Incremental Method for Spectral Clustering of Increasing Orders

Pin-Yu Chen
University of Michigan
pinyu@umich.edu

Baichuan Zhang
Indiana University - Purdue University Indianapolis
bz3@umail.iu.edu

Mohammad Al Hasan
Indiana University - Purdue University Indianapolis
alhasan@cs.iupui.edu

Alfred Hero
University of Michigan
hero@umich.edu

ABSTRACT

The smallest eigenvalues and the associated eigenvectors (i.e., eigenpairs) of a graph Laplacian matrix have been widely used for spectral clustering and community detection. However, in real-life applications the number of clusters or communities (say, K) is generally unknown a-priori. Consequently, the majority of the existing methods either choose K heuristically or they repeat the clustering method with different choices of K and accept the best clustering result. The first option, more often, yields suboptimal result, while the second option is computationally expensive. In this work, we propose an incremental method for constructing the eigenspectrum of the graph Laplacian matrix. This method leverages the eigenstructure of graph Laplacian matrix to obtain the K-th eigenpairs of the Laplacian matrix given a collection of all the $K-1$ smallest eigenpairs. Our proposed method adapts the Laplacian matrix such that the batch eigenvalue decomposition problem transforms into an efficient sequential leading eigenpair computation problem. As a practical application, we consider user-guided spectral clustering. Specifically, we demonstrate that users can utilize the proposed incremental method for effective eigenpair computation and determining the desired number of clusters based on multiple clustering metrics.

1. INTRODUCTION

Over the past two decades, the graph Laplacian matrix and its variants have been widely adopted for solving various research tasks, including graph partitioning [23], data clustering [14], community detection [5, 28], consensus in networks [19], dimensionality reduction [2], entity disambiguation [33], link prediction [32], graph signal processing [27], centrality measures for graph connectivity [4], interconnected physical systems [24], network vulnerability assessment [7], image segmentation [26], among others. The fundamental task is to represent the data of interest as a graph for analysis, where a node represents an entity (e.g., a pixel in an image or a user in an online social network) and an edge represents similarity between two multivariate data samples or actual relation (e.g., friendship) between nodes [14]. More often the K eigenvectors associated with the K smallest eigenvalues of the graph Laplacian matrix are used to cluster the entities into K clusters of high similarity. For brevity, throughout this paper we will call these eigenvectors as the K smallest eigenvectors.

The success of graph Laplacian matrix based methods for graph partitioning and spectral clustering can be explained by the fact that acquiring K smallest eigenvectors is equivalent to solving a relaxed graph cut minimization problem, which partitions a graph into K clusters by minimizing various objective functions including min cut, ratio cut or normalized cut [14]. Generally, in clustering K is selected to be much smaller than n (the number of data points), making full eigen decomposition (such as QR decomposition) unnecessary. An efficient alternative is to use methods that are based on power iteration, such as Arnoldi method or Lanczos method, which computes the leading eigenpairs through repeated matrix vector multiplication. ARPACK [13] library is a popular parallel implementation of different variants of Arnoldi and Lanczos method, which is used by many commercial software including Matlab.

However, in most situations the best value of K is unknown and a heuristic is used by clustering algorithms to determine the number of clusters, e.g., fixing a maximum number of clusters $K_{\max}$ and detecting a large gap in the values of the $K_{\max}$ largest sorted eigenvalues or normalized cut score [16,21]. Alternatively, this value of K can be determined based on domain knowledge [1]. For example, a user may require that the largest cluster size be no more than 10% of the total number of nodes or that the total inter-cluster edge weight be no greater than a certain amount. In these cases, the desired choice of K cannot be determined *a priori*. Over-estimation of the upper bound $K_{\max}$ on the number of clusters is expensive as the cost of finding K eigenpairs using the power iteration method grows rapidly with K. On the other hand, choosing an insufficiently large value for $K_{\max}$ runs the risk of severe bias. Setting $K_{\max}$ to the number of data points n is generally computationally infeasible, even for a moderate-sized graph. Therefore, an incremental eigenpair computation method that effectively

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

computes the K-th smallest eigenpair of graph Laplacian matrix by utilizing the previously computed $K-1$ smallest eigenpairs is needed. Such an iterative method obviates the need to set an upper bound $K_{\max}$ on K, and its efficiency can be explained by the adaptivity to increments in K.

By exploiting the special matrix properties and graph characteristics of a graph Laplacian matrix, we propose an efficient method for computing the $(K+1)$-th eigenpair given all of the K smallest eigenpairs, which we call the Incremental method of Increasing Orders (Incremental-IO). For each increment, given the previously computed smallest eigenpairs, we show that computing the next smallest eigenpair is equivalent to computing a leading eigenpair of a particular matrix, which transforms potentially tedious numerical computation (such as the iterative tridiagonalization and eigen-decomposition steps in the Lanczos algorithm [11]) to simple matrix power iterations of known computational efficiency [10]. We then compare the performance of Incremental-IO with a batch computation method which computes all of the K smallest eigenpairs in a single batch, and an incremental method adapted from the Lanczos algorithm, which we call the Lanczos method of Increasing Orders (Lanczos-IO). For a given number of eigenpairs K iterative matrix-vector multiplication of Lanczos procedure yields a set of Lanczos vectors ($\mathbf{Q}_\ell$), and a tridiagonal matrix ($\mathbf{T}_\ell$), followed by a full eigen-decomposition of $\mathbf{T}_\ell$ (ℓ is a value much smaller than the matrix size). Lanczos-IO saves the Lanczos vectors that were obtained while K eigenpairs were computed and use those to generate new Lanczos vectors for computing $(K+1)$-th eigenpair.

Comparing to the batch method, our experimental results show that for a given order K, Incremental-IO provides a significant reduction in computation time. Also, as K increases, the gap between Incremental-IO and the batch approach widens, providing an order of magnitude speedup. Experiments on real-life datasets show that the performance of Lanczos-IO is overly sensitive to the selection of augmented Lanczos vectors, a parameter that cannot be optimized *a priori*—for some of our experimental datasets, Lanczos-IO performs even worse than the batch method (see Sec. 6). Moreover, Lanczos-IO consumes significant amount of memory as it has to save the Lanczos vectors ($\mathbf{Q}_\ell$) for making the incremental approach realizable. In summary, Lanczos-IO, although an incremental eigenpair computation algorithm, falls short in terms of robustness.

To illustrate the real-life utility of incremental eigenpair computation methods, we design a user-guided spectral clustering algorithm which uses Incremental-IO. The algorithm provides clustering solution for a sequence of K values efficiently, and thus enable a user to compare these clustering solutions for facilitating the selection of the most appropriate clustering.

The contributions of this paper are summarized as follows.

1. We propose an incremental eigenpair computation method (Incremental-IO) for both unnormalized and normalized graph Laplacian matrices, by transforming the original eigenvalue decomposition problem into an efficient sequential leading eigenpair computation problem. Simulation results show that Incremental-IO generates the desired eigenpair accurately and has superior performance over the batch computation method in terms of computation time.
2. We show that Incremental-IO is robust in comparison to Lanczos-IO, which is an incremental eigenpair method that we design by adapting the Lanczos method.
3. We use several real-life datasets to demonstrate the utility of Incremental-IO. Specifically, we show that Incremental-IO is suitable for user-guided spectral clustering which provides a sequence of clustering results for unit increment of the number K of clusters, and updates the associated cluster evaluation metrics for helping a user in decision making.

