lambda_i \leq 2h_G + \frac{\epsilon^2}{2}.$$

Proof. Let $f_1, \cdots, f_d$ be the orthogonal set of vectors defined in the proof of Theorem 3.1, each with $\mathcal{R}(f) \leq \epsilon^2/2$. Then, $\|f(v)\|^2 = 1$ for all v. Given $A \subset V$ and $B = \bar{A}$, define $g_i : V \to \mathbb{R}^d$ by

$$g_i(v) = \begin{cases} \dfrac{1}{\operatorname{vol} A} f_i(v) & \text{for} \quad v \in A \\ -\dfrac{1}{\operatorname{vol} B} f_i(v) & \text{for} \quad v \in B. \end{cases}$$

For ease of notation we will simply write g and f for g_i and f_i. Note that if both $u, v \in A$, then $\|g(u)O_{uv} - g(v)\|^2 = \|\frac{1}{\operatorname{vol} A} f(u)O_{uv} - \frac{1}{\operatorname{vol} A} f(v)\|^2 \leq \frac{1}{(\operatorname{vol} A)^2}\epsilon^2$. Similarly, if both $u, v \in B$, $\|g(u)O_{uv} - g(v)\|^2 \leq \frac{1}{(\operatorname{vol} B)^2}\epsilon^2$. For $u \in A$ and $v \in B$, we have $\|g(u)O_{uv} - g(v)\|^2 = \|\frac{1}{\operatorname{vol} A} f(u)O_{uv} + \frac{1}{\operatorname{vol} B} f(v)\|^2 \leq (\frac{1}{\operatorname{vol} A} + \frac{1}{\operatorname{vol} B})^2$ by the triangle inequality.

Therefore,

$$\begin{aligned} \mathcal{R}(g) &= \frac{\sum_{(u,v)\in E} w_{uv}\|g(u)O_{uv} - g(v)\|^2}{\sum_{v\in V} \|g(v)\|^2 d_v} \\ &\leq \frac{\frac{1}{2}\operatorname{vol} A \frac{1}{(\operatorname{vol} A)^2}\epsilon^2 + \frac{1}{2}\operatorname{vol} B \frac{1}{(\operatorname{vol} B)^2}\epsilon^2 + \left(\frac{1}{\operatorname{vol} A} + \frac{1}{\operatorname{vol} B}\right)^2 |E(A,B)|}{\sum_{v\in A} \frac{1}{(\operatorname{vol} A)^2} d_v + \sum_{v\in B} \frac{1}{(\operatorname{vol} B)^2} d_v} \\ &= \frac{\frac{1}{2}\epsilon^2 \left(\frac{1}{\operatorname{vol} A} + \frac{1}{\operatorname{vol} B}\right) + \left(\frac{1}{\operatorname{vol} A} + \frac{1}{\operatorname{vol} B}\right)^2 |E(A,B)|}{\frac{1}{\operatorname{vol} A} + \frac{1}{\operatorname{vol} B}} \\ &\leq \frac{1}{2}\epsilon^2 + 2h_G(A). \end{aligned}$$

Therefore, we have d orthogonal vectors $g_1, ..., g_d$ satisfying this bound, each orthogonal to $f_1, ..., f_d$, which clearly satisfy the bound, so the result follows. □

We remark that the work of [4] gives a different but related notion of "almost consistent" for a connection graph, which they call the *frustration constant*, denoted $\eta_{\mathbb{G}}$, defined by

$$\eta_{\mathbb{G}} = \min_{f:V\to\mathbb{S}^{d-1}} \frac{\sum_{(u,v)\in E} w_{uv}||f(u)O_{uv} - f(v)||^2}{\sum_v d_v||f(v)||^2},$$

where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. So the frustration constant restricts only functions whose entries have norm 1, and as remarked in [4], computation of $\lambda_1(\hat{\mathcal{L}})$ is a relaxation of the computation of $\eta_{\mathbb{G}}$. The proof of Theorem 3.1 utilized only functions $f : V \to \mathbb{R}^d$ whose entries have norm 1, so the proof shows that if $\mathbb{G}$ is an ϵ-consistent connection graph, then

$$\eta_{\mathbb{G}} \leq \frac{\epsilon^2}{2}.$$

4. CONSISTENT AND ϵ-CONSISTENT SUBSETS

In this section, we will consider the case where a connection graph has been created in which some subset of the data is errorfree (or close to it), leading to a consistent or ϵ-consistent induced subgraph. We will define functions on the vertex set in such a way that the Rayleigh quotient will keep track of the edges leaving the consistent subset. In this way, we will obtain bounds on the spectrum of the normalized connection Laplacian involving the Cheeger ratio of such subsets.

Theorem 4.1. *Let $\mathbb{G}$ be a connection graph of dimension d with normalized connection Laplacian $\hat{\mathcal{L}}$, and $S \subset V$ a subset of the vertex set that is ϵ-consistent for given $\epsilon \geq 0$. Then for $i = 1, ..., d$,*

$$\lambda_i(\hat{\mathcal{L}}) \leq \frac{\epsilon^2}{2} + h_G(S).$$

Proof. Fix a spanning tree T of the subgraph induced by S. Define f as follows. For a fixed vertex u of S, define $f(u) = x$ where $||x|| = 1$, and for $v \in S$, define f to be consistent with the subtree T. For $v \notin S$, define $f(v) = 0$. Fix an edge $uv \in E$ and note that for $u, v \notin S$, $||f(u)O_{uv} - f(v)|| = 0$, for $u, v \in S$, $||f(u)O_{uv} - f(v)|| = ||f(v)\left(O_{P_{vu}}O_{uv} - I\right)|| < \epsilon$, and for $u \in S$, $v \notin S$, $||f(u)O_{uv} - f(v)|| = 1$. Therefore,

