Spectral Embedding Norm: Looking Deep into the Spectrum of the Graph Laplacian

Xiuyuan Cheng and Gal Mishne

Abstract

The extraction of clusters from a dataset which includes multiple clusters and another significant portion of “background” samples is a task of practical importance. The traditional spectral clustering algorithm, relying on the leading $K$ eigenvectors to detect the $K$ clusters, fails in such cases. This paper proposes the spectral embedding norm which sums the squared values of the first $I$ (normalized) eigenvectors, where $I$ can be larger than $K$. We prove that the quantity can be used to separate clusters from the background under generic conditions motivated by applications such as anomaly detection. The performance of the algorithm is not sensitive to the choice of $I$, and we present experiments on synthetic and real-world datasets.

1 Introduction

In unsupervised data analysis and machine learning, one of the most common goals is to group the data points into clusters. A variant task is to extract interesting clusters from the data when, in practice, data points do not perfectly fall into $K$ clusters. Often, data points can be roughly divided into two parts, one consists of interesting sub-groups, namely “clusters”, and the other one contains points which are less structured or of less interest, which we call “background”. Theoretically, one may view the background component as an extra cluster, and this poses challenges for traditional clustering algorithms. In this paper, we consider the situation where the background takes a significant portion of the data compared to the size of the clusters. This unbalanced case is a classical scenario where traditional spectral clustering fails to identify the existing clusters [1–3]. For example, the task of anomaly (or outlier) detection is such a case, where anomalous samples (small clusters) in the dataset differ from the normal ones (background) and indicate that something important has happened or a problem has occurred. By the very nature of the problem, most data points belong to the background and only a small fraction of data points are anomalies. Other examples include image segmentation and saliency detection where the clusters are certain regions of interest to detect, and background consists of the rest of the image, see applications in [4,5].

As a popular clustering method, the spectral clustering algorithm [6–8] reduces the dimensionality of the data using a spectral embedding, and then performs clustering in the low-dimensional space. The method originally proposed to cluster data into $K$ clusters with the leading $K$ eigenvectors (the low-lying eigenvectors of the graph Laplacian) computed from an affinity matrix built from the data [6,7,9,10]. A question is then how to set the parameter $K$, and this is especially important for exploratory data analysis when the number of clusters underlying the data is not known a priori. The traditional solution is to use the spectral gap of the eigenvalues to determine $K$ [11], yet in practical settings, such a gap may not exist. In particular, as has been pointed out by [1] and others, when the cluster sizes are unbalanced, or a large background component is present, there is no spectral gap after the $K$-th eigenvalue and the leading eigenvectors tend to be supported mostly on the large component due to the slow mixing time of the diffusion process restricted to it. This is indeed the case for anomaly detection, which poses a challenge for spectral methods. In practice, spectral embeddings have been used for anomaly detection in a modified way, based on the first non-trivial eigenvector of an affinity matrix [12,13], eigenvector selection [14,15].
out-of-sample extension [16][18], and multiscale constructions of spectral embeddings [16][19], and usually requiring tuning of multiple parameters.

In this paper we consider a cluster-background splitting model of the graph, including anomaly detection as a special case. The model is motivated by applications, and experiments on real-world datasets will be shown. We propose a quantity called the spectral embedding norm, which maps each node in the graph to a positive number, and separates clusters from background with a theoretical guarantee. The idea is closely related to the “localization” pattern of the eigenvectors, namely where they are supported on - specifically, either mainly on the cluster block or the background block - and this pattern maintains even when the spectral gap vanishes. Viewing the affinity matrix as a perturbed one from a baseline affinity where the background and clusters are completely disconnected, one can analyze the consequent deformation of the spectrum of the graph Laplacian matrix. However, the instability of eigenvectors under the deformation poses difficulty to the use of individual eigenvectors in this environment. The spectral embedding norm, on the other hand, improves the stability by using a summation over multiple eigenmodes and provides guaranteed detection of the clusters by a simple thresholding of the quantity. The algorithm involves a parameter which is the number of eigenvectors summed over, and we will show that the method is not sensitive to the choice of this parameter.

This paper is organized as follows. We introduce the spectral embedding norm in Section 2, starting with a toy example which illustrates the main idea of the paper. The theoretical result is presented in Section 3, including extensions of the basic result. Experiments on synthetic and real-world datasets are shown in Section 4, proofs in Section 5 and further remarks in the final section.

1.1 Related works

As spectral clustering and variants have been intensively studied in literature, we only list the most relevant works and the references given here may not be exhaustive.

Spectral clustering and limitations. Different papers have highlighted the limitations of spectral clustering, mainly that, in practice, the top $K$ eigenvectors may not localize on $K$ given clusters, causing spectral clustering to fail [1][9]. Previous works have analyzed the limitations and successes of spectral clustering in view of a diffusion process in a multi-well potential or a mixture of Gaussians model [1][10][20], or nonparametric mixture model [21]. While Schieber et al. [21] showed that under certain conditions, the embedded points lie in an orthogonal cone structure and $k$-means succeeds in clustering the data, Nadler and Galun [1] showed that even for well separated Gaussians the top $K$ eigenvectors do not necessarily localize on $K$ clusters. Different methods, like [9][22], attempted to align the eigenvectors axes with the different clusters and improve the robustness of cluster identification. Zelnik-Manor and Perona [9] proposed to estimate the number of clusters from the eigenvectors instead of from the spectral gap, and empirically demonstrated improved performance when a background cluster is present (called “clutter”), while Damle, Minden and Ying [22] considered the case of balanced block-like affinity matrix. The embedding norm studied in the current work differs from the above approaches, and it involves a simple algorithm with theoretical guarantee under the specified settings.

Selection of eigenvectors. Several papers proposed to perform eigenvector selection to determine eigenvectors that localize on clusters or specifically on anomalies, using unsupervised spectral ranking [13], kurtosis [15], relevance learning [3], entropy [2][23], the L1 norm [14], and local linear regression [24]. Both [2][3] proposed calculating $Km$ eigenvectors where $m > 1$ and then select informative eigenvectors. The theoretical analysis of [15] focused on the adjacency matrix of a graph, while [14] considered the modularity matrix of a graph. We will analyze the spectrum of the (normalized) random-walk Laplacian matrix which has a more stable spectrum with finite samples [23]. In [26], the authors presented a two-level tensor product construction to model the localization of eigenvectors on low-order eigenvectors and produced empirical improvements, but no theoretical analysis was provided to explain the model.

Spectral methods for anomaly detection. The following papers have employed spectral embeddings for anomaly detection in different applications: in images [16][27], fraud detection [13], in graphs [14][15],...
sensor networks [19], computer systems [12] and in general applications [17, 18, 28]. For a general review of anomaly detection methods, the interested reader is referred to [29–31]. As has been shown in [13–15, 32], and we will demonstrate, the eigenvectors in an anomaly setting can lie deep within the spectrum of the affinity matrix. This gives rise to the notion of abandoning the guideline of focusing on $K$ eigenvectors and rather choosing to look deeper into the spectrum, both for clustering and for anomaly detection. Similar to the embedding norm which we propose, the square sum of principle components (called “leverage score”) has been used in a different setting as a indicator of outlier sample points for linear regression and other applications in statistics [33].