2. RELATED WORKS

2.1 Incremental eigenvalue decomposition

The proposed method (Incremental-IO) aims to incrementally compute the smallest eigenpair of a given graph Laplacian matrix. There are several works that are named as incremental eigenvalue decomposition methods [8,9,17,18,25]. However, these works focus on updating the eigenstructure of graph Laplacian matrix of dynamic graphs when nodes (data samples) or edges are inserted or deleted from the graph, which are different from incremental computation of eigenpairs of increasing orders.

2.2 Cluster Count Selection for Spectral Clustering

Many spectral clustering algorithms utilize the eigenstructure of graph Laplacian matrix for selecting number of clusters. In [21], a value K that maximizes the gap between the K-th and the $(K+1)$-th smallest eigenvalue is selected as the number of clusters. In [16], a value K that minimizes the sum of cluster-wise Euclidean distance between each data point and the centroid obtained from K-means clustering on K smallest eigenvectors is selected as the number of clusters. In [31], the smallest eigenvectors of normalized graph Laplacian matrix are rotated to find the best alignment that reflects the true clusters. A model based method for determining the number of clusters is proposed in [22]. Note that aforementioned methods use only one single clustering metric to determine the number of clusters and often implicitly assume an upper bound on K (namely $K_{\max}$).

3. INCREMENTAL EIGENPAIR COMPUTATION FOR GRAPH LAPLACIAN MATRICES

3.1 Background

Throughout this paper bold uppercase letters (e.g., $\mathbf{X}$) denote matrices and $\mathbf{X}_{ij}$ (or $[\mathbf{X}]_{ij}$) denotes the entry in i-th row and j-th column of $\mathbf{X}$, bold lowercase letters (e.g., $\mathbf{x}$ or $\mathbf{x}_i$) denote column vectors, $(\cdot)^T$ denotes matrix or vector transpose, italic letters (e.g., x or x_i) denote scalars, and calligraphic uppercase letters (e.g., $\mathcal{X}$ or $\mathcal{X}_i$) denote sets. The $n \times 1$ vector of ones (zeros) is denoted by $\mathbf{1}_n$ ($\mathbf{0}_n$). The matrix $\mathbf{I}$ denotes an identity matrix and the matrix $\mathbf{O}$ denotes the matrix of zeros.

We use two $n \times n$ symmetric matrices, $\mathbf{A}$ and $\mathbf{W}$, to denote the adjacency and weight matrices of an undirected weighted simple graph G with n nodes and m edges. $\mathbf{A}_{ij} = 1$ if there is an edge between nodes i and j, and $\mathbf{A}_{ij} = 0$ otherwise. $\mathbf{W}$ is a nonnegative symmetric matrix such that $\mathbf{W}_{ij} \geq 0$ if

Table 1: Utility of the lemmas, corollaries, and theorems.

Graph Type / Laplacian Matrix	Unnormalized	Normalized
Connected Graphs	**Lemma** 1 **Theorem** 1	**Corollary** 1 **Corollary** 3
Disconnected Graphs	**Lemma** 2 **Theorem** 2	**Corollary** 2 **Corollary** 4

$\mathbf{A}_{ij} = 1$ and $\mathbf{W}_{ij} = 0$ if $\mathbf{A}_{ij} = 0$. Let $s_i = \sum_{j=1}^{n} \mathbf{W}_{ij}$ denote the strength of node i. Note that when $\mathbf{W} = \mathbf{A}$, the strength of a node is equivalent to its degree. $\mathbf{S} = \text{diag}(s_1, s_2, \ldots, s_n)$ is a diagonal matrix with nodal strength on its main diagonal and the off-diagonal entries being zero.

The (unnormalized) graph Laplacian matrix is defined as

$$\mathbf{L} = \mathbf{S} - \mathbf{W}. \tag{1}$$

One popular variant of the graph Laplacian matrix is the normalized graph Laplacian matrix defined as

$$\mathbf{L}_{\mathcal{N}} = \mathbf{S}^{-\frac{1}{2}}\mathbf{L}\mathbf{S}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{S}^{-\frac{1}{2}}\mathbf{W}\mathbf{S}^{-\frac{1}{2}}, \tag{2}$$

where $\mathbf{S}^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{s_1}}, \frac{1}{\sqrt{s_2}}, \ldots, \frac{1}{\sqrt{s_n}})$. The i-th smallest eigenvalue and its associated unit-norm eigenvector of $\mathbf{L}$ are denoted by $\lambda_i(\mathbf{L})$ and $\mathbf{v}_i(\mathbf{L})$, respectively. That is, the eigenpair $(\lambda_i, \mathbf{v}_i)$ of $\mathbf{L}$ has the relation $\mathbf{L}\mathbf{v}_i = \lambda_i\mathbf{v}_i$, and $\lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \ldots \leq \lambda_n(\mathbf{L})$. The eigenvectors have unit Euclidean norm and they are orthogonal to each other such that $\mathbf{v}_i^T\mathbf{v}_j = 1$ if $i = j$ and $\mathbf{v}_i^T\mathbf{v}_j = 0$ if $i \neq j$. The eigenvalues of $\mathbf{L}$ are said to be distinct if $\lambda_1(\mathbf{L}) < \lambda_2(\mathbf{L}) < \ldots < \lambda_n(\mathbf{L})$. Similar notations are used for $\mathbf{L}_{\mathcal{N}}$.

3.2 Theoretical foundations of the proposed method (Incremental-IO)

The following lemmas and corollaries provide the cornerstone for establishing the proposed incremental eigenpair computation method (Incremental-IO). The main idea is that we utilize the eigenspace structure of graph Laplacian matrix to inflate specific eigenpairs via a particular perturbation matrix, without affecting other eigenpairs. Incremental-IO can be viewed as a specialized Hotelling's deflation method [20] designed for graph Laplacian matrices by exploiting their spectral properties and associated graph characteristics. It works for both connected, and disconnected graphs using both normalized and unnormalized graph Laplacian matrices. For illustration purposes, in Table 1 we group the established lemmas, corollaries, and theorems under different graph types and different graph Laplacian matrices. Because of the page limit, the proofs of the established lemmas, theorems and corollaries are given in the extended version[1].

LEMMA 1. *Assume that G is a connected graph and $\mathbf{L}$ is the graph Laplacian with s_i denoting the sum of the entries in the i-th row of the weight matrix $\mathbf{W}$. Let $s = \sum_{i=1}^{n} s_i$ and define $\widetilde{\mathbf{L}} = \mathbf{L} + \frac{s}{n}\mathbf{1}_n\mathbf{1}_n^T$. Then the eigenpairs of $\widetilde{\mathbf{L}}$ satisfy $(\lambda_i(\widetilde{\mathbf{L}}), \mathbf{v}_i(\widetilde{\mathbf{L}})) = (\lambda_{i+1}(\mathbf{L}), \mathbf{v}_{i+1}(\mathbf{L}))$ for $1 \leq i \leq n-1$ and $(\lambda_n(\widetilde{\mathbf{L}}), \mathbf{v}_n(\widetilde{\mathbf{L}})) = (s, \frac{\mathbf{1}_n}{\sqrt{n}})$.*