$$\begin{aligned}\mathcal{R}(f) &= \frac{\sum_{(u,v)\in E} w_{uv}||f(u)O_{uv} - f(v)||^2}{\sum_v d_v||f(v)||^2}\\ &< \frac{\sum_{\substack{uv\in E\\ u,v\in S}} w_{uv}\epsilon^2}{\text{vol}(S)} + \frac{\sum_{\substack{uv\in E\\ u\in S, v\notin S}} w_{uv}}{\text{vol}(S)}\\ &\leq \tfrac{\epsilon^2}{2} + h_G(S).\end{aligned}$$

There are d orthogonal choices for the initial choice of x leading to d orthogonal vectors satisfying this bound, so by the Courant–Fischer theorem, the result follows. □

In the next result, we consider the situation when most of the edges are close to being consistent except for some edges in the edge boundary of a subset.

Theorem 4.2. *Suppose G is an ϵ_1-consistent graph for some $\epsilon_1 > 0$, and suppose that $S \subset V$ is a set such that the subgraphs induced by S and $\bar{S}$ are both ϵ_2-consistent, with $0 \leq \epsilon_2 < \epsilon_1$, and* $\text{vol}(S) \leq \frac{1}{2}\text{vol}(G)$. *Let $\hat{\mathcal{L}}$ be the normalized connection Laplacian. Then for $i = 1, ..., d$,*

$$\lambda_i(\hat{\mathcal{L}}) < \frac{\epsilon_2^2}{2} + \frac{\epsilon_1^2}{2}h_G(S).$$

Proof. We will construct a function $f : V \to \mathbb{R}^d$ whose Rayleigh quotient will bound λ_1. Fix a spanning tree T of S and T' of $\bar{S}$, and fix a vertex $w \in S$. Choose a unit vector $x \in \mathbb{R}^d$, and assign $f(w) = x$. For $v \in S$, assign $f(v)$ for each vertex $v \in S$ such that $f(v) = f(u)O_{uv}$ moving along edges uv of T. Now choose an arbitrary edge $e = yz \in E(S, \bar{S})$ such that $y \in S$ and $z \in \bar{S}$. Assign $f(z) = f(y)O_{yz}$. Assign the remaining vertices of $\bar{S}$ so that $f(v) = f(u)O_{uv}$ moving along edges uv of T'. Note that f is consistent with both T and T'.

Let us examine the Dirichlet sum $\sum_{uv \in E} w_{uv}||f(u)O_{uv} - f(v)||^2$. Consider an edge $f = uv \in E(S, \bar{S})$, $f \neq e$. We may, without loss of generality, assume that both S and $\bar{S}$ are connected. (If one or both is not, then we may alter our definition of f to be consistent along even more edges). Therefore, there is a cycle, $C = v_1 v_2 ... v_k v_1$, where $v_1 = u$, $v_k = v$, C contains the edges e and f, and all other edges have endpoints lying in either S or $\bar{S}$. By construction, $f(v) = f(u)O_{P_{uv}}$, so by the ϵ-consistency condition, we have

$$\begin{aligned} ||f(u)O_{uv} - f(v)|| &= \left\| f(v)O_{P_{vu}}O_{uv} - f(v) \right\| \\ &= \left\| f(v)\left(\prod_{i=1}^{k-1} O_{v_i v_{i+1}} O_{v_k v_1} - I\right) \right\| \\ &\leq \epsilon_1 ||f(v)|| = \epsilon_1. \end{aligned}$$

In a similar manner, we have that $||f(u)O_{uv} - f(v)|| \leq \epsilon_2$ for each edge uv with both u and v in S or both in $\bar{S}$.

Therefore,

$$\begin{aligned} \lambda_1 \leq \mathcal{R}(f) &= \frac{\sum_{(u,v)\in E} w_{uv}||f(u)O_{uv} - f(v)||^2}{\sum_v d_v ||f(v)||^2} \\ &\leq \frac{\sum_{(u,v)\in E} w_{uv}\epsilon_2^2}{\sum_v d_v} + \frac{\sum_{\substack{u\sim v \\ u\in S, v\in \bar{S}}} w_{uv}\epsilon_1^2}{\sum_v d_v} \\ &\leq \frac{\epsilon_2^2 |E(G)|}{\mathrm{vol}(G)} + \frac{\epsilon_1^2 |E(S,\bar{S})|}{2\,\mathrm{vol}(S)} \\ &= \frac{\epsilon_2^2}{2} + \frac{\epsilon_1^2}{2} h_G(S). \end{aligned}$$

We have d orthogonal choices for the initial assignment of x, which leads to d orthogonal vectors satisfying this inequality. Therefore, $\lambda_1, ..., \lambda_d$ all satisfy this bound. □

Our next result is similar to Theorem 3.2, but in a setting similar to the previous theorem.