**Notations.** $|\cdot|$ stands for the cardinal number of a set. $A^c$ means the complement of a set $A$.

## 2 Spectral Embedding Norm

### 2.1 Cluster-Background splitting in the graph

Given $n$ data points in the feature space, an undirected weighted graph can be constructed which has $n$ nodes, denoted by $V$, and the weight on edge $(x, y)$ is the affinity between nodes $x$ and $y$ denoted by $W(x, y)$. $W(x, y) = W(y, x) \geq 0$. $W$ is an $n$-by-$n$ real-symmetric matrix of non-negative entries, called the graph affinity matrix. In applications, $W$ is built as a pairwise affinity between data points in a feature space, e.g., $W(x, y) = k(x, y)$, where $k$ is a symmetric kernel function applied to the feature vectors of data points $x$ and $y$. In our analysis we assume that $W$ has been constructed.

Suppose that $V$ can be divided into two disjoint subsets, background and clusters, denoted by $B$ and $C$ respectively. The typical scenario which we consider is when data points in $C$ are concentrated in the feature space and well-clustered into $K$ sub-clusters, whereas those in $B$ can be “manifold-like” and spread over the space. The precise assumptions will be formulated in terms of the graph-Laplacian spectra constrained to the subgraphs of $C$ and $B$ (see Assumption 1). We also assume the connections between $C$ and $B$ are weak. As a result, the submatrix of $W$ constrained to $C$ is close to having $K$ blocks, and is almost separated from the submatrix of $B$. Define matrix $W_0$ by removing all the connections between $B$ and $C$ from $W$, i.e. $W_0$ is a block-diagonal matrix consisting of two blocks of $C$ and $B$ respectively. We introduce a pseudo-dynamic parametrized by time $t$ as

$$W(t) = W_0 + tE, \quad t \in [0, 1]$$

so that $W(0) = W_0$ and $W(1) = W$. The analysis is of the deformation of the spectrum of $W(t)$ (specifically, that of the normalized graph Laplacian) over time $t$.

To simplify the analysis, we assume that the $K$ sub-clusters are of equal size, that is,

$$|C| = \delta |V|, \quad |B| = (1 - \delta) |V|,$$

and each of the $K$ clusters in $C$ has $\frac{\delta |V|}{K}$ nodes. Our result extends to the unequal-size case (Section 3.3).

### 2.2 Graph Laplacian and embedding norm

Consider the random-walk graph Laplacian of $W$

$$L = I - D^{-1}W := I - P,$$

where $D$ is a diagonal matrix defined by $D_{ii} = \sum_j W_{ij}$, and $P$ is a Markov matrix. We shall see that $D$ is always invertible. $P$ is similar to $D^{-1/2}WD^{-1/2}$ which is real-symmetric, thus $P$ is diagonalizable and has $n$ real eigenvalues. Let

$$P\psi_k = \lambda_k \psi_k, \quad k = 1, \ldots, n, \quad \psi_k^T D \psi_j = \delta_{kj},$$

(2)
where \( \{\lambda_k\}_k \) are the eigenvalues of \( P \), \( \{\psi_k\}_k \) are the corresponding right eigenvectors and \( \delta_{kj} = 1 \) when \( k = j \) and 0 otherwise. Given that \( |\lambda_k| \leq 1 \) (Perron-Frobenius theorem), and when \( W \) is a positive-definite kernel matrix then all the eigenvalues are between 0 and 1. The largest eigenvalue is 1 and the associated eigenvectors is the constant vector. Note that time dependence is omitted in the above notations: As we introduced the deformation of \( W \) in \([1]\), \( D, P \) and consequently \( \psi_k \) and \( \lambda_k \) also depend on \( t \). We assume that at \( t = 1 \) the eigenvalues are sorted to be decreasing, and for other \( t \), the indexing \( k \) is arranged so that \( \psi_k \) and \( \lambda_k \) are differentiable with respect to \( t \) (see Section \([3]\)).

The spectral embedding norm of every node \( x \in \mathcal{V} \) is defined to be

\[
S(x) = \sum_{k \in I} \psi_k(x)^2
\]

(3)

where \( I \) is a subset of the eigenvalue indices \( \{1, \cdots, n\} \). \( S(x) \) is the (squared) Euclidean norm of the embedded vector of \( x \) in the spectral embedding space using eigenvectors of indices in \( I \). When needed, we include the time dependence in the notation written as \( S(x,t), t \in [0,1] \). A typical choice of \( I \) is

\[I = \{1, \cdots, |I|\},\]

when the eigenvalues are sorted to be descending. The cardinal number \( |I| \) is chosen to be around \( K \) and up to a few multiples of \( K \), and in practice, estimates of \( K \) can be used if \( K \) is not known. We will show that the result is not sensitive to the choice of \( |I| \), in analysis and experiments.

Under certain assumptions, the embedding norm \( S \) is able to separate \( C \) from \( B \) by a provable margin (Theorem \([3,1]\)). A general weighted form of \( S \) is introduced in Section \([3,3]\) and the main result extends directly. The quantity (3) is naturally related to the diffusion distance \([3,4]\), we will elaborate more on this in the last section.

2.3 A prototypical toy example

The prototypical scenario which motivates the proposed method is illustrated in the toy example in Figure \([1]\) (a) shows data points in \( \mathbb{R}^2 \) consisting of two groups: a large group, denoted by \( \mathcal{B} \), which lie close to the the unit circle (blue) and a small one, denoted by \( \mathcal{C} \), which form a small cluster lying close to the circle (red), and \( \delta_{ij} = \delta = 0.01 \). Variations of this model involve multiple sub-clusters in \( \mathcal{C} \) (see Figure 4) and the application in Section \([1,3]\) and the qualitative picture is the same. The affinity matrix \( W \) built from the data is shown in (b). The first few eigenvalues, changing over time \( t \), are shown in (c), and the associated eigenvectors evaluated on two nodes, one in \( \mathcal{C} \) and one in \( \mathcal{B} \), are plotted in (d) and (e). We sort the eigenvalues of \( P \) from large to small, and the first eigenvalue is always 1. The embedding norm \( S(x) \) takes the squared-sum of the first 40 eigenvectors on each node (c.f. (3)) and is plotted in (f) over time.

The figure demonstrates that

1. Though two blocks \( \mathcal{C} \) and \( \mathcal{B} \) exists in the graph, there is no clear eigen-gap between the second and third eigenvalues. Actually, the leading eigenvalues are all very close to 1 throughout time 5 (the eigth eigenvalue is greater than 0.998).

2. While the second eigenvector \( \psi_2 \) distinguishes \( \mathcal{C} \) at short time \( t \), once \( t \) is greater than 0.01 \( \psi_2 \) fails to indicate the cluster \( \mathcal{C} \) ((d) shows the value on one node and it is typical for the value of \( \psi_k \)'s on \( \mathcal{C} \). The transition actually happens when the initial gap between \( \lambda_2 \) and \( \lambda_3 \) almost vanishes. However, the high-index eigenvectors may take large value on \( \mathcal{C} \) (the eigth eigenvector starts to take large magnitude on \( \mathcal{C} \) around \( t = 0.1 \), and the trend of high-indexed eigenvectors localizing on \( \mathcal{C} \) continues, which is not shown). This is evident by \( S(x) \) consistently distinguishing \( \mathcal{C} \) from \( \mathcal{B} \) over time (shown in (f)).
Figure 1: Plots of eigenvalue and eigenvectors of $P = D^{-1}W$ over time. (a) $n = 5000$ data points in $\mathbb{R}^2$ sampled on $B \cup C$, where points in $B$ lie close to a circle (blue) and points in $C$ form the small cluster lying close to the circle (red). In this case, $K = 1$, $\delta = 0.01$. (b) The affinity matrix $W(t)$ at $t = 1$, c.f. (1). (c) Plot of the first 8 eigenvalues as $t$ increases from 0 to 1 (excluding $\lambda_1 = 1$). (d) The absolute values of the associated first 8 eigenvectors at $x_1 \in C$ over time. (e) Same plot at $x_2 \in B$. (f) The values of the embedding norm $S(x)$ defined in (3) at $x_1$ and $x_2$ over time, where $|I| = 40$.