COROLLARY 1. *For a normalized graph Laplacian matrix $\mathbf{L}_{\mathcal{N}}$, assume G is a connected graph and let $\widetilde{\mathbf{L}}_{\mathcal{N}} = \mathbf{L}_{\mathcal{N}} + \frac{2}{s}\mathbf{S}^{\frac{1}{2}}\mathbf{1}_n\mathbf{1}_n^T\mathbf{S}^{\frac{1}{2}}$. Then $(\lambda_i(\widetilde{\mathbf{L}}_{\mathcal{N}}), \mathbf{v}_i(\widetilde{\mathbf{L}}_{\mathcal{N}})) = (\lambda_{i+1}(\mathbf{L}_{\mathcal{N}}), \mathbf{v}_{i+1}(\mathbf{L}_{\mathcal{N}}))$ for $1 \leq i \leq n-1$ and $(\lambda_n(\widetilde{\mathbf{L}}_{\mathcal{N}}), \mathbf{v}_n(\widetilde{\mathbf{L}}_{\mathcal{N}})) = (2, \frac{\mathbf{S}^{\frac{1}{2}}\mathbf{1}_n}{\sqrt{s}})$.*

[1] http://arxiv.org/abs/1512.07349

LEMMA 2. *Assume that G is a disconnected graph with $\delta \geq 2$ connected components, and let $s = \sum_{i=1}^{n} s_i$, let $\mathbf{V} = [\mathbf{v}_1(\mathbf{L}), \mathbf{v}_2(\mathbf{L}), \ldots, \mathbf{v}_\delta(\mathbf{L})]$, and let $\widetilde{\mathbf{L}} = \mathbf{L} + s\mathbf{V}\mathbf{V}^T$. Then $(\lambda_i(\widetilde{\mathbf{L}}), \mathbf{v}_i(\widetilde{\mathbf{L}})) = (\lambda_{i+\delta}(\mathbf{L}), \mathbf{v}_{i+\delta}(\mathbf{L}))$ for $1 \leq i \leq n-\delta$, $\lambda_i(\widetilde{\mathbf{L}}) = s$ for $n-\delta+1 \leq i \leq n$, and $[\mathbf{v}_{n-\delta+1}(\widetilde{\mathbf{L}}), \mathbf{v}_{n-\delta+2}, (\widetilde{\mathbf{L}}), \ldots, \mathbf{v}_n(\widetilde{\mathbf{L}})] = \mathbf{V}$.*

COROLLARY 2. *For a normalized graph Laplacian matrix $\mathbf{L}_{\mathcal{N}}$, assume G is a disconnected graph with $\delta \geq 2$ connected components. Let $\mathbf{V}_\delta = [\mathbf{v}_1(\mathbf{L}_{\mathcal{N}}), \mathbf{v}_2(\mathbf{L}_{\mathcal{N}}), \ldots, \mathbf{v}_\delta(\mathbf{L}_{\mathcal{N}})]$, and let $\widetilde{\mathbf{L}}_{\mathcal{N}} = \mathbf{L}_{\mathcal{N}} + 2\mathbf{V}_\delta\mathbf{V}_\delta^T$. Then $(\lambda_i(\widetilde{\mathbf{L}}_{\mathcal{N}}), \mathbf{v}_i(\widetilde{\mathbf{L}}_{\mathcal{N}})) = (\lambda_{i+\delta}(\mathbf{L}_{\mathcal{N}}), \mathbf{v}_{i+\delta}(\mathbf{L}_{\mathcal{N}}))$ for $1 \leq i \leq n-\delta$, $\lambda_i(\widetilde{\mathbf{L}}_{\mathcal{N}}) = 2$ for $n-\delta+1 \leq i \leq n$, and $[\mathbf{v}_{n-\delta+1}(\widetilde{\mathbf{L}}_{\mathcal{N}}), \mathbf{v}_{n-\delta+2}, (\widetilde{\mathbf{L}}_{\mathcal{N}}), \ldots, \mathbf{v}_n(\widetilde{\mathbf{L}}_{\mathcal{N}})] = \mathbf{V}_\delta$.*

REMARK 1. *note that the columns of any matrix $\mathbf{V}' = \mathbf{V}\mathbf{R}$ with an orthonormal transformation matrix $\mathbf{R}$ (i.e., $\mathbf{R}^T\mathbf{R} = \mathbf{I}$) are also the largest δ eigenvectors of $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{L}}_{\mathcal{N}}$ in **Lemma** 2 and **Corollary** 2. Without loss of generality we consider the case $\mathbf{R} = \mathbf{I}$.*

3.3 Incremental method of increasing orders

Given the K smallest eigenpairs of a graph Laplacian matrix, we prove that computing the $(K+1)$-th smallest eigenpair is equivalent to computing the leading eigenpair (the eigenpair with the largest eigenvalue in absolute value) of a certain perturbed matrix. The advantage of this transformation is that the leading eigenpair can be efficiently computed via matrix power iteration methods [11, 13].

Let $\mathbf{V}_K = [\mathbf{v}_1(\mathbf{L}), \mathbf{v}_2(\mathbf{L}), \ldots, \mathbf{v}_K(\mathbf{L})]$ be the matrix with columns being the K smallest eigenvectors of $\mathbf{L}$ and let $\mathbf{\Lambda}_K = \text{diag}(s - \lambda_1(\mathbf{L}), s - \lambda_2(\mathbf{L}), \ldots, s - \lambda_K(\mathbf{L}))$ be the diagonal matrix with $\{s - \lambda_i(\mathbf{L})\}_{i=1}^{K}$ being its main diagonal. The following theorems show that given the K smallest eigenpairs of $\mathbf{L}$, the $(K+1)$-th smallest eigenpair of $\mathbf{L}$ is the leading eigenvector of the original graph Laplacian matrix perturbed by a certain matrix.

THEOREM 1. (connected graphs) *Given $\mathbf{V}_K$ and $\mathbf{\Lambda}_K$, assume that G is a connected graph. Then the eigenpair $(\lambda_{K+1}(\mathbf{L}), \mathbf{v}_{K+1}(\mathbf{L}))$ is a leading eigenpair of the matrix $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^T + \frac{s}{n}\mathbf{1}_n\mathbf{1}_n^T - s\mathbf{I}$. In particular, if $\mathbf{L}$ has distinct eigenvalues, then $(\lambda_{K+1}(\mathbf{L}), \mathbf{v}_{K+1}(\mathbf{L})) = (\lambda_1(\widetilde{\mathbf{L}}) + s, \mathbf{v}_1(\widetilde{\mathbf{L}}))$.*

The next theorem describes an incremental eigenpair computation method when the graph G is a disconnected graph with δ connected components. The columns of the matrix $\mathbf{V}_\delta$ are the δ smallest eigenvectors of $\mathbf{L}$. Note that $\mathbf{V}_\delta$ has a canonical representation that the nonzero entries of each column are a constant and their indices indicate the nodes in each connected component [6, 14], and the columns of $\mathbf{V}_\delta$ are the δ smallest eigenvectors of $\mathbf{L}$ with eigenvalue 0 [6]. Since the δ smallest eigenpairs with the canonical representation are trivial by identifying the connected components in the graph, we only focus on computing the $(K+1)$-th smallest eigenpair given K smallest eigenpairs, where $K \geq \delta$. The columns of the matrix $\mathbf{V}_{K,\delta} = [\mathbf{v}_{\delta+1}(\mathbf{L}), \mathbf{v}_{\delta+2}(\mathbf{L}), \ldots, \mathbf{v}_K(\mathbf{L})]$ are the $(\delta+1)$-th to the K-th smallest eigenvectors of $\mathbf{L}$ and the matrix $\mathbf{\Lambda}_{K,\delta} = \text{diag}(s - \lambda_{\delta+1}(\mathbf{L}), s - \lambda_{\delta+2}(\mathbf{L}), \ldots, s - \lambda_K(\mathbf{L}))$ is the diagonal matrix with $\{s - \lambda_i(\mathbf{L})\}_{i=\delta+1}^{K}$ being its main diagonal. If $K = \delta$, $\mathbf{V}_{K,\delta}$ and $\mathbf{\Lambda}_{K,\delta}$ are defined as the matrix with all entries being zero, i.e., $\mathbf{O}$.