Theorem 4.3. *Let $\mathbb{G}$ be a connection graph, and suppose $S \subset V$ is a set such that the subgraphs induced by S and $\bar{S}$ are ϵ-consistent, with $\mathrm{vol}(S) \leq \frac{1}{2}\mathrm{vol}(G)$. Let $\hat{\mathcal{L}}$ be the normalized connection Laplacian with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_{nd}$. Then, for $i = d+1, ..., 2d$,*

$$\lambda_i \leq \frac{\epsilon^2}{2} + 2h_G(S).$$

Proof. Let $f_1, ... f_d$ be d orthogonal vectors defined as in the proof of the preceding theorem. Each of these has $\mathcal{R}(f_i) \le \frac{\epsilon^2}{2} + 2h_G(S)$ and $||f(v)||^2 = 1$ for all v. Define $g_i : V \to \mathbb{R}^d$ by

$$g_i(v) = \begin{cases} \dfrac{1}{\operatorname{vol} S} f_i(v) & \text{for} \quad v \in S \\ -\dfrac{1}{\operatorname{vol} \bar{S}} f_i(v) & \text{for} \quad v \in \bar{S}. \end{cases}$$

For ease of notation, we will simply write g and f for g_i and f_i. Then,

$$\begin{aligned} \mathcal{R}(g) &= \frac{\sum_{u \sim v} w_{uv} \|g(u)O_{uv} - g(v)\|_2^2}{\sum\limits_{v \in V} \|g(v)\|^2 d_v} \\ &\le \frac{\frac{1}{2}\left(\frac{1}{\operatorname{vol} S} + \frac{1}{\operatorname{vol} \bar{S}}\right)\epsilon^2 + \sum_{\substack{u \sim v \\ u \in S, v \in \bar{S}}} w_{uv} \| \frac{1}{\operatorname{vol} S} f(u)O_{uv} + \frac{1}{\operatorname{vol} \bar{S}} f(v) \|}{\frac{1}{\operatorname{vol} S} + \frac{1}{\operatorname{vol} \bar{S}}} \\ &\le \frac{\epsilon^2}{2} + \left(\frac{1}{\operatorname{vol} S} + \frac{1}{\operatorname{vol} \bar{S}}\right) |E(S, \bar{S})| \le \frac{\epsilon^2}{2} + 2h_G(S). \end{aligned}$$

We have d orthogonal vectors $g_1, \cdots, g_d$ satisfying this bound and observe that each is orthogonal to the vectors $f_1, \cdots, f_d$. Therefore, the result follows.

□

We remark that this theorem is a stronger result than that in Theorem 3.2, because the hypothesis does not require that the full graph be ϵ-consistent. That is, the result still holds even if the edges going from S to $\bar{S}$ involve inconsistencies that cause the full graph to fail to be ϵ-consistent.

5. IDENTIFYING SUBSETS

In this section, we follow ideas from [2] and [3] to relate connection PageRank vectors to the Cheeger ratio of ϵ-consistent subsets of a connection graph. We will give an algorithm, which runs in time nearly linear in the size of the vertex set, which outputs a subset of the vertex set (if one exists) and which has small Cheeger ratio and is ϵ-consistent.

5.1. PageRank Vectors and ϵ-Consistent Subsets

We define, for $S \subset V$, $f(S) = \sum_{v \in S} ||f(v)||_2$. Given a vertex v of $\mathbb{G}$, define a connection characteristic function χ_v to be any vector satisfying $||\chi_v(v)||_2 = 1$ and $\chi_v(u) = 0$ for $u \ne v$. Likewise, for a subset S of V, define a characteristic function χ_S to be a function such that $||\chi_S(v)||_2 = 1$ for $v \in S$, and $\chi_S(v) = 0$ for $v \notin S$.

Recall the definition of connection PageRank [8]. Given a seed vector $\hat{s} : V \to \mathbb{R}^d$ is the vector $\widehat{\operatorname{pr}}(\alpha, \hat{s}) : V \to \mathbb{R}^d$ that satisfies

$$\widehat{\operatorname{pr}}(\alpha, \hat{s}) = \alpha \hat{s} + (1 - \alpha) \widehat{\operatorname{pr}}(\alpha, \hat{s}) \mathbb{Z},$$

where $\mathbb{Z} = \frac{1}{2}(I + \mathbb{D}^{-1}\mathbb{A})$ is the matrix for the random walk. Define $\mathbb{R}_\alpha = \alpha(I - (1 - \alpha)\mathbb{Z})^{-1} = \alpha \sum_{t=0}^{\infty}(1-\alpha)^t \mathbb{Z}^t$ and note that $\widehat{\mathrm{pr}}(\alpha, \hat{s}) = \hat{s}\mathbb{R}_\alpha$.

Lemma 5.1. *Let $S \subset V$ be a subset of the vertex set of a connection graph, and let χ_S be a characteristic function for S. Then*

$$\|\chi_S \mathbb{D}\mathbb{R}_\alpha(v)\| \le d_v$$

for all $v \in V$.

Proof. First, we will show that

$$\left\|\chi_S \mathbb{D}\mathbb{Z}^k(v)\right\| \le d_v$$

for all k by induction. For $k = 1$,

$$\|\chi_S \mathbb{D}\mathbb{Z}(v)\| = \frac{1}{2}\|\chi_S \mathbb{D}(I + \mathbb{D}^{-1}\mathbb{A})(v)\| \le \frac{1}{2}\left(d_v + \sum_{\substack{u \in S \\ u \sim v}} w_{uv}\|\chi_S(u) O_{uv}\|\right) \le d_v.$$

By the induction hypothesis,

$$\begin{aligned}
\left\|\chi_S \mathbb{D}\mathbb{Z}^{k+1}(v)\right\| = \left\|\chi_S \mathbb{D}\mathbb{Z}^k \mathbb{Z}(v)\right\| &= \left\|\sum_{u \in V} \chi_S \mathbb{D}\mathbb{Z}^k(u)\mathbb{Z}(u, v)\right\| \\
&\le \sum_{u \in V} \left\|\chi_S \mathbb{D}\mathbb{Z}^k(u)\right\|_2 \|\mathbb{Z}(u, v)\| \\
&\le \sum_{u \in V} d_u \frac{1}{2}\|I(u, v) + \mathbb{D}^{-1}\mathbb{A}(u, v)\| \\
&\le \frac{d_v}{2} + \frac{1}{2}\sum_{u \in V} d_u \left\|\frac{1}{d_u} w_{uv} O_{uv}\right\| \\
&= \frac{d_v}{2} + \frac{1}{2}\sum_{u \in V} w_{uv} = d_v,
\end{aligned}$$

so this claim follows by induction.