This suggests that when the leading eigenvectors fail to identify the cluster $C$, the information of the location of $C$ may be contained in higher-indexed eigenvectors, and looking “deep” into the spectrum may be helpful. However, the selection of “informative” eigenvectors is generally a challenging problem. In particular, as shown in (d)(e), the deformation of eigenvectors is not stable when eigenvalues get close, which makes it difficult to study them individually. Instead, we consider in the embedding norm the summation over a group of eigenvectors, and intuitively this should improve the stability over time. This is indeed the case here: The quantity $S$ varies smoothly over time and preserves a gap between $C$ and $B$, as shown in (f), thus one can detect $C$ from $B$ by thresholding the value of $S$. Note that 40 eigenvectors are used in the summation, which is much larger than 2. We will justify this in the analysis.

3 Analysis

At $t = 0$ in (1), the matrix $W_0$ has the two-block structure, and the spectrum of the graph Laplacian of $W_0$ also splits into two groups, one residing on $C$ and the other on $B$ respectively. However, as $t$ increases, interactions among the eigenvectors develop and the the perfect splitting pattern is no longer preserved. In this section we show that the embedding norm varies more stably than individual eigenvectors over time, and serves as a measure by which to separate $C$ from $B$ up to time $t = 1$.  


Figure 2: Diagram showing the evolution of eigenvalues of the Markov matrix as $W(t)$ changes over time as in (1). At $t = 0$, circles indicate $B$ eigenvalues, and crosses indicate $C$ ones. In this example, $K = 3$, and $|I| = 10$. Eigenvalues of the $B$-submatrix are shown in circles, and those of $C$-submatrix in crosses. Note that $B$ can have eigenvalues close to 1 even at $t = 0$. As $t$ increases, at most time the eigenvalues are all of multiplicity one. Eigen-crossings may happen within $I$ and $I^c$ but not in between, and the $I$-spectral gap denoted by $\Delta(t)$ is preserved (Proposition 2). The differential equation (19) is obtained by by the contour integral $\Gamma$, which exists for all $t$ due to positive $\Delta$.

3.1 Initial separation by $S$ and assumptions

Since we will use $S$ to separate $B$ and $C$, we need it to do so at least at $t = 0$ when the two blocks $B$ and $C$ are perfectly separated by removing all the edges connecting them. Note that this does not necessarily happen unless certain assumptions are made: Because the eigenvalues of the $B$ block can be close to 1 and the clustering in the $C$ block may not be perfect, the first $|I|$ eigenvectors may be supported either on $C$ or on $B$, and there is generally no guarantee that the squared sum (3) will distinguish the two blocks. The assumptions needed here are the following two:

1. At $t = 0$, the eigenvectors in $I$ which are supported on $B$ are sufficiently delocalized ("flat") and those on $C$ are close to the well-clustered case;
2. The fraction $\delta$ of $|C|$ is sufficiently small so that the eigenvectors on $C$ are of sufficiently larger magnitude than those on $B$, due to the eigenvector normalization (2). The precise condition depends on the choice of $|I|$, the node degrees and so on.

To proceed, we define the degree function to be

$$d(x, t) = \sum_{y \in V} W(x, y; t)$$

which depends on time $t$. We introduce the notation $\nu(A, t)$ to stand for the "volume" of set $A$ at time $t$, where "volume" means the sum of the degrees,

$$\nu(A, t) = \sum_{x \in A} d(x, t).$$

We also define lower and upper bounds

$$d_0 := \min_{x \in V} d(x, 0), \quad d_0 := \max_{x \in V} d(x, 0),$$

and assume that $d_0 > 0$. By construction (1), the degree $d(x, t)$ of any node monotonically increases over time. Thus $d_0$ is the universal degree lower-bound:

**Lemma 1.** For all $x \in V$ and all $0 \leq t \leq 1$, $d(x, t) \geq d_0 > 0$.

At $t = 0$, since the affinity matrix falls into two separated blocks $B$ and $C$, so do the eigenvectors. We call them initial eigenvectors, and the set of eigenvectors which are only supported on $B$ are called the $B$-eigenvectors, denoted by $\Psi^B$, and similarly for $C$-eigenvectors and $\Psi^C$. The assumption on these eigenvectors and the index set $I$ is the following:
Assumption 1 (B and C-eigenvectors). At $t = 0$,

(a) The index $I$ includes $K$ C-eigenvectors and $|I| - K$ B-eigenvectors.

(b) Each of the $K$ eigenvectors in $I \cap \Psi^C$ (up to a $K$-by-$K$ rotation of these $K$ vectors) is associated with one of the $K$ clusters in the following sense: There exists $0 \leq \varepsilon_1 < 1$, and for each $\psi \in I \cap \Psi^C$, there is a unique $j$, $1 \leq j \leq K$, s.t.

$$
\frac{1 - \varepsilon_1}{\nu(C_j, 0)} \leq \psi(x)^2 \leq \frac{1 + \varepsilon_1}{\nu(C_j, 0)}, \quad \forall x \in C_j,
$$

$$
\psi(x)^2 \leq \frac{\varepsilon_1}{\nu(C, 0)}, \quad \forall x \in C \setminus C_j,
$$

(c) There exists $\varepsilon_2 \geq 0$, s.t. for any $\psi \in I \cap \Psi^B$,

$$
\psi(x)^2 \leq \frac{1 + \varepsilon_2}{\nu(B, 0)}, \quad \forall x \in B.
$$

The above assumption, while appearing to be complicated, poses only generic conditions on the subgraphs $B$ and $C$.

In the perfectly separated case the largest $K$ C-eigenvales are 1, and the $(K+1)$-th one is strictly less than 1 and depends on the mixing time of the Markov chain within each cluster. Since we primarily work with well-clustered $C$ which takes a small fraction of nodes and are localized in the graph (typically $C$ is an outlier cluster, or several localized regions of interest, see the applications in Section 4), this spectral gap is usually significant. As a result, even when the clustering is not perfect, the $(K+1)$-th C-eigenvalue is still sufficiently far away from the first $K$ ones, and they can be excluded from the index set $I$, since $I$ selects the largest $|I|$ eigenvalues. This fulfills (a).

If the $K$ clusters in $C$ are perfectly separated, one can verify that $\varepsilon_1 = 0$ in (b). Thus (b) holds when $C$ is not far from being well-clustered.

Assumption 1(c) requires that the eigenvector $\psi$ is sufficiently delocalized, or “flattened” on $B$: Recall that

$$
\sum_{x \in B} \psi(x)^2 d(x, 0) = 1,
$$

and the first eigenvector (associated with eigenvalue 1) takes the constant value $\psi(x)^2 = \frac{1}{\nu(B, 0)}$. If all the other eigenvectors are flattened, then (c) holds with some small $\varepsilon_2$. The delocalization widely applies when $B$ are built from data vectors lying on certain regular manifolds: assuming that the discrete eigenvectors well approximate the continuous limits which are eigenfunctions of the manifold Laplacian, the delocalization of the former inherit from that of the latter (quantum ergodicity theorem [35,36]). When the spectral convergence is poor, the finite-sample effects may create some localized pattern in the “noisy” eigenvectors, however, since $|I|$ is typically a small number compared to $n$, we assume that the selected B-eigenvectors are sufficiently close to the population ones.