THEOREM 2. (disconnected graphs) *Assume that G is a disconnected graph with $\delta \geq 2$ connected components, given $\mathbf{V}_{K,\delta}$, $\mathbf{\Lambda}_{K,\delta}$ and $K \geq \delta$, the eigenpair $(\lambda_{K+1}(\mathbf{L}), \mathbf{v}_{K+1}(\mathbf{L}))$ is a leading eigenpair of the matrix $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{V}_{K,\delta}\mathbf{\Lambda}_{K,\delta}\mathbf{V}_{K,\delta}^T + s\mathbf{V}_\delta\mathbf{V}_\delta^T - s\mathbf{I}$. In particular, if $\mathbf{L}$ has distinct nonzero eigenvalues, then $(\lambda_{K+1}(\mathbf{L}), \mathbf{v}_{K+1}(\mathbf{L})) = (\lambda_1(\widetilde{\mathbf{L}}) + s, \mathbf{v}_1(\widetilde{\mathbf{L}}))$.*

Following the same methodology for proving **Theorem** 1 and using **Corollary** 1, for normalized graph Laplacian matrices, let $\mathbf{V}_K = [\mathbf{v}_1(\mathbf{L}_\mathcal{N}), \mathbf{v}_2(\mathbf{L}_\mathcal{N}), \ldots, \mathbf{v}_K(\mathbf{L}_\mathcal{N})]$ be the matrix with columns being the K smallest eigenvectors of $\mathbf{L}_\mathcal{N}$ and let $\mathbf{\Lambda}_K = \text{diag}(2-\lambda_1(\mathbf{L}_\mathcal{N}), 2-\lambda_2(\mathbf{L}_\mathcal{N}), \ldots, 2-\lambda_K(\mathbf{L}_\mathcal{N}))$. The following corollary provides the basis for incremental eigenpair computation for normalized graph Laplacian matrix of connected graphs.

COROLLARY 3. *For the normalized graph Laplacian matrix $\mathbf{L}_\mathcal{N}$ of a connected graph G, given $\mathbf{V}_K$ and $\mathbf{\Lambda}_K$, the eigenpair $(\lambda_{K+1}(\mathbf{L}_\mathcal{N}), \mathbf{v}_{K+1}(\mathbf{L}_\mathcal{N}))$ is a leading eigenpair of the matrix $\widetilde{\mathbf{L}}_\mathcal{N} = \mathbf{L}_\mathcal{N} + \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^T + \frac{2}{s}\mathbf{S}^{\frac{1}{2}}\mathbf{1}_n\mathbf{1}_n^T\mathbf{S}^{\frac{1}{2}} - 2\mathbf{I}$. In particular, if $\mathbf{L}_\mathcal{N}$ has distinct eigenvalues, then $(\lambda_{K+1}(\mathbf{L}_\mathcal{N}), \mathbf{v}_{K+1}(\mathbf{L}_\mathcal{N})) = (\lambda_1(\widetilde{\mathbf{L}}_\mathcal{N}) + 2, \mathbf{v}_1(\widetilde{\mathbf{L}}_\mathcal{N}))$.*

For disconnected graphs with δ connected components, let $\mathbf{V}_{K,\delta} = [\mathbf{v}_{\delta+1}(\mathbf{L}_\mathcal{N}), \mathbf{v}_{\delta+2}(\mathbf{L}_\mathcal{N}), \ldots, \mathbf{v}_K(\mathbf{L}_\mathcal{N})]$ with columns being the $(\delta+1)$-th to the K-th smallest eigenvectors of $\mathbf{L}_\mathcal{N}$ and let $\mathbf{\Lambda}_{K,\delta} = \text{diag}(2 - \lambda_{\delta+1}(\mathbf{L}_\mathcal{N}), 2 - \lambda_{\delta+2}(\mathbf{L}_\mathcal{N}), \ldots, 2 - \lambda_K(\mathbf{L}_\mathcal{N}))$. Based on **Corollary** 2, the following corollary provides an incremental eigenpair computation method for normalized graph Laplacian matrix of disconnected graphs.

COROLLARY 4. *For the normalized graph Laplacian matrix $\mathbf{L}_\mathcal{N}$ of a disconnected graph G with $\delta \geq 2$ connected components, given $\mathbf{V}_{K,\delta}$, $\mathbf{\Lambda}_{K,\delta}$ and $K \geq \delta$, the eigenpair $(\lambda_{K+1}(\mathbf{L}_\mathcal{N}), \mathbf{v}_{K+1}(\mathbf{L}_\mathcal{N}))$ is a leading eigenpair of the matrix $\widetilde{\mathbf{L}}_\mathcal{N} = \mathbf{L}_\mathcal{N} + \mathbf{V}_{K,\delta}\mathbf{\Lambda}_{K,\delta}\mathbf{V}_{K,\delta}^T + \frac{2}{s}\mathbf{S}^{\frac{1}{2}}\mathbf{1}_n\mathbf{1}_n^T\mathbf{S}^{\frac{1}{2}} - 2\mathbf{I}$. In particular, if $\mathbf{L}_\mathcal{N}$ has distinct eigenvalues, then $(\lambda_{K+1}(\mathbf{L}_\mathcal{N}), \mathbf{v}_{K+1}(\mathbf{L}_\mathcal{N}))$ $= (\lambda_1(\widetilde{\mathbf{L}}_\mathcal{N}) + 2, \mathbf{v}_1(\widetilde{\mathbf{L}}_\mathcal{N}))$.*

3.4 Computational complexity analysis

Here we analyze the computational complexity of Incremental-IO and compare it with the batch computation method. Incremental-IO utilizes the existing K smallest eigenpairs to compute the $(K+1)$-th smallest eigenpair as described in Sec. 3.3, whereas the batch computation method recomputes all eigenpairs for each value of K. Both methods can be easily implemented via well-developed numerical computation packages such as ARPACK [13].

Following the analysis in [10], the average relative error of the leading eigenvalue from the Lanczos algorithm [11] has an upper bound of the order $O\left(\frac{(\ln n)^2}{t^2}\right)$, where n is the number of nodes in the graph and t is the number of iterations for Lanczos algorithm. Therefore, when one sequentially computes from $k = 2$ to $k = K$ smallest eigenpairs, for Incremental-IO the upper bound on the average relative error of K smallest eigenpairs is $O\left(\frac{K(\ln n)^2}{t^2}\right)$ since in each increment computing the corresponding eigenpair can be transformed to a leading eigenpair computation process as described in Sec. 3.3. On the other hand, for the batch computation method, the upper bound on the average relative error of K smallest eigenpairs is $O\left(\frac{K^2(\ln n)^2}{t^2}\right)$ since for the k-th increment ($k \leq K$) it needs to compute all k smallest eigenpairs from scratch. These results also imply that to reach the same average relative error ϵ for sequential computation of K smallest eigenpairs, Incremental-IO requires $\Omega\left(\sqrt{\frac{K}{\epsilon}}\ln n\right)$ iterations, whereas the batch method requires $\Omega\left(\frac{K\ln n}{\sqrt{\epsilon}}\right)$ iterations. It is difficult to analyze the computational complexity of Lanczos-IO, as its convergence results heavily depend on the quality of previously generated Lanczos vectors.