Then, from this claim,

$$\|\chi_S \mathbb{D}\mathbb{R}_\alpha(v)\| = \left\|\chi_S \mathbb{D}\alpha \sum_{k=0}^{\infty}(1-\alpha)^k \mathbb{Z}^k\right\| \le \alpha \sum_{k=0}^{\infty}(1-\alpha)^k \left\|\chi_S \mathbb{D}\mathbb{Z}^k(v)\right\| \le d_v.$$

□

Lemma 5.2. *Let $S \subset V$ be a subset of the vertices such that the subgraph of $\mathbb{G}$ induced by S is ϵ-consistent. Let χ_S be some connection characteristic function for S that is consistent with some spanning subtree T of S. Define $\hat{f}_S$ by $\hat{f}_S(v) = \frac{d_v}{\mathrm{vol}(S)}\chi_S(v)$. The function $\hat{f}_S$ is the expected value for a characteristic function χ_u when a vertex u is chosen from S at* *random with probability $d_u/\operatorname{vol}(S)$. Then*

$$\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(S) \geq 1 - \frac{1-\alpha}{\alpha}(h(S)+\epsilon).$$

Proof. We have

$$\begin{aligned}
\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(S) &= \sum_{v\in S} ||\,\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(v)|| = \sum_{v\in S} \|\,\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(v)\|\,\|\chi_S(v)\| \\
&\geq \sum_{v\in S} \widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(v)\chi_S(v)^T = \widehat{\mathrm{pr}}(\alpha, \hat{f}_S)\chi_S^T = \hat{f}_S\mathbb{R}_\alpha\chi_S^T \\
&= \hat{f}_S\left(I - \frac{(1-\alpha)(I-\mathbb{Z})}{I-(1-\alpha)\mathbb{Z}}\right)\chi_S^T = 1 - \left(\hat{f}_S\frac{(1-\alpha)(I-\mathbb{Z})}{I-(1-\alpha)\mathbb{Z}}\right)\chi_S^T \\
&= 1 - \left(\frac{(1-\alpha)\chi_S\mathbb{D}}{\alpha\operatorname{vol}(S)}\frac{\alpha I}{I-(1-\alpha)\mathbb{Z}}(I-\mathbb{Z})\right)\chi_S^T \\
&= 1 - \frac{1-\alpha}{\alpha\operatorname{vol}(S)}\left(\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}\frac{(\mathbb{D}-\mathbb{A})}{2}\right)\chi_S^T \\
&= 1 - \frac{1-\alpha}{2\alpha\operatorname{vol}(S)}\sum_{uv\in E} w_{uv}\left(\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(u)O_{uv} - \chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(v)\right) \\
&\quad \times ((\chi_S(u)O_{uv})^T - \chi_S(v)^T).
\end{aligned}$$

Here, the first inequality follows from the Cauchy–Schwarz inequality. Note that χ_S is a characteristic function, so all the terms in the sum corresponding to $u, v \notin S$ are 0, for $v \in S$ and $u \notin S$ we are left with just $\chi_s(v)$, and for $u, v \in S$, since S is ϵ-consistent and χ_S was chosen to be consistent with a spanning subtree of S, then we have $\chi_S(u)O_{uv} - \chi_S(v)$ has norm less than ϵ. Applying this, the Cauchy–Schwarz inequality, and the triangle inequality to the above, we have

$$\begin{aligned}
\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(S) \geq 1 - \frac{1-\alpha}{2\alpha\operatorname{vol}(S)}\Bigg(&\sum_{\substack{u\sim v\\ v\in S, u\in\bar{S}}} w_{uv}\left\|\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(u)O_{uv} - \chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(v)\right\| \\
&+ \sum_{\substack{u\sim v\\ u,v\in S}} w_{uv}\left\|\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(u)O_{uv} - \chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(v)\right\|\,\|\chi_S(u)O_{uv} - \chi_S(v)\|\Bigg) \\
\geq 1 - \frac{1-\alpha}{2\alpha\operatorname{vol}(S)}\Bigg(&\sum_{\substack{u\sim v\\ v\in S, u\in\bar{S}}} w_{uv}\left(||\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(u)O_{uv}|| + ||\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(v)||\right) \\
&+ \sum_{\substack{u\sim v\\ u,v\in S}} w_{uv}\left(||\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(u)O_{uv}|| + ||\chi_S\mathbb{D}\mathbb{R}_\alpha\mathbb{D}^{-1}(v)||\right)\epsilon\Bigg).
\end{aligned}$$

Using Lemma 5.1 we can conclude that

$$\widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(S) \geq 1 - \frac{1-\alpha}{\alpha\operatorname{vol}(S)}(|\partial S| + \epsilon|E(S,S)|) \geq 1 - \frac{1-\alpha}{\alpha}(h(S)+\epsilon).$$