The second assumption is the following: Recall that $\delta = \frac{|C|}{|V|}$.

Assumption 2. The constants $\delta$, $|I|$ and $K$ satisfy that

$$
\frac{\delta}{1 - \delta} \frac{|I| - K}{K} < \frac{d_0(1 - \varepsilon_2)}{d_0(1 + \varepsilon_1)},
$$

where $\varepsilon_1, \varepsilon_2$ are as in Assumption 1.
Proposition 1 (Initial separation by $S(x)$). Under Assumption 1 at time $t = 0$,
\[
\frac{1}{n\delta_0} K (1 - \varepsilon_1) \leq S(x) \leq \frac{1}{n\delta_0} K (1 + 2\varepsilon_1), \quad \forall x \in \mathcal{C}.
\]
(10)
\[
S(x) \leq \frac{1}{n\delta_0} \frac{(1 + \varepsilon_2)(|I| - K)}{1 - \delta}, \quad \forall x \in \mathcal{B},
\]
(11)
If furthermore, Assumption 2 holds, then
(1) The initial gap between $\mathcal{B}$ and $\mathcal{C}$ is at least
\[
g_0 := \frac{1}{n\delta_0} K (#), \quad (#) := \frac{d_0(1 - \varepsilon_1)}{d_0} - \frac{\delta}{1 - \delta} \frac{|I| - K}{K} (1 + \varepsilon_2) > 0,
\]
(12)
that is, $\forall x \in \mathcal{C}$ and $y \in \mathcal{B}$, $S(x) - S(y) \geq g_0 > 0$.
(2) At $t = 0$,
\[
\sup_{x \in \mathcal{V}} S(x) \leq \frac{1}{n\delta_0} \frac{K}{\delta} (1 + 2\varepsilon_1).
\]
(13)

3.2 Stability of the embedding norm under deformation

We will prove the stability of $S(x, t)$ over time making use of the Hadamard variation formula for the eigenvalues and eigenvectors, after properly indexing them. Specifically, since we assume that $d_0 > 0$, the diagonal matrix $D$ is invertible throughout time, and the Markov matrix $P = D^{-1}W$ is diagonalizable and similar to $D^{-1/2}WD^{-1/2}$. Under the matrix perturbation model (1) which is linear in $t$, the $n$ eigenvalues of the Markov matrix $P$ can be indexed as
\[
\lambda_1(t), \cdots, \lambda_n(t)
\]
so that they are descending at $t = 0$, i.e. $\lambda_{k+1}(0) \leq \lambda_k(0)$, and differentiable with respect to $t$ for $0 \leq t \leq 1$ (Chapter 2 of [37]). Similar to the classical Hadamard variation formula, the evolution equation of $\lambda_k$ can be shown to be
\[
\dot{\lambda}_k = \psi_k^T (\dot{W} - \lambda_k \dot{D}) \psi_k,
\]
(14)
and the equation of the associated eigenvector $\psi_k$ is, when valid,
\[
\dot{\psi}_k = -\frac{1}{2} (\psi_k^T \dot{D} \psi_k) \psi_k + \sum_{j \neq k} \frac{\psi_j^T (\dot{W} - \lambda_k \dot{D}) \psi_k}{\lambda_k - \lambda_j} \psi_j,
\]
(15)
that is, (15) holds on time intervals when no eigen-crossing of any pair of $\lambda_k$ and $\lambda_j$ happens. The derivation of (14), (15) is left to Appendix B.

Though the $n$ eigenvalues are ordered from large to small at $t = 0$, an eigen-crossing (or neighboring eigenvalues becoming very close) may happen as $t$ increases, as illustrated in the diagram in Figure 2 and numerically in the toy example in Figure 1. This voids a direct adoption of (15) unless one shows that the singularity does not affect the differentiability of the eigenvector branches before and after the crossing, which is still possible in our setting [37]. However, even if (15) can be made valid with such an effort, when an eigen-crossing or a near crossing happens there is generally no control on the speed of change of the associated pair of eigenvectors. Some steep changes of eigenvectors are shown in the toy example in Figure 1 at times of (near) eigen-crossings. This instability of eigenvectors under matrix perturbation underlies the main difficulty to justify the use of leading eigenvectors in this environment, not only for theoretical analysis but for algorithms as well.
The main observation of this work is to overcome such instability by considering the spectral embedding norm instead of individual eigenvectors. A key quantity needed in the stability bounds (of both the eigenvalues and the embedding norm) is the $C$-$B$ “connection strength”, measured by
\[ C := \sum_{x \in B, y \in C} W(x, y). \] (16)

The analysis needs $C$ to be a small compared to the magnitude of the node degrees, specifically, $\frac{C}{d_0}$ needs to be a small constant. We note that the condition may be much stronger than encountered in applications due to the reliance on a spectral gap between $I$ and $I^c$ eigenvalues. To be specific, we define the $I$-eigen-gap (depending on time $t$) to be
\[ \Delta(t) := \min_{i \in I, j \notin I} |\lambda_i(t) - \lambda_j(t)|, \quad t \in [0, 1]. \] (17)

Such an “$I$-eigen-gap” prevents eigenvalues from $I$ and $I^c$ to get too close, but allows arbitrary eigen-crossings within $I$ and $I^c$. While needed in the perturbation analysis, we note that $\Delta(t)$ should be viewed as an artifact due to the limitation of our theory (see Remark 1). However, this is essentially different from the traditional spectral gap assumed after the $K$-th eigenvalue.

The following proposition proves the preserved $I$-eigen-gap assuming an initial one, based upon the stable evolution of eigenvalues c.f. (14). Proof in Section 5.

**Proposition 2** (Preservation of $I$-eigen-gap). Under (A1), $C$ as in (16), if for some constant $\Delta > 0$, $\Delta(0) \geq 2\Delta$ and
\[ \frac{C}{d_0} \leq \frac{1}{8} \Delta, \] (18)

then
\[ \Delta(t) \geq \Delta, \quad \forall 0 \leq t \leq 1. \]

The significance of the preserved $I$-eigen-gap is that we can derive the evolution equation of the embedding norm $S(x, t)$ without being concerned with the eigen-crossings within $I$ (and within $I^c$). This is possible by relying on $S(x, t)$ being the $(x, x)$-th diagonal entry of the spectral projection matrix $P_t := \sum_{i \in I} \psi_i \psi_i^T$, which can be written in form of a contour integral of the resolvent in the complex plane where the contour circles the eigenvalues in $I$ throughout $t \in [0, 1]$, as illustrated in Figure 2. The evolution equation below only requires eigenvalue difference $\lambda_k - \lambda_j$ to be non-vanishing when one is from $I$ and the other is from $I^c$. Actually, this difference is bounded from below by the constant $\Delta$ by Proposition 2. Proof in Section 5.

**Proposition 3** (Evolution of $S$). When Proposition 2 applies, for $0 < t < 1$,
\[ \frac{\partial}{\partial t} S(x, t) = -\sum_{k \in I, j \in I} (\psi_j^T \hat{D} \psi_k) \psi_k(x) \psi_j(x) + 2 \sum_{k \in I, j \notin I} \frac{\psi_j^T (\hat{W} - \lambda_k \hat{D}) \psi_k}{\lambda_k - \lambda_j} \psi_k(x) \psi_j(x). \] (19)

We are ready to derive the main result, proved in Section 5.