4. APPLICATION: USER-GUIDED SPECTRAL CLUSTERING WITH INCREMENTAL-IO

Based on the developed incremental eigenpair computation method (Incremental-IO) in Sec. 3, we propose an incremental algorithm for user-guided spectral clustering as summarized in **Algorithm** 1. This algorithm sequentially computes the smallest eigenpairs via Incremental-IO (steps 1-3) for spectral clustering and provides a sequence of clusters with the values of user-specified clustering metrics.

The input graph is a connected undirected weighted graph $\mathbf{W}$ and we convert it to the reduced weighted graph $\mathbf{W}_\mathcal{N} = \mathbf{S}^{-\frac{1}{2}}\mathbf{W}\mathbf{S}^{-\frac{1}{2}}$ to alleviate the effect of imbalanced edge weights. The entries of $\mathbf{W}_\mathcal{N}$ are properly normalized by the nodal strength such that $[\mathbf{W}_\mathcal{N}]_{ij} = \frac{[\mathbf{W}]_{ij}}{\sqrt{s_i \cdot s_j}}$. We then obtain the graph Laplacian matrix $\mathbf{L}$ for $\mathbf{W}_\mathcal{N}$ and incrementally compute the eigenpairs of $\mathbf{L}$ via Incremental-IO (steps 1-3) until the user decides to stop further computation.

Starting from $K = 2$ clusters, the algorithm incrementally computes the K-th smallest eigenpair $(\lambda_K(\mathbf{L}), \mathbf{v}_K(\mathbf{L}))$ of $\mathbf{L}$ with the knowledge of the previous $K-1$ smallest eigenpairs via **Theorem** 1 and obtains matrix $\mathbf{V}_K$ containing K smallest eigenvectors. By viewing each row in $\mathbf{V}_K$ as a K-dimensional vector, K-means clustering is implemented to separate the rows in $\mathbf{V}_K$ into K clusters. For each increment, the identified K clusters are denoted by $\{\widehat{G}_k\}_{k=1}^K$, where $\widehat{G}_k$ is a graph partition with $\widehat{n}_k$ nodes and $\widehat{m}_k$ edges.

In addition to incremental computation of smallest eigenpairs, for each increment the algorithm can also be used to update clustering metrics such as normalized cut, modularity, and cluster size distribution, in order to provide users with clustering information to stop the incremental computation process. The incremental computation algorithm allows users to efficiently track the changes in clusters as the number K of hypothesized clusters increases.

Note that **Algorithm** 1 is proposed for connected graphs and their corresponding unnormalized graph Laplacian matrices. The algorithm can be easily adapted to disconnected graphs or normalized graph Laplacian matrices by modifying steps 1-3 based on the developed results in **Theorem** 2, **Corollary** 3 and **Corollary** 4.

5. IMPLEMENTATION

We implement the proposed incremental eigenpair computation method (Incremental-IO) using Matlab R2015a's "eigs" function, which is based on ARPACK package [13]. Note that this function takes a parameter K and returns K leading eigenpairs of the given matrix. The *eigs* function is implemented in Matlab with a Lanczos algorithm that computes the leading eigenpairs (the implicitly-restarted Lanczos method [3]). This Matlab function iteratively generates

Algorithm 1 Incremental algorithm for user-guided spectral clustering using Incremental-IO (steps 1-3)

Input: connected undirected weighted graph $\mathbf{W}$
Output: K clusters $\{\widehat{G}_k\}_{k=1}^K$
Initialization: $K=2$. $\mathbf{V}_1=\mathbf{\Lambda}_1=\mathbf{O}$. Flag $=1$.
$\mathbf{S}=\text{diag}(\mathbf{W}\mathbf{1}_n)$. $\mathbf{W}_{\mathcal{N}}=\mathbf{S}^{-\frac{1}{2}}\mathbf{W}\mathbf{S}^{-\frac{1}{2}}$.
$\mathbf{L}=\text{diag}(\mathbf{W}_{\mathcal{N}}\mathbf{1}_n)-\mathbf{W}_{\mathcal{N}}$. $s=\mathbf{1}_n^T\mathbf{W}_{\mathcal{N}}\mathbf{1}_n$.
while Flag$=1$ **do**
1. $\widetilde{\mathbf{L}}=\mathbf{L}+\mathbf{V}_{K-1}\mathbf{\Lambda}_{K-1}\mathbf{V}_{K-1}^T+\frac{s}{n}\mathbf{1}_n\mathbf{1}_n^T-s\mathbf{I}$.
2. Compute the leading eigenpair $(\lambda_1(\widetilde{\mathbf{L}}),\mathbf{v}_1(\widetilde{\mathbf{L}}))$ and set $(\lambda_K(\mathbf{L}),\mathbf{v}_K(\mathbf{L}))=(\lambda_1(\widetilde{\mathbf{L}})+s,\mathbf{v}_1(\widetilde{\mathbf{L}}))$.
3. Update K smallest eigenpairs of $\mathbf{L}$ by $\mathbf{V}_K=[\mathbf{V}_{K-1}\ \mathbf{v}_K]$ and $[\mathbf{\Lambda}_K]_{KK}=s-\lambda_K(\mathbf{L})$.
4. Perform K-means clustering on the rows of $\mathbf{V}_K$ to obtain K clusters $\{\widehat{G}_k\}_{k=1}^K$.
5. Compute user-specified clustering metrics.
if user decides to stop **then** Flag$=0$
Output K clusters $\{\widehat{G}_k\}_{k=1}^K$
else
Go back to step 1 with $K=K+1$.
end if
end while

Lanczos vectors starting from an initial vector (the default setting is a random vector) with restart. Following **Theorem** 1, Incremental-IO works by sequentially perturbing the graph Laplacian matrix $\mathbf{L}$ with a particular matrix and computing the leading eigenpair of the perturbed matrix $\widetilde{\mathbf{L}}$ (see **Algorithm** 1) by calling $eigs(\widetilde{\mathbf{L}},1)$. For the batch computation method, we use $eigs(\mathbf{L},K)$ to compute the desired K eigenpairs from scratch as K increases.

For implementing Lanczos-IO, we extend the Lanczos algorithm of fixed order (K is fixed) using the PROPACK package [12]. As we have stated earlier, Lanczos-IO works by storing all previously generated Lanczos vectors and using them to compute new Lanczos vectors for each increment in K. The general procedure of computing K leading eigenpairs of a real symmetric matrix $\mathbf{M}$ using Lanczos-IO is described in **Algorithm** 2. The operation of Lanczos-IO is similar to the explicitly-restarted Lanczos algorithm [29], which restarts the computation of Lanczos vectors with a subset of previously computed Lanczos vectors. Note that the Lanczos-IO consumes additional memory for storing all previously computed Lanczos vectors when compared with the proposed incremental method in **Algorithm** 1, since the *eigs* function uses the implicitly-restarted Lanczos method that spares the need of storing Lanczos vectors for restart.