□

Theorem 5.3. *Let $S \subset V$ be a subset of the vertex set such that the subgraph induced by S is ϵ-consistent. Let χ_S be some connection characteristic function for S that is consistent with some spanning subtree T of S. For each vertex $v \in S$, define $\chi_v : V \to \mathbb{R}^d$ by $\chi_v(v) = \chi_S(v)$ and $\chi_v(u) = 0$ for $u \neq v$. Then, for any $\alpha \in (0, 1]$, there is a subset $S_\alpha \subset S$ with volume* $\mathrm{vol}(S_\alpha) \geq \mathrm{vol}(S)/2$ *such that for any vertex $v \in S_\alpha$, the PageRank vector $\widehat{\mathrm{pr}}(\alpha, \chi_v)$ satisfies*

$$\widehat{\mathrm{pr}}(\alpha, \chi_v)(S) \geq 1 - \frac{2(h(S) + \epsilon)}{\alpha}.$$

Proof. Let v be a vertex of S chosen randomly from the distribution given by $\hat{f}_S$ of the previous result. Define the random variable $X = \widehat{\mathrm{pr}}(\alpha, \chi_v)(\bar{S})$ and note that the definition of PageRank and linearity of expectation implies that $E[X] = \widehat{\mathrm{pr}}(\alpha, \hat{f}_S)$. Therefore, by the preceding result,

$$E[X] = \widehat{\mathrm{pr}}(\alpha, \hat{f}_S)(\bar{S}) \leq \frac{1-\alpha}{\alpha\,\mathrm{vol}(S)}(h(S) + \epsilon) \leq \frac{h(S) + \epsilon}{\alpha}.$$

Define

$$S_\alpha = \left\{ v : \widehat{\mathrm{pr}}(\alpha, \chi_v)(S) \geq 1 - \frac{2(h(S) + \epsilon)}{\alpha} \right\}.$$

Then Markov's inequality implies

$$\Pr[v \notin S_\alpha] \leq \Pr[X > 2E[X]] \leq \frac{1}{2}.$$

Therefore, $\Pr[v \in S_\alpha] \geq \frac{1}{2}$, so $\mathrm{vol}(S_\alpha) \geq \frac{1}{2}\mathrm{vol}(S)$. □

5.2. A Local Partitioning Algorithm

We will follow ideas from [3] to produce an analogue of the Sharp Drop Lemma. Given any function $p : V \to \mathbb{R}^d$, define $q^{(p)} : V \to \mathbb{R}^d$ by $q^{(p)}(u) = p(u)/d_u$ for all $u \in V$. Order the vertices such that $\|q^{(p)}(v_1)\| \geq \|q^{(p)}(v_2)\| \geq \cdots \geq \|q^{(p)}(v_n)\|$. Define $S_j = \{v_1, ..., v_j\}$. The following lemma will be the basis of our algorithm.

Lemma 5.4. (**Sharp Drop Lemma**). *Let $v \in V(\mathbb{G})$ and let $p = \widehat{\mathrm{pr}}(\alpha, \chi_v)$ for some $\alpha \in (0, 1)$, let $q = q^{(p)}$ and let $\phi \in (0, 1)$ be a real number. Then, for any index $j \in [1, n]$, S_j either satisfies*

$$h(S_j) < 2\phi,$$

or there exists some index $k > j$ such that

$$\mathrm{vol}(S_k) \geq \mathrm{vol}(S_j)(1 + \phi) \text{ and } \|\mathrm{q(v_k)}\| \geq \|\mathrm{q(v_j)}\| - \frac{2\alpha}{\phi\,\mathrm{vol(S_j)}}.$$

Proof. Let $S \subset V$ be a subset of the vertex set that contains v. We have

$$\begin{aligned}
p\mathbb{Z}(S) &= \sum_{u\in S} \|p\mathbb{Z}(u)\| \\
&= \sum_{u\in S} \left\| \frac{1}{2}p(u) + \frac{1}{2}q\mathbb{A}(u) \right\| \leq \frac{1}{2}\left(\sum_{u\in S}\|p(u)\| + \sum_{u\in S}\left\| \sum_{v\sim u} q(v)O_{uv} \right\| \right) \\
&\leq \frac{1}{2}\left(\sum_{u\in S}\|p(u)\| + \sum_{u\in S}\sum_{v\sim u}\|q(v)\| \right) \\
&= \frac{1}{2}\left(2\sum_{u\in S}\|p(u)\| - \sum_{(u,v)\in E(S,\bar{S})} (\|q(u)\| - \|q(v)\|) \right) \\
&= p(S) - \frac{1}{2}\sum_{(u,v)\in E(S,\bar{S})} (\|q(u)\| - \|q(v)\|)\,.
\end{aligned}$$

Since $p = \widehat{\mathrm{pr}}(\alpha, \chi_v)$, we have that p satisfies $p\mathbb{Z} = \alpha\chi_v + (1-\alpha)p\mathbb{Z}$, therefore,

$$\|p\mathbb{Z}(u)\| = \frac{1}{1-\alpha}\|p(u) - \alpha\chi_v(u)\| \geq \|p(u)\| - \alpha\|\chi_v(u)\|$$

for any u. Therefore,

$$p\mathbb{Z}(S) \geq p(S) - \alpha.$$

Combining these, we see that

$$\sum_{(u,v)\in E(S,\bar{S})} (\|q(u)\| - \|q(v)\|) \leq 2\alpha. \tag{5.1}$$

Now we will consider S_j. If $\mathrm{vol}(S_j)(1+\phi) > \mathrm{vol}(G)$, then

$$|E(S_j, \bar{S}_j)| \leq \mathrm{vol}(\bar{S}_j) \leq \mathrm{vol}(G)\left(1 + \frac{1}{1+\phi}\right) \leq \phi\,\mathrm{vol}(S_j)$$