**Theorem 3.1** (Separation at $t = 1$). Under (A1)-(A3), if for some constant $\Delta > 0$, the following conditions are satisfied:
(i) $\Delta(0) \geq 2\Delta$, $\Delta(t)$ as in (17),
(ii) $C$ as in (16),
\[ \frac{C}{d_0} \leq \frac{\Delta}{8} \left( 1 - \frac{\frac{\#}{2}}{1 + 2\epsilon_1} \right) \] (20)

where $(\#)$ is defined in (12), and $(\#) > 0$ under Assumption 3.
The main result Theorem 3.1 can be directly extended to cover the following cases, and proofs omitted:

**3.3 Extension of the results**

more restrictive than what occurs in practical applications. Analysis should be able to relax the constraint. This also means that the requirement on \( C \) for \( k \)

Remark 1. In practice, the requirement of a spectral gap \( \Delta \) between the \(|I|\)-th and \(|I|+1\)-th eigenvalue may not be satisfied, yet the embedding norm \( S \) can still produce a good separation of \( C \) from \( B \), see e.g. Fig. 4. The is because that the analysis is not tight: for one thing, the relaxation of the term \( \frac{\Delta}{8} \frac{1+\frac{d}{5}}{1+2\varepsilon_1} \) is crude, and can be improved, e.g., under proper assumptions of the eigenvalue distribution. In this aspect, the need for an I-eigen gap in the theory is still somehow artificial, and further analysis should be able to relax the constraint. This also means that the requirement on \( C \) in (20) may be more restrictive than what occurs in practical applications.

3.3 Extension of the results

The main result Theorem 3.1 can be directly extended to cover the following cases, and proofs omitted:

1. **Weighed embedding norm.** The definition of the embedding norm \( S \) can be generalized as

\[
S(x) = \sum_{k \in I} f(\lambda_k)\psi_k(x)^2,
\]

where \( f(\lambda) \) is a (complex) analytic function which is real-valued on real \( \lambda \). We have been addressing the special case where \( f = 1 \). The results extend to analytic functions \( f(\lambda) \), e.g., \( f(\lambda) = \lambda^p \) where \( p \) is a positive integer. To see this, consider the contour integral of \( f(z)R(z) \), \( R \) being the resolvent (defined in (37)), and then the time-evolution equation of \( S(x,t) \) can be shown to be

\[
\dot{S}(x) = 2 \sum_{k \in I, j \notin I} \frac{f(\lambda_k)}{\lambda_k - \lambda_j} \psi_k^T (W - \lambda_k \hat{D}) \psi_j \psi_k(x) \psi_j(x)
\]

\[
+ \sum_{k,j \in I, k \neq j} \left( \frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} \psi_k^T W \psi_j - \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} \psi_k^T \hat{D} \psi_j \right) \psi_k(x) \psi_j(x)
\]

\[
+ \sum_{k \in I} (f'\lambda_k) \psi_k^T W \psi_k - (zf(z))' (\lambda_k) \psi_k^T \hat{D} \psi_k \psi_k(x)^2,
\]

where in case that the eigenvalues \( \lambda_k \) and \( \lambda_j \) coincide, the term \( \frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} \) is replaced by \( f'(\lambda_k) \) and \( \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} \) by \( (zf(z))' (\lambda_k) \). So the r.h.s. is well-defined when an eigen-crossing within \( I \) happens, and
the terms \( \frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} \) and \( \frac{\lambda_k f(\lambda_k) - \lambda_j f(\lambda_j)}{\lambda_k - \lambda_j} \) are uniformly bounded due to the analyticity of \( f \). When \( f = 1 \), the equation reduces to [19]. Proceeding with the same technique as in the proof of the main result, the deformation bound of \( S(x, t) \) will then involve constant factors which depend on the boundedness of \( f \) and \( f' \) on \([0, 1]\). Specifically, the constant \( C \) will need to be redefined to be \((e_1 + \frac{\epsilon_2}{\delta \delta'})^2C \) where \( e_1 \) and \( \epsilon_2 \) are absolute constants. E.g., when \( f(\lambda) = \lambda^p \), \( p > 0 \), \( e_2 \) remains \( 4 \) (which is the dominating term with small \( \Delta \)) and \( e_1 = (p + 1) \). We compare to diffusion distance [34] in the last section.

2. Unequal cluster size in \( \mathcal{C} \). The requirement of equal cluster size of the \( K \) clusters in \( \mathcal{C} \) can be relaxed. Specifically, suppose that the \( K \) clusters have varying sizes \( |C_j| = \delta_jn \), and \( \sum_{k=1}^{K} \delta_j = \delta \). Let \( \delta_{\min} = \min_{1 \leq j \leq K} \delta_j \), and similarly define \( \delta_{\max} \). Then under Assumption 1 [10] and [11] become

\[
\frac{1 - \epsilon_1}{nd_0 \delta_j} \leq S(x, 0) \leq \frac{1}{nd_0} \left( \frac{\epsilon_1}{\delta/K} + \frac{1 + \epsilon_1}{\delta_j} \right), \quad \forall x \in C_j \text{ for some } j.
\]

\[
S(x, 0) \leq \frac{1}{nd_0} \left( \frac{1 + \epsilon_2 (|I| - K)}{1 - \delta} \right), \quad \forall x \in \mathcal{B},
\]

Define

\[
g_{j,0} := \frac{1}{nd_0} \left( \frac{d_0 (1 - \epsilon_1) - (1 + \epsilon_2 (|I| - K))}{1 - \delta} \right), \quad j = 1, \ldots, K \tag{25}
\]

and the minimum of \( g_{j,0} \) is

\[
g_{\min,0} = \min_{1 \leq j \leq K} \left( \frac{d_0 (1 - \epsilon_1) - (1 + \epsilon_2 (|I| - K))}{1 - \delta} \right).
\]

Modify Assumption 2 to be that \( g_{\min,0} > 0 \), then the initial separation of \( S(x, 0) \) on \( \mathcal{C} \) and \( \mathcal{B} \) is at least \( g_{\min,0} \) (and more precisely \( g_{j,0} \) between \( C_j \) and \( \mathcal{B} \)), and [13] becomes

\[
S(0) = \sup_{x \in \mathcal{V}} S(x) \leq \frac{1}{nd_0} \left( \frac{\epsilon_1}{\delta/K} + \frac{1 + \epsilon_1}{\delta_{\min}} \right).
\]

Note that Proposition 2, Proposition 3 and claims (1) (2) in the proof of Theorem 3.1 do not rely on Assumption 1 or Assumption 2 and are valid. As a result, it can be shown that the \( t = 1 \) separation between \( \mathcal{C} \) and \( \mathcal{B} \) by \( S \) holds as long as

\[
g_{\min,0} \geq 2(e^{\tilde{C}} - 1) \frac{1}{nd_0} \left( \frac{\epsilon_1}{\delta/K} + \frac{1 + \epsilon_1}{\delta_{\min}} \right), \quad \tilde{C} = \left( 1 + \frac{4}{\Delta} \right) \frac{2C}{d_0} \tag{27}
\]

This condition is more restrictive when the cluster sizes in \( \mathcal{C} \) are less balanced, namely when the difference \( \delta_{\max} - \delta_{\min} \) becomes larger. In our numerical experiments, all the sub-clusters are of comparable sizes (in the outlier detection in images, \( K = 1 \) or \( 2 \), and in image segmentation the clusters are of similar sizes), while we note that an extremely unbalanced cluster size, e.g., very small \( \delta_{\min} \), could affect the performance of the method.