To apply Lanczos-IO to spectral clustering of increasing orders, we can set $\mathbf{M}=\mathbf{L}+\frac{s}{n}\mathbf{1}_n\mathbf{1}_n^T-s\mathbf{I}$ to obtain the smallest eigenvectors of $\mathbf{L}$. Throughout the experiments the parameters in **Algorithm** 2 are set to be $Z_{ini}=20$ and Tolerence $=\epsilon\cdot\|\mathbf{M}\|$, where ϵ is the machine precision, $\|\mathbf{M}\|$ is the operator norm of $\mathbf{M}$, and these settings are consistent with the settings used in *eigs* function [13]. The number of augmented Lanczos vectors Z_{aug} is set to be 10, and the effect of Z_{aug} on the computation time is discussed in Sec. 6.2. The Matlab implementation of the aforementioned batch method, Lanczos-IO, and Incremental-IO are available at https://sites.google.com/site/pinyuchenpage/codes.

Algorithm 2 Lanczos method of Increasing Orders (Lanczos-IO)

Input: real symmetric matrix $\mathbf{M}$, # of initial Lanczos vectors Z_{ini}, # of augmented Lanczos vectors Z_{aug}
Output: K leading eigenpairs $\{\lambda_i,\mathbf{v}_i\}_{i=1}^K$ of $\mathbf{M}$
Initialization: Compute Z_{ini} Lanczos vectors as columns of $\mathbf{Q}$ and the corresponding tridiagonal matrix $\mathbf{T}$ of $\mathbf{M}$. Flag $=1$. $K=1$. $Z=Z_{ini}$.
while Flag$=1$ **do**
1. Obtain the K leading eigenpairs $\{t_i,\mathbf{u}_i\}_{i=1}^K$ of $\mathbf{T}$. $\mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_K]$.
2. Residual error $=|\mathbf{T}(Z-1,Z)\cdot\mathbf{U}(Z,K)|$
while Residual error $>$ Tolerence **do**
2-1. $Z=Z+Z_{aug}$
2-2. Based on $\mathbf{Q}$ and $\mathbf{T}$, compute the next Z_{aug} Lanczos vectors as columns of $\mathbf{Q}_{aug}$ and the augmented tridiagonal matrix $\mathbf{T}_{aug}$
2-3. $\mathbf{Q}\leftarrow[\mathbf{Q}\ \mathbf{Q}_{aug}]$ and $\mathbf{T}\leftarrow\begin{bmatrix}\mathbf{T} & \mathbf{O}\\ \mathbf{O} & \mathbf{T}_{aug}\end{bmatrix}$
2-4. Go back to step 1
end while
3. $\{\lambda_i\}_{i=1}^K=\{t_i\}_{i=1}^K$. $[\mathbf{v}_1,\ldots,\mathbf{v}_K]=\mathbf{QU}$.
if user decides to stop **then** Flag$=0$
Output K leading eigenpairs $\{\lambda_i,\mathbf{v}_i\}_{i=1}^K$
else
Go back to step 1 with $K=K+1$.
end if
end while

6. EXPERIMENTAL RESULTS

We perform several experiments: first, compare the computation time between Incremental-IO, Lanczos-IO, and the batch method; second, numerically verify the accuracy of Incremental-IO; third, demonstrate the usages of Incremental-IO for user-guided spectral clustering. For the first experiment, we generate synthetic Erdos-Renyi random graphs of various sizes. For the second experiment, we compare the consistency of eigenpairs obtained from Incremental-IO and the batch method. For the third experiment, we use six popular graph datasets as summarized in Table 2.

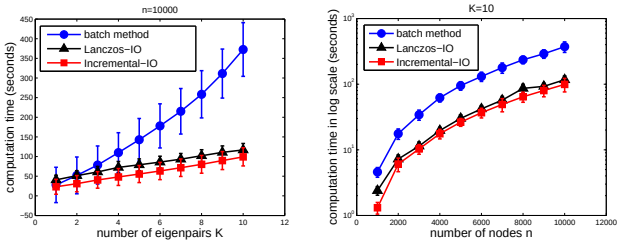
6.1 Comparison of computation time on simulated graphs

To illustrate the advantage of Incremental-IO, we compare its computation time with the other two methods, the batch method and Lanczos-IO, for varying order K and varying graph size n. The Erdos-Renyi random graphs that we build are used for this comparison. Fig. 1 (a) shows the computation time of Incremental-IO, Lanczos-IO, and the batch computation method for sequentially computing from $K=2$ to $K=10$ smallest eigenpairs. It is observed that the computation time of Incremental-IO and Lanczos-IO grows linearly as K increases, whereas the computation time of the batch method grows superlinearly with K.

Fig. 1 (b) shows the computation time of all three methods with respect to different graph size n. It is observed that the difference in computation time between the batch method and the two incremental methods grow exponentially as n increases, which suggests that in this experiment Incremental-IO and Lanczos-IO are more efficient than the batch computation method, especially for large graphs. It is

Table 2: Statistics of datasets

Dataset	Nodes	Edges	Density
Minnesota Road[2]	2640 intersections	3302 roads	0.095%
Power Grid[3]	4941 power stations	6594 power lines	0.054%
CLUTO[4]	7674 data points	748718 kNN edges	2.54%
Swiss Roll[5]	20000 data points	81668 kNN edges	0.041%
Youtube[6]	13679 users	76741 interactions	0.082%
BlogCatalog[7]	10312 bloggers	333983 interactions	0.63%



(a) (b)

Figure 1: Sequential eigenpair computation time on Erdos-Renyi random graphs with edge connection probability $p = 0.1$. The marker represents averaged computation time of 50 trials and the error bar represents standard deviation. (a) Computation time with $n = 10000$ and different number of eigenpairs K. (b) Computation time with $K = 10$ and different number of nodes n.

worth noting that although Lanczos-IO has similar performance in computation time as Incremental-IO, it requires additional memory allocation for storing all previously computed Lanczos vectors.

6.2 Comparison of computation time on real-life datasets

Fig. 2 shows the time improvement of Incremental-IO relative to the batch method for the real-life datasets listed in Table 2, where the difference in computation time is displayed in log scale to accommodate large variance of time improvement across datasets that are of widely varying size. It is observed that the gain in computational time via Incremental-IO is more pronounced as cluster count K increases, which demonstrates the merit of the proposed incremental method.

On the other hand, although Lanczos-IO is also an incremental method, in addition to the well-known issue of requiring memory allocation for storing all Lanczos vectors, the experimental results show that it does not provide performance robustness as Incremental-IO does, as it can perform even worse than the batch method for some cases. Fig. 3 shows that Lanczos-IO actually results in excessive computation time compared with the batch method for four out of the six datasets, whereas in Fig. 2 Incremental-IO is superior than the batch method for all these datasets, which demonstrates the robustness of Incremental-IO over Lanczos-IO. The reason of lacking robustness for Lanczos-IO can be explained by the fact the previously computed Lanczos vectors may not be effective in minimizing the Ritz approximation error of the desired eigenpairs. In contrast, Incremental-IO and the batch method adopt the implicitly-restarted Lanczos method, which restarts the Lanczos algorithm when the generated Lanczos vectors fail to meet the Ritz approximation criterion, and may eventually lead to faster convergence. Furthermore, Fig. 4 shows that Lanczos-IO is overly sensitive to the number of augmented Lanczos vectors Z_{aug}, which is a parameter that cannot be optimized *a priori*.

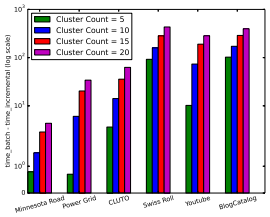


Figure 2: Computation time improvement of Incremental-IO relative to the batch method. Incremental-IO outperforms the batch method for all cases, and has improvement w.r.t. K.