and the result holds. Assume $\mathrm{vol}(S_j)(1+\phi) \leq \mathrm{vol}(G)$. Then, there exists a unique index $k > j$ such that

$$\mathrm{vol}(S_{k-1}) \leq \mathrm{vol}(S_j)(1+\phi) \leq \mathrm{vol}(S_k).$$

If $e(S_j, \bar{S}_j) < 2\phi\,\mathrm{vol}(S_j)$, then we are done. If $e(S_j, \bar{S}_j) \geq 2\phi\,\mathrm{vol}(S_j)$, then we note that we can also get a lower bound on $e(S_j, \bar{S}_{k-1})$, namely,

$$e(S_j, \bar{S}_{k-1}) \geq e(S_j, \bar{S}_j) - \mathrm{vol}(S_{k-1} \setminus S_j) \geq 2\phi\,\mathrm{vol}(S_j) - \phi\,\mathrm{vol}(S_j) = \phi\,\mathrm{vol}(S_j).$$

Therefore, by Equation (5.1)

$$\begin{aligned}
2\alpha &\geq \sum_{(u,v)\in E(S_j,\bar{S}_j)} (\|q(u)\| - \|q(v)\|) \\
&\geq \sum_{(u,v)\in E(S_j,\bar{S}_{k-1})} (\|q(u)\| - \|q(v)\|) \\
&\geq e(S_j, \bar{S}_{k-1})(\|q(v_j)\| - \|q(v_k)\|) \\
&\geq \phi \operatorname{vol}(S_j)(\|q(v_j)\| - \|q(v_k)\|).
\end{aligned}$$

This implies that $\|q(v_j)\| - \|q(v_k)\| \leq 2\alpha/\phi \operatorname{vol}(S_j)$, and the result follows. □

For our algorithm, we will employ an efficient algorithm for computing an approximate connection PageRank vector called ApproximatePR. The specifics of the algorithm as well as its runtime analysis can be found in [7], and a version for connection graphs is found in [17]. We will state the relevant result from [17] as the following:

Theorem 5.5. *For $\alpha, \epsilon \in (0,1)$ and $v \in V$, the algorithm ApproximatePR(v, α, ϵ) outputs a vector $\hat{p} = \widehat{\mathrm{pr}}(\alpha, \chi_v - \hat{r})$ such that*

$$\frac{\|\hat{r}(v)\|_2}{d_v} \leq \epsilon$$

for all vertices v. The running time of the algorithm is $O\left(\frac{d^2}{\epsilon\alpha}\right)$.

We note that

$$\frac{\|\hat{p}(u)\|}{d_u} \geq \frac{\|\widehat{\mathrm{pr}}(\alpha, \chi_v)(u)\|}{d_u} - \epsilon$$

for all u.

We are now ready to present the algorithm ConnectionPartition that utilizes PageRank vectors to come up with an ϵ-consistent subset of a small Cheeger ratio.

Theorem 5.6. *Suppose $\mathbb{G}$ is a connection graph with a subset C such that $\operatorname{vol}(C) \leq \frac{1}{2}\operatorname{vol}(\mathbb{G})$, and $h(C) \leq \alpha/64\gamma$ with α as chosen in the algorithm. Assume further that C is ϵ-consistent for some $\epsilon < h(C)$. Let $C_\alpha = \{v \in C : \widehat{\mathrm{pr}}(\alpha, \chi_v)(\bar{C}) \leq \frac{2(h(C)+\epsilon)}{\alpha}\}$. Then, for $v \in C_\alpha$, $\phi < 1$, and $x \geq \operatorname{vol}(C)$, the algorithm ConnectionPartition outputs a set S satisfying the following properties:*

1. *$h(S) \leq 2\phi$.*
2. *$\operatorname{vol}(S) \leq (2/3)\operatorname{vol}(\mathbb{G})$.*
3. *$\operatorname{vol}(S \cap C) \geq (3/4)\operatorname{vol}(S)$.*

Proof.

Claim 5.7. *There exist's an index j such that $\|q(v_j)\| \geq \frac{1}{\gamma \operatorname{vol}(S_j)}$.*

```
ConnectionPartition($v, \phi, x$):
The input into the algorithm is a vertex $v \in V$, a target Cheeger ratio $\phi \in (0,1)$,
and a target volume $x \in [0, 2m]$.
1. Set $\gamma = \frac{1}{8} + \sum_{k=1}^{2m} \frac{1}{k}$ where $m$ is the number of edges, $\alpha = \frac{\phi^2}{8\gamma}$, and $\delta = \frac{1}{16\gamma x}$.
2. Compute $p = \mathsf{ApproximatePR}(v, \alpha, \delta)$ (which approximates $\widehat{\mathrm{pr}}(\alpha, \chi_v)$).

   Set $q(u) = p(u)/d_u$ for each $u$ and order the vertices $v_1, ..., v_n$ so that
   $\|q(v_1)\| \geq \|q(v_2)\| \geq \cdots \geq \|q(v_n)\|$ and for each $j \in [1, n]$ define
   $S_j = \{v_1, ..., v_j\}$.
3. Choose a starting index $k_0$ such that $\|q(v_{k_0})\| \geq \frac{1}{\gamma \operatorname{vol}(S_{k_0})}$.
   If no such starting vertex exists, output Fail: No starting vertex.
4. While the algorithm is running:
   (a) If $(1+\phi)\operatorname{vol}(S_{k_i}) > \operatorname{vol}(G)$, output Fail: No cut found.
   (b) Otherwise, let $k_{i+1}$ be the smallest index such that $\operatorname{vol}(S_{k_{i+1}}) \geq (1+\phi)$
       $\operatorname{vol}(S_{k_i})$.
   (c) If $\|q(v_{k_{i+1}})\| \leq \|q(v_{k_i})\| - 2\alpha/\phi \operatorname{vol}(S_{k_i})$, then output $S = S_{k_i}$ and quit.
       Otherwise repeat the loop.
```