3. Detection of parts of \( \mathcal{C} \). The above argument leads to a “personalized” detection condition for each cluster \( C_j \) in \( \mathcal{C} \), that is, even when \( S(x, 1) \) fails to separate some clusters in \( \mathcal{C} \) from \( \mathcal{B} \) it may still successfully detect the part. To see this, note that the proof of the theorem actually gives the following: For any subsets \( E_1 \) and \( E_2 \) of \( \mathcal{V} \), if

\[
\min_{x \in E_1} S(x, 0) - \max_{x \in E_2} S(x, 0) \geq g_{E_1, E_2} > 0
\]

then \( E_1 \) and \( E_2 \) can be separated by \( S(x, 1) \) by some threshold as long as

\[
g_{E_1, E_2} \geq 2(e^{\tilde{C}} - 1)S(0).
\]
Figure 3: Detection of $\mathcal{C}$ from a manifold-like $\mathcal{B}$: (a) $n = 5000$ data points in $\mathbb{R}^2$ sampled on $\mathcal{B} \cup \mathcal{C}$, where $\mathcal{B}$ points lie close to a circle (blue) and $\mathcal{C}$ points form $K = 2$ clusters lying nearby (red), $\delta = 0.02$. (b) The affinity matrix $W(t)$ at $t = 1$, c.f. (1). (c) The first 100 eigenvalues of the Markov matrix. (d) The plot of $S$. (e) $k$-th Eigenvectors of multiple $k$’s of the Markov matrix. (f) F1-score of the detection of $\mathcal{C}$ by thresholding the values of $S_I$, where $|I|$ varies from 2 to 100, and for multiple choice of self-tuning parameter ($k$-nearest neighbor in self-tuning, denoted by $k_{ST}$). Mean and standard deviation of F1-score are shown, and optimal value of $|I|$ are indicated by a red cross.

The previous results corresponds to $E_1 = \mathcal{C}$ and $E_2 = \mathcal{B}$. Let $E_i$ be any individual cluster $C_j$, then since $\bar{S}(0)$ is upper-bounded by (26), we have that each cluster $C_j$ can be separated from $\mathcal{B}$ by $S(x, 1)$ if $g_{j,0}$ as defined in (25) is larger than the r.h.s. of (27).

4. Initial inclusion of $I$. The Assumption (1)(a) can be relaxed by only requiring $K' \mathcal{C}$-eigenvectors in $I$, $0 < K' \leq K$, as long as they contribute to a sufficiently large $S(x, 0)$ on $\mathcal{C}$, or any subset of $\mathcal{C}$ such as an individual cluster $C_j$. The separation guarantee at time 1 follows the same argument as in item 2. above, where the quantities (23) (24) and consequently (25) (26) need to be modified. The precise condition is not pursued here. In practice, this means that even if less than $K$ “nearly” $\mathcal{C}$-eigenvectors are included in $I$, the method may still be able to detect part of $\mathcal{C}$ from $\mathcal{B}$.

4 Experiments

In this section we will apply the spectral embedding norm to both synthetic and real-world datasets, in scenarios of both single outliers and multiple clusters in a cluttered background.
4.1 Manifold data toy example

We begin with a simulated dataset composed of a manifold-like background $B$ and clusters in $C$ following the model: in $\mathbb{R}^2$, $B$ consists of i.i.d samples $x_i$ distributed as $x_i = y_i + n_i$, where $y_i$ are uniformly distributed on the unit circle, which is a one-dimensional manifold, and $n_i \sim N(0, \epsilon_B^2 I)$, with $\epsilon_B = 0.01$. $C$ contain $K$ equal-sized sub-clusters, each has i.i.d. samples drawn from $N(\mu_j, \epsilon_C^2 I)$, where $\mu_j$ are centered close to the circle, and $\epsilon_C = 0.02$. $n = 5000$ points are generated, and the number of points in $C$ is set to be $\delta n$ for positive $\delta$, rounded so that each sub-cluster has integer many points. To measure the accuracy of the detection of $C$ we compute the F1 score:

$$F1 = \frac{2pr}{p + r}, \quad p := \frac{TP}{TP + FP}, \quad r := \frac{TP}{TP + FN}.$$  

Figure 3 shows results for $K = 2$, $\delta = 0.02$. For $K = 2$, the eigenvalues do not reveal any clear eigen-gap, and the first 5 eigenvectors do not all give a clear indication of the cluster $C$, but contain a significant amount of “component” from $B$-eigenvectors. Examining the first $k \sim 20$, certain eigenvectors are more localized on $C$, e.g. $k = 8$ and 10 in this example. The embedding norm $S_I$ clearly separates $C$ from $B$. The results are not sensitive to parameter choices. Let $k_{ST}$ be the $k$-nearest neighbor used to set the local self-tuning scale $[9]$ in constructing the affinity matrix $W$. Then, throughout varying values of the parameter $k_{ST}$, the F1-score of the detection by thresholding $S$ reveals a “plateau” of valid values of $|I|$, e.g., when $k_{ST} = 8$, the range of $|I|$ is about $18 \sim 30$, with the optimal F1 score obtained at $|I| = 23$. The best F1 score for $k_{ST} = 4, 8$ or 16 are all greater than $0.99$.

Similar results are obtained for $K = 10$, $\delta = 0.1$, as shown in Figure 4. The condition in Assumption 2, $|I| \sim |I|$ suggests that $|I|$ is chosen to be proportional to $\frac{K}{\delta}$, and this is revealed in Figure 3 and Figure 4 (in these two examples $\frac{K}{\delta}$ is kept to be the same) as the “plateaus” of valid $|I|$ are at about the same range, across values of $k_{ST}$.
4.2 Side-scan sonar imagery

Anomaly detection can be seen as a special case of clustering in which there is a vast imbalance in the size of clusters, i.e., background vs. anomaly, and the density of each cluster. In image anomaly detection, the goal is to detection a small region (subset of pixels) that differs from the “normal image background. This is an important task in a variety of applications such as target detection in remote sensing imagery, detecting abnormalities such as tumors in biomedical imagery and for quality inspection in production lines. An automated solution highlighting only suspicious regions to be reviewed by an expert would save greatly on time. We demonstrate on real-world images that eigenvectors localizing on the anomaly can be buried deep within the spectrum of the image, and that by calculating the spectral embedding norm we can identify the anomalies.

We examine two side-scan sonar images containing a single sea-mine, displayed in Fig. 5(a) and (c),
Figure 7: Spectral embedding norm $S_I$ for the side-scan sonar image in Fig. 5(a), with increasing $|I|$. The sea-mine first “appears” for $|I| = 6$. For $|I| = 13,50$ the sea-mine separates cleanly from the background, however as $|I|$ increases to 100 spectral norm starts to reveal background components as well.

Figure 8: Same figure as Figure 7 for the side-scan sonar image in Fig. 5(c). The target sea-mine (indicated by the red circle in Fig. 5(c)) first “appears” for $|I| = 5$, however it then recedes into the background as $|I|$ increases to 7. For $|I| = 50,100$, this sea-mine is well separated from the background.

where we consider the sea-mine to be an anomaly (indicated by a red circle). The first image contains a noisy periodic background, and the presence of the sea-mine is indicated by its shadow which appears as a dark cluster. The second image contains a more varying background and the sea-mine appears as a bright highlight with dark shadow to its right. To construct the graph we extract all overlapping patches of size 8, and build a nearest neighbor graph with 64 neighbors and set $k_{ST} = 32$. In Fig. 5(b) and (d) we display $S(x)$ for $|I| = 50$, reshaped as an image, separating the sea-mine from the background.