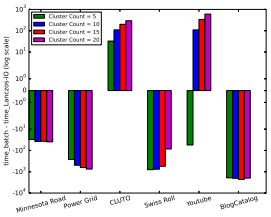


Figure 3: Computation time improvement of Lanczos-IO relative to the batch method. Negative values mean that Lanczos-IO requires more computation time than the batch method.

Theorem 1 establishes that the proposed incremental method exactly computes the K-th eigenpair using 1 to $(K-1)$-th eigenpairs, yet for the sake of experiments with real datasets, we have computed the normed eigenvalue difference and eigenvector correlation of the K smallest eigenpairs using the batch method and Incremental-IO as displayed in Fig. 5. The K smallest eigenpairs are identical as expected; to be specific, using Matlab library, on the Minnesota road dataset for $K = 20$, the normed eigenvalue difference is 7×10^{-12} and the associated eigenvectors are identical up to differences in sign. Results for the other datasets are reported in the extended version[1].

6.3 Clustering metrics for user-guided spectral clustering

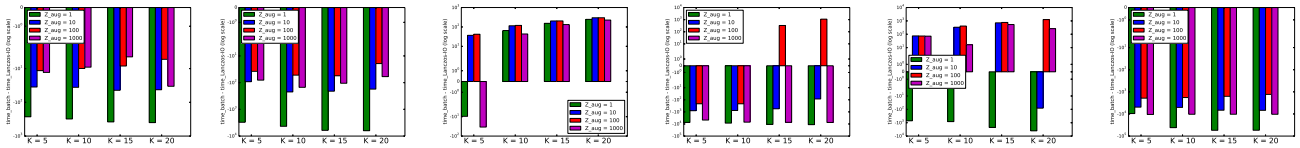
In real-life, an analyst can use Incremental-IO for clustering along with a mechanism for selecting the best choice of K starting from $K = 2$. To demonstrate this, in the experiment we use five clustering metrics that can be used for online decision making regarding the value of K. These metrics are commonly used in clustering unweighted and weighted graphs and they are summarized as follows.

1. Modularity: modularity is defined as

$$\text{Mod} = \sum_{i=1}^{K} \left(\frac{W(\mathcal{C}_i, \mathcal{C}_i)}{W(\mathcal{V}, \mathcal{V})} - \left(\frac{W(\mathcal{C}_i, \mathcal{V})}{W(\mathcal{V}, \mathcal{V})} \right)^2 \right), \qquad (3)$$

where $\mathcal{V}$ is the set of all nodes in the graph, $\mathcal{C}_i$ is the i-th cluster, $W(\mathcal{C}_i, \mathcal{C}_i)$ ($W(\mathcal{C}_i, \overline{\mathcal{C}_i})$) denotes the sum of weights of all internal (external) edges of the i-th cluster, $W(\mathcal{C}_i, \mathcal{V}) = W(\mathcal{C}_i, \mathcal{C}_i) + W(\mathcal{C}_i, \overline{\mathcal{C}_i})$, and $W(\mathcal{V}, \mathcal{V}) = \sum_{j=1}^{n} s_j = s$ denotes the total nodal strength.

[2] http://www.cs.purdue.edu/homes/dgleich/nmcomp/matlab/minnesota
[3] http://www-personal.umich.edu/ mejn/netdata
[4] http://glaros.dtc.umn.edu/gkhome/views/cluto
[5] http://isomap.stanford.edu/datasets.html
[6] http://socialcomputing.asu.edu/datasets/YouTube
[7] http://socialcomputing.asu.edu/datasets/BlogCatalog



(a) Minnesota Road (b) Power Grid (c) CLUTO (d) Swiss Roll (e) Youtube (f) BlogCatalog

Figure 4: The effect of number of augmented Lanczos vectors Z_{aug} of Lanczos-IO in **Algorithm** 2 on computation time improvement relative to the batch method. Negative values mean that the computation time of Lanczos-IO is larger than that of the batch method. The results show that Lanczos-IO is not a robust incremental eigenpair computation method. Intuitively, small Z_{aug} may incur many iterations in the second step of **Algorithm** 2, whereas large Z_{aug} may pose computation burden in the first step of **Algorithm** 2, and therefore both cases lead to the increase in computation time.

2. Scaled normalized cut (SNC): NC is defined as [30]

$$\text{NC} = \sum_{i=1}^{K} \frac{W(\mathcal{C}_i, \overline{\mathcal{C}_i})}{W(\mathcal{C}_i, \mathcal{V})}. \tag{4}$$

SNC is NC divided by the number of clusters, i.e., NC/K.

3. Scaled median (or maximum) cluster size: Scaled medium (maximum) cluster size is the medium (maximum) cluster size of K clusters divided by the total number of nodes n of a graph.

4. Scaled spectrum energy: scaled spectrum energy is the sum of the K smallest eigenvalues of the graph Laplacian matrix $\mathbf{L}$ divided by the sum of all eigenvalues of $\mathbf{L}$, which can be computed by

$$\text{scaled spectrum energy} = \frac{\sum_{i=1}^{K} \lambda_i(\mathbf{L})}{\sum_{j=1}^{n} \mathbf{L}_{jj}}, \tag{5}$$

where $\lambda_i(\mathbf{L})$ is the i-th smallest eigenvalue of $\mathbf{L}$ and $\sum_{j=1}^{n} \mathbf{L}_{jj} = \sum_{i=1}^{n} \lambda_i(\mathbf{L})$ is the sum of diagonal elements of $\mathbf{L}$.

These metrics provide alternatives for gauging the quality of the clustering method. For example, Mod and NC reflect the trade-off between intracluster similarity and intercluster separation. Therefore, the larger the value of Mod, the better the clustering quality, and the smaller the value of NC, the better the clustering quality. Scaled spectrum energy is a typical measure of cluster quality for spectral clustering [16, 21, 31], and smaller spectrum energy means better separability of clusters. For Mod and scaled NC, a user might look for a cluster count K such that the increment in the clustering metric is not significant, i.e., the clustering metric is saturated beyond such a K. For scaled median and maximum cluster size, a user might require the cluster count K to be such that the clustering metric is below a desired value. For scaled spectrum energy, a user might look for a noticeable increase in the clustering metric between consecutive values of K.

6.4 Demonstration

Here we use Minnesota Road data to demonstrate how users can utilize the clustering metrics in Sec. 6.3 to determine the number of clusters. The five metrics evaluated for Minnesota Road clustering with respect to different cluster counts K are displayed in Fig. 6. Starting from $K = 2$ clusters, these metrics are updated by the incremental user-guided spectral clustering algorithm (**Algorithm** 1) as K increases. If the user imposes that the maximum cluster size should be less than 30% of the total number of nodes, then the algorithm returns clustering results with a number of clusters of $K = 6$ or greater. Inspecting the modularity

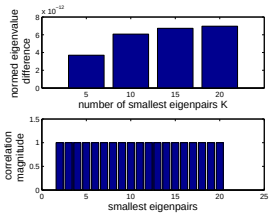


Figure 5: Consistency of eigenpairs computed by the batch computation method and Incremental-IO for Minnesota Road dataset.

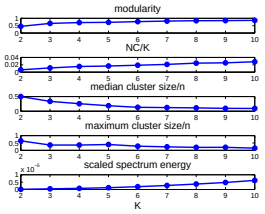
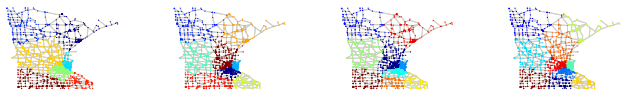


Figure 6: Five clustering metrics computed incrementally via **Algorithm** 1 for Minnesota Road.