Proof. Suppose that $\|q(v_j)\| < \frac{1}{\gamma \operatorname{vol}(S_j)}$ for every index j. Since $v \in C_\alpha$, $\epsilon < h(C)$, and $h(C) \leq \alpha/64\gamma$, then we know that

$$p(C) \geq \widehat{\mathrm{pr}}(\alpha, \chi_v)(C) - \delta \operatorname{vol}(C) \geq 1 - \frac{2(h(C)+\epsilon)}{\alpha} - \frac{1}{16\gamma x}\operatorname{vol}(C) \geq 1 - \frac{1}{16\gamma} - \frac{1}{16\gamma}$$
$$= 1 - \frac{1}{8\gamma},$$

since $x \geq \operatorname{vol}(C)$. However, under our assumption,

$$\begin{aligned} p(C) \leq p(V) = \sum_{i=1}^{n} \|p(v_i)\| &= \sum_{i=1}^{n} \|q(v_i)\| d_{v_i} \\ &< \sum_{i=1}^{n} \frac{d_{v_i}}{\gamma \operatorname{vol}(S_j)} \\ &\leq \frac{1}{\gamma} \sum_{k=1}^{2m} \frac{1}{k}. \end{aligned}$$

Putting these together, we have

$$1 - \frac{1}{8\gamma} < \frac{1}{\gamma} \sum_{k=1}^{2m} \frac{1}{k}.$$

With the choice of $\gamma = \frac{1}{8} + \sum_{k=1}^{2m} \frac{1}{k}$ as in the algorithm, this yields a contradiction. Therefore there exists some index j with $\|q(v_j)\| \geq \frac{1}{\gamma \operatorname{vol}(S_j)}$ and the claim is proved. □

It follows from Claim 5.7, that the algorithm will not fail to find a starting vertex.

Let k_f be the final vertex considered by the algorithm.

Claim 5.8. *If $k_0, \ldots, k_f$ is a sequence of vertices satisfying both*

- $\|q(v_{k_{i+1}})\| \geq \|q(v_{k_i})\| - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_i})}$
- $\operatorname{vol}(S_{k_{i+1}}) \geq (1+\phi)\operatorname{vol}(S_{k_i})$,

then

$$\|q(k_f)\| \geq \|q(k_0)\| - \frac{4\alpha}{\phi^2 \operatorname{vol}(S_{k_0})}.$$

Proof. We note that $\operatorname{vol}(S_{k_{i+1}}) \geq (1+\phi)^i \operatorname{vol}(S_{k_0})$, and so we have

$$\begin{aligned}
\|q(k_f)\| &\geq \|q(k_0)\| - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_0})} - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_1})} - \cdots - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_{f-1}})} \\
&\geq \|q(k_0)\| - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_0})}\left(1 + \frac{1}{1+\phi} + \cdots + \frac{1}{(1+\phi)^{f-1}}\right) \\
&\geq \|q(k_0)\| - \frac{2\alpha}{\phi \operatorname{vol}(S_{k_0})}\frac{1+\phi}{\phi} \\
&\geq \|q(k_0)\| - \frac{4\alpha}{\phi^2 \operatorname{vol}(S_{k_0})}
\end{aligned}$$

and the claim follows. □

Now we will use this claim, the choice of $\alpha = \phi^2/8\gamma$, and the condition on the starting vertex $\|q(k_0)\| \geq 1/\gamma \operatorname{vol}(S_{k_0})$ to obtain a lower bound on $\|q(k_f)\|$,

$$\begin{aligned}
\|q(k_f)\| &\geq \|q(k_0)\| - \frac{4\alpha}{\phi^2 \operatorname{vol}(S_{k_0})} \\
&\geq \frac{1}{\gamma \operatorname{vol}(S_{k_0})} - \frac{1}{2\gamma \operatorname{vol}(S_{k_0})} \\
&\geq \frac{1}{2\gamma \operatorname{vol}(S_{k_0})}.
\end{aligned}$$

As in the proof of Claim 5.7, we have that $p(C) \geq 1 - \frac{1}{8\gamma}$, and therefore, $p(\bar{C}) \leq \frac{1}{8\gamma}$.