In Fig. 6 we plot examples of eigenvectors of the Laplacian, where the eigenvectors in the top row localize on the background $B$, revealing its periodic nature at different scales and orientations, while the eigenvectors in the bottom row localize on the sea-mine $C$. To demonstrate the stability of $S(x)$, in
Figures 7-8 we plot $S(x)$ for increasing values of $|I|$, demonstrating that for both images, the sea-mine is revealed consistently for a wide range of values, while the background is suppressed. Note that in both cases this requires looking deep enough in the spectrum, and the first few eigenvectors bring out background structures.

4.3 Calcium Imaging

Calcium imaging is an experimental method in neuroscience that enables imaging the individual activity of hundreds of neurons in an awake behaving animal, at cellular resolution [38]. The acquired data is composed of a spatiotemporal volume, where, after motion correction, the neuron locations are fixed and the temporal activity consists of hundreds to tens of thousands of time-frames. There is also varying temporal activity in the background (neuropil). Thus, this data can also be viewed as an image whose pixels lie in a high-dimensional space (time-frames), consisting of hundreds of clusters (neurons) in an image plane with a non-trivial background, which matches our problem setting.

It is of interest to provide a 2D image that depicts the structure that exists in this volume and highlights the existing neurons. Such images serve for manual segmentation, to align imaged volumes across days (where the field of view may shift), to display neurons detected by automatic and manual means, etc. Several ROI extraction algorithms are also initialized based on such images [39,40]. Typically the temporal correlation image is used [41], or the temporal mean image. Here we show that the spectral embedding norm provides a meaningful visualization of the data, with sharp morphology and suppression of noise from the background clutter.

We analyze a publicly available dataset from Neurofinder [42]. The images are $512 \times 512$ pixel and 8000 time frames have been recorded at 8 Hz. Ground truth labels provided with the dataset includes 197 identified neurons, however note that recent papers point out that the ground truth on Neurofinder datasets is probably lacking, i.e. not all neurons are labeled [39,43]. To calculate the affinity matrix, we calculate a nearest neighbor graph for all pixels, represented as high-dimensional vectors in time, using 50
nearest neighbors. To accelerate the nearest neighbor search dimensionality is reduced from 8000 to 300 using PCA. In Fig. 9 we display examples of eigenvectors from both the background (top) and localizing on single neurons (bottom).

Figure 10 compares the spectral embedding norm (right) to the temporal mean (left) and temporal correlation image (middle). In each image, the values (mean/correlation/norm) appear in the green channel, while we overlay in the red channel a mask of the ground truth labels that were manually selected (where the two overlap it appears as yellow). The mean image exhibits a strong background, while neurons appear as typical “donuts” [40]. In the correlation image, the background mostly appears as noise. In comparison, the background has been suppressed in the spectral embedding norm image, while neurons which are barely or not at all visible in the correlation image appear as bright clusters.

To quantify, the separation of background and clusters, we segment the spectral embedding norm image for increasing \(|I|\), and compare the overlap between the segmented clusters and the given ground-truth mask. In Fig. 11 we plot the F1 score and demonstrate a plateau of stable F1-score values around 200-250. We also display the image of \(S(x)\) for multiple values of \(|I|\), to demonstrate the property of the spectral embedding norm to detect parts of \(C\). Note that we are not performing clustering here, but rather demonstrating how the embedding norm can be used to separate meaningful structure from background clutter. Beyond visualization, this approach can then serve to remove the background, and focus only on the remaining clusters in \(C\), thus simplifying a clustering task.

5 Proofs

5.1 Proposition 1 (Initial separation by \(S\))

Proof of Proposition 1. It suffices to prove (10) and (11), because Assumption 2 implies that the r.h.s of (11) is strictly less than the l.h.s of (10) by \(g_0\), and then claims (1) and (2) directly follow.
Figure 11: F1-score of segmenting neurons from background based on the spectral embedding norm image, for increasing $|I|$ values. We can see that for a range of values (200-250) the F1 score plateaus, and then decreases as the number of included eigenvectors increases. We display $S(x)$ for $|I| = 20, 49, 100, 250$ to demonstrate how more and more clusters are revealed.

To prove (10): Note that for any $x \in C_j$, at $t = 0$, 

$$S(x) = \sum_{k \in I} \psi_k(x)^2 = \sum_{k \in I \cap \Psi} \psi_k(x)^2.$$ 

By Assumption 1 (a), up to a possible $K$-by-$K$ rotation among the $K$ eigenvectors in $I \cap \Psi^C$, we assume that $\psi_j$ is the eigenvector associated with the sub-cluster $C_j$, $j = 1, \ldots, K$, and then 

$$S(x) = \sum_{j=1}^{K} \psi_j(x)^2,$$ 

as the rotation preserves the squared sum. Furthermore, suppose that $x \in C_{j_x}$, Assumption 1 (b) gives 

$$\psi_j(x)^2 \in \frac{1}{\nu(C_{j_x}, 0)}[1 - \varepsilon_1, 1 + \varepsilon_1], \quad j = j_x,$$

$$\psi_j(x)^2 \leq \frac{\varepsilon_1}{\nu(C, 0)}, \quad j \neq j_x.$$ 

Plugging into (28), it shows that 

$$S(x) \leq \frac{1 + \varepsilon_1}{\nu(C_{j_x}, 0)} + (K - 1) \frac{\varepsilon_1}{\nu(C, 0)},$$

and together with $\nu(C_j) = \sum_{x \in C_j} d(x, 0) \geq d_0 |C_j| = d_0 \delta n / K$ for any $j$ (the $K$ sub-clusters are equal-size) and similarly $\nu(C) \geq d_0 |C| = d_0 \delta n$, it gives the upper bound in (10). Consider the lower bound, (28) continues as 

$$S(x) \geq \psi_{j_x}(x)^2 \geq \frac{1 - \varepsilon_1}{\nu(C_{j_x}, 0)}.$$
Combined with \( \nu(C_j) = \sum_{x \in C_j} d(x, 0) \leq \frac{d_0}{n} |C_j| = \frac{d_0}{n} \delta_n \kappa \) for any \( j \), this gives the lower bound in (10).

To prove (11): For any \( x \in B \), at \( t = 0 \),

\[
S(x) = \sum_{k \in I} \psi_k(x)^2 = \sum_{k \in I \cap \Psi^B} \psi_k(x)^2 \leq \sum_{k \in I \cap \Psi^B} \frac{1 + \varepsilon_2}{\nu(B, 0)} \quad \text{(by Assumption 1(c))}
\]

(29)

where \( \nu(B, 0) = \sum_{x \in B} d(x, 0) \geq d_0 |B| = d_0 n (1 - \delta) \). Meanwhile, \( |I \cap \Psi^B| = |I| - K \) by Assumption 1(a).

Plugging into (29), this proves (11).

\[\square\]

5.2 Proposition 2 (Preservation of the \( I \)-eigen-gap)

Proof of Proposition 2 We will establish that for any \( k = 1, \ldots, n \),

\[
|\lambda_k(t) - \lambda_k(0)| \leq \frac{4C}{d_0} t, \quad 0 \leq t \leq 1.
\]

(30)

Given that this inequality holds, then by (18),

\[
|\lambda_k(t) - \lambda_k(0)| \leq \frac{4C}{d_0} \leq \frac{1}{2} \Delta, \quad \forall 1 \leq k \leq n.
\]

(31)

This means that it is impossible for \( \Delta(t) < \Delta \): Otherwise, there exist \( |\lambda_{k_1}(t) - \lambda_{k_2}(t)| < \Delta \), where \( k_1 \in I \) and \( k_2 \notin I \), and then (31) implies that \( |\lambda_{k_1}(0) - \lambda_{k_2}(0)| < 2\Delta \) which contradicts the assumption that \( \Delta(0) \geq 2\Delta \).