(a) $K = 7$ (b) $K = 8$ (c) $K = 9$ (d) $K = 10$

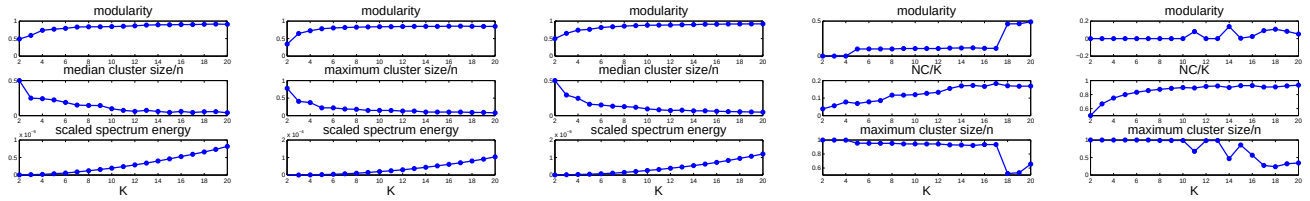
Figure 7: Visualization of user-guided spectral clustering on Minnesota Road with respect to selected cluster count K. Colors represent different clusters.

one sees it saturates at $K = 7$, and the user also observes a noticeable increase in scaled spectrum energy when $K = 7$. Therefore, the algorithm can be used to incrementally generate four clustering results for $K = 7, 8, 9$, and 10. The selected clustering results in Fig. 7 are shown to be consistent with geographic separations of different granularity.

We also apply the proposed incremental user-guided spectral clustering algorithm (**Algorithm** 1) to Power Grid, CLUTO, Swiss Roll, Youtube, and BlogCatalog. In Fig.8, we show how the value of clustering metrics changes with K, for each dataset. The incremental method enables us to efficiently generate all clustering results with $K = 2, 3, 4 \ldots$ and so on. Due to space limitation, for each dataset we only show the trend of the three clustering metrics that exhibit the highest variation for different K; thus, the chosen clustering metrics can be different for different datasets. This suggests that selecting the correct number of clusters is a difficult task and a user might need to use different clustering metrics for a range of K values, and Incremental-IO is an effective tool to support such an endeavor.

7. CONCLUSION

In this paper we present Incremental-IO, an efficient incremental eigenpair computation method for graph Laplacian matrices which works by transforming a batch eigenvalue



(a) Power Grid (b) CLUTO (c) Swiss Roll (d) Youtube (e) BlogCatalog

Figure 8: Three selected clustering metrics of each dataset. The complete clustering metrics can be found in the extended version (available at http://arxiv.org/abs/1512.07349).

decomposition problem into a sequential leading eigenpair computation problem. The method is elegant, robust and easy to implement using a scientific programming language, such as Matlab. We provide analytical proof of its correctness. We also demonstrate that it achieves significant reduction in computation time when compared with the batch computation method. Particularly, it is observed that the difference in computation time of these two methods grows exponentially as the graph size increases.

To demonstrate the effectiveness of Incremental-IO, we also show experimental evidences that obtaining such an incremental method by adapting the existing leading eigenpair solvers (such as, the Lanczos algorithm) is non-obvious and such efforts generally do not lead to a robust solution.

Finally, we demonstrate that the proposed incremental eigenpair computation method (Incremental-IO) is an effective tool for a user-guided spectral clustering task, which effectively updates clustering results and metrics for each increment of the cluster count.

Acknowledgments

This work is partially supported by the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0002534, by ARO grant W911NF-15-1-0479, by ARO grant W911NF-15-1-0241, and by Mohammad Hasan's NSF CAREER Award (IIS-1149851).

8. REFERENCES

[1] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *SDM*, volume 4, pages 333–344, 2004.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] D. Calvetti, L. Reichel, and D. C. Sorensen. An implicitly restarted lanczos method for large symmetric eigenvalue problems. *Electronic Transactions on Numerical Analysis*, 2(1), 1994.

[4] P.-Y. Chen and A. Hero. Deep community detection. *IEEE Trans. Signal Process.*, 63(21):5706–5719, Nov. 2015.

[5] P.-Y. Chen and A. Hero. Phase transitions in spectral community detection. *IEEE Trans. Signal Process.*, 63(16):4339–4347, Aug 2015.

[6] P.-Y. Chen and A. O. Hero. Node removal vulnerability of the largest component of a network. In *GlobalSIP*, pages 587–590, 2013.

[7] P.-Y. Chen and A. O. Hero. Assessing and safeguarding network resilience to nodal attacks. *IEEE Commun. Mag.*, 52(11):138–143, Nov. 2014.

[8] C. Dhanjal, R. Gaudel, and S. Clémençon. Efficient eigen-updating for spectral graph clustering. *Neurocomputing*, 131:440–452, 2014.

[9] P. Jia, J. Yin, X. Huang, and D. Hu. Incremental laplacian eigenmaps by preserving adjacent information between data points. *Pattern Recognition Letters*, 30(16):1457–1463, 2009.

[10] J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122, 1992.

[11] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4), 1950.

[12] R. M. Larsen. Computing the svd for large and sparse matrices. *SCCM, Stanford University, June*, 16, 2000.

[13] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, volume 6. Siam, 1998.

[14] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007.

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

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2002.

[17] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. S. Huang. Incremental spectral clustering with application to monitoring of evolving blog communities. In *SDM*, pages 261–272, 2007.

[18] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. S. Huang. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113–127, 2010.

[19] R. Olfati-Saber, J. Fax, and R. Murray. Consensus and cooperation in networked multi-agent systems. *Proc. IEEE*, 95(1):215–233, 2007.

[20] B. N. Parlett. *The symmetric eigenvalue problem*, volume 7. SIAM, 1980.

[21] M. Polito and P. Perona. Grouping and dimensionality reduction by locally linear embedding. In *NIPS*, 2001.

[22] L. K. M. Poon, A. H. Liu, T. Liu, and N. L. Zhang. A model-based approach to rounding in spectral clustering. In *UAI*, pages 68–694, 2012.

[23] A. Pothen, H. D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 11(3):430–452, 1990.

[24] F. Radicchi and A. Arenas. Abrupt transition in the structural formation of interconnected networks. *Nature Physics*, 9(11):717–720, Nov. 2013.

[25] G. Ranjan, Z.-L. Zhang, and D. Boley. Incremental computation of pseudo-inverse of laplacian. In *Combinatorial Optimization and Applications*, pages 729–749. Springer, 2014.

[26] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[27] 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.*, 30(3):83–98, 2013.

[28] S. White and P. Smyth. A spectral clustering approach to finding communities in graph. In *SDM*, volume 5, pages 76–84, 2005.

[29] K. Wu and H. Simon. Thick-restart lanczos method for large symmetric eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 22(2):602–616, 2000.

[30] M. J. Zaki and W. M. Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[31] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004.

[32] B. Zhang, S. Choudhury, M. A. Hasan, X. Ning, K. Agarwal, and S. P. andy Paola Pesantez Cabrera. Trust from the past: Bayesian personalized ranking based link prediction in knowledge graphs. In *SDM MNG Workshop*, 2016.

[33] B. Zhang, T. K. Saha, and M. A. Hasan. Name disambiguation from link data in a collaboration graph. In *ASONAM*, pages 81–84, 2014.