Now observe that $\operatorname{vol}(S_{k_f} \cap \bar{C}) \leq \dfrac{p(\bar{C})}{\|q(k_f)\|}$. This follows since

$$\|q(k_f)\| \operatorname{vol}(S_{k_f} \cap \bar{C}) = \sum_{v \in S_{k_f} \cap \bar{C}} \|q(k_f)\| d_v \leq \sum_{v \in S_{k_f} \cap \bar{C}} \|q(v)\| d_v \leq \sum_{v \in \bar{C}} \|p(v)\| = p(\bar{C}).$$

Thus,

$$\begin{aligned}
\operatorname{vol}(S_{k_f} \cap \bar{C}) &\leq \frac{p(\bar{C})}{\|q(k_f)\|} \\
&\leq \frac{2\gamma \operatorname{vol}(S_{k_f})}{8\gamma} \\
&= \frac{1}{4}\operatorname{vol}(S_{k_f}).
\end{aligned}$$

Therefore, $\text{vol}(S_{k_f}) \leq \text{vol}(C)+\text{vol}(S_{k_f}\cap\bar{C}) \leq \text{vol}(C)+\frac{1}{4}\text{vol}(S_{k_f})$, implying that $\text{vol}(S_{k_f}) \leq \frac{4}{3}\text{vol}(C)$. Using that fact that $\text{vol}(C) \leq \frac{1}{2}\text{vol}(G)$,

$$\text{vol}(S_{k_f}) \leq \frac{4}{3}\text{vol}(C) \leq \frac{2}{3}\text{vol}(G) \leq \frac{\text{vol}(G)}{1+\phi}.$$

This last step follows under the assumption that $\phi \leq 1/2$. We can do this without loss of generality since the guarantee on $h(S)$ in the theorem is trivial for $\phi > 1/2$.

The above shows that the algorithm will not experience a failure due to the volume becoming too large, and we have seen that conditions (2) and (3) will be satisfied by the output.

Finally, to show condition (1), we apply the Sharp Drop Lemma. We know that k_f is the smallest index such that $\text{vol}(S_{k_f+1}) \geq (1+\phi)\text{vol}(S_{k_f})$, and $\|q(v_{k_f+1})\| \leq \|q(v_{k_f})\| - 2\alpha/\phi\,\text{vol}(S_{k_i})$. Therefore the Sharp Drop Lemma guarantees that $h(S_{k_f}) < 2\phi$, and the proof is complete.

□

Theorem 5.9. *The running time for the algorithm* ConnectionPartition *is* $O(d^2 x \frac{\log^2 m}{\phi^2})$.

Proof. The running time is dominated by the computation of the PageRank vector. According to Theorem 5.5, the running time for this is $O(\frac{d^2}{\delta\alpha})$. In the algorithm, we have $\alpha = \frac{\phi^2}{8\gamma}$, $\delta = \frac{1}{16\gamma x}$, and $\gamma = \frac{1}{8} + \sum_{k=1}^{2m} \frac{1}{k} = \Theta(\log m)$. Therefore, $\alpha = O(\frac{\phi^2}{\log m})$ and $\delta = O(\frac{1}{x \log m})$. Therefore, the running time is as claimed. □

REFERENCES

1. S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. “Building Rome in a Day.” In *Proceedings of the 12th IEEE International Conference on Computer Vision*, IEEE, 2009. 72–79.
2. R. Andersen, F. Chung and K. Lang. “Using PageRank to Locally Partition a Graph.” *Internet Math.* 4:1 (2007), 35–64.
3. R. Andersen and F. Chung. “Detecting Sharp Drops in PageRank and a Simplified Local Partitioning Algorithm.” In *Theory and Applications of Models of Computation*, pp. 1–12, Proceedings of TAMC 2007, LNCS 4484. Berlin, Heidelberg: Springer, 2007.
4. A. S. Bandeira, A. Singer and D. A. Spielman. “A Cheeger Inequality for the Graph Connection Laplacian,.” Available online (http://arxiv.org/pdf/1204.3873v1.pdf), 2012.
5. S. Brin and L. Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” *Computer Networks and ISDN Systems* 30:1-7 (1998), 107–117.
6. F. Chung. *Spectral Graph Theory*. AMS Publications, 1997.
7. F. Chung and W. Zhao. “A Sharp PageRank Algorithm with Applications to Edge Ranking and Graph Sparsification.” *WAW 2010*, pp. 2–14, LNCS 6516. Springer, 2010.
8. F. Chung, M. Kempton, and W. Zhao. “Ranking and Sparsifying a Connection Graph.” preprint, 2013.
9. F. Chung and S. Sternberg. “Laplacian and Vibrational Spectra for Homogeneous Graphs.” *J. Graph Theory* 16 (1992), 605–627.
10. M. Cucuringu, Y. Lipman, and A. Singer. “Sensor Network Localization by Eigenvector Synchronization over the Euclidean Group.” *ACM Transactions on Sensor Networks* 8:3 (2012), 19.

11. R. Hadani and A. Singer. “Representation Theoretic Patterns in Three Dimensional Cryo-Electron Microscopy I - The Intrinsic Reconstitution Algorithm.” *Annals of Mathematics* 174:2 (2011), 1219–1241.
12. R. Horn and C. Johnson. *Matrix Analysis*. New York, NY: Cambridge University Press, 1985.
13. I. T. Jolliffe. *Principal Component Analysis*, Springer Series in Statistics, 2nd ed. New York, NY: Springer-Verlag, 2002.
14. A. Singer. “Angular Synchronization by Eigenvectors and Semidefinite Programming.” *Applied and Computational Harmonic Analysis* 30:1 (2011), 20–36.
15. A. Singer, Z. Zhao, Y. Shkolnisky, and R. Hadani. “Viewing Angle Classification of Cryo-Electron Microscopy Images Using Eigenvectors.” *SIAM Journal on Imaging Sciences* 4:2 (2011), 723–759.
16. A. Singer and H.-T. Wu. “Vector Diffusion Maps and the Connection Laplacian.” *Communications on Pure and Applied Mathematics* 65:8 (2012), 1067–1144.
17. W. Zhao. “Ranking and Sparsifying Edges of a Graph.” PhD Thesis, University of California, San Diego, CA, 2012.