It suffices to show (30) to finish the proof. To do so, we prove the following bound

\[
|\dot{\lambda}_k| \leq \frac{4C}{d_0}, \quad \forall k, \forall t.
\]

(32)

From (14),

\[
|\dot{\lambda}_k| = |\psi_k^T (\dot{W} - \lambda_k \dot{D}) \psi_k| \leq |\psi_k^T \dot{W} \psi_k| + |\lambda_k| |\psi_k^T \dot{D} \psi_k|.
\]

As \( |\lambda_k| \leq 1 \) (Perron-Frobenius), then

\[
|\dot{\lambda}_k| \leq |\psi_k^T \dot{W} \psi_k| + |\psi_k^T \dot{D} \psi_k|.
\]

(33)

If the following claim is true, then (32) follows directly from (33):

\[
|\psi_l^T \dot{W} \psi_k|, |\psi_l^T \dot{D} \psi_k| \leq \frac{2C}{d_0}, \quad \forall k, l = 1, \ldots, n, \forall 0 \leq t \leq 1.
\]

(34)

Proof of (34): To bound \( |\psi_l^T \dot{W} \psi_k| \), note that \( \dot{W} = E \), and then

\[
|\psi_l^T E \psi_k| \leq \sum_{x \in \mathcal{C}, y \in B} |W(x, y)| |\psi_l(x)||\psi_k(y)| + \sum_{x \in B, y \in \mathcal{C}} |W(x, y)| |\psi_l(x)||\psi_k(y)|.
\]

(35)

We use the following more relaxed bound

\[
|\psi_k(x)|^2 \leq \sum_{j=1}^n |\psi_j(x)|^2 = \frac{1}{d(x)} \leq \frac{1}{d_0}, \quad \forall x \in \mathcal{V} = \mathcal{B} \cup \mathcal{C}, \forall k, \forall t,
\]

(36)

where the last inequality relies on \( d(x) \geq d_0 \) throughout time (Lemma 1). The second equality relies on \( D \Psi^T = I \), thus \( \Psi^T D^{-1} \). Then (35) continues as

\[
|\psi_l^T E \psi_k| \leq 2 \sum_{x \in \mathcal{C}, y \in B} |W(x, y)| \frac{1}{d_0} = \frac{2C}{d_0}.
\]

(35)
For $|\psi^T_k \hat{D}_k|$, similarly,

\[
|\psi^T_k \hat{D}_k| \leq \sum_{x \in V} |\psi_l(x)||\psi_k(x)||d(x)| = \sum_{x \in V} |\psi_l(x)||\psi_k(x)|| \sum_y E(x, y)| 
\leq \frac{1}{d_0} \sum_{x \in F} \sum_{y \in E} W(x, y) + \frac{1}{d_0} \sum_{x \in E} \sum_{y \in B} W(x, y) = \frac{2C}{d_0},
\]

where the bound (36) is used to bound each $|\psi_l(x)|$ and $|\psi_k(x)|$ in the 2nd inequality.

Note that while time dependence has been omitted in all the notations, the above arguments hold throughout time $t \in [0, 1]$.

5.3 Proposition 3 (Evolution of $S$)

Proof of Proposition 3. As explained in the text, one may first establish the validity of (15) and then verify the formula (19) based on the former by observing the cancelation of terms. As an alternative approach, we use the contour integral of the resolvent.

For $z \in C$ and not an eigenvalue of $P = D^{-1}W$, define

\[ R(z) = (W - zD)^{-1} \]

where the time dependence is omitted. By that $P = \Psi \Lambda \Phi^T$, $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$, $\Phi = D\Psi$ and $\Psi^T \Phi = I$, one can verify the equivalent form of $R$ as

\[ R(z) = \Psi(\Lambda - zI)^{-1}\Psi = \sum_{k=1}^{n} \frac{\psi_k \psi_k^T}{\lambda_k - z}. \tag{37} \]

This means that

\[ P_I = \sum_{k \in I} \psi_k \psi_k^T = -\frac{1}{2\pi i} \oint_{\Gamma} R(z) dz, \]

where the contour $\Gamma$ is such that the eigenvalues in $I$ (I') stay inside (outside) $\Gamma$ throughout time $t$ (Figure 2), and such $\Gamma$ exists due to Proposition 2. Thus the above expression of $P_I$ holds for all time $t$, and as a result

\[ \dot{P}_I = -\frac{1}{2\pi i} \oint_{\Gamma} \dot{R}(z) dz. \]

By differencing both sides of \[ (W - zD)R = I, \]
one obtains that

\[ \dot{R} = -R(\dot{W} - z\dot{D})R. \]

This means that

\[
\dot{P}_I = \frac{1}{2\pi i} \oint_{\Gamma} R(z)(\dot{W} - z\dot{D})R(z) dz \\
= \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{2\pi i} \oint_{\Gamma} \frac{\psi_k^T (\dot{W} - z\dot{D}) \psi_l}{(\lambda_k - z)(\lambda_l - z)} \psi_k \psi_l^T dz \\
= \sum_{k=1}^{n} \sum_{l=1}^{n} (\alpha_{kl} \dot{W} - \beta_{kl} \dot{D}) \psi_l) \psi_k \psi_l^T, \tag{38}
\]
where
\[ \alpha_{kl} = \frac{1}{2\pi i} \int_{\Gamma} \frac{1}{(\lambda_k - z)(\lambda_l - z)} \, dz, \quad \beta_{kl} = \frac{1}{2\pi i} \int_{\Gamma} \frac{z}{(\lambda_k - z)(\lambda_l - z)} \, dz. \]

By Cauchy’s integral formula, one can verify the following:

(a) When \( k \in I, l \in I \), \( \alpha_{kl} = 0 \), \( \beta_{kl} = 1 \).

(b) When \( k \in I, l \notin I \), \( \alpha_{kl} = \frac{1}{\lambda_k - \lambda_l} \), \( \beta_{kl} = \frac{\lambda_k}{\lambda_k - \lambda_l} \).

(c) When \( k \notin I, l \in I \), \( \alpha_{kl} = \frac{1}{\lambda_k - \lambda_l} \), \( \beta_{kl} = \frac{\lambda_l}{\lambda_k - \lambda_l} \).

(d) When \( k \notin I, l \notin I \), \( \alpha_{kl} = 0 \), \( \beta_{kl} = 0 \). Then (38) continues as

\[
\dot{P}_t = \sum_{k \in I, l \in I} - (\psi_k^T \dot{D} \psi_l) \psi_k \psi_l^T + \sum_{k \in I, l \notin I} \frac{\psi_k^T (\dot{W} - \lambda_k \hat{D}) \psi_l}{\lambda_k - \lambda_l} \psi_k \psi_l^T + \sum_{k \notin I, l \in I} \frac{\psi_k^T (\dot{W} + \lambda_l \hat{D}) \psi_l}{\lambda_k - \lambda_l} (\psi_k \psi_l^T + \psi_l \psi_k^T).
\]

Since \( S(x) = P_t(x, x) \), the claim follows by evaluating at the entry \((x, x)\) on both sides. \( \square \)

5.4 Theorem 3.1 (Separation at \( t = 1 \))

Proof of Theorem 3.1 We firstly show that condition (ii) implies (18): Note that

\[ (\#) \leq \frac{d_0(1 - \varepsilon)}{d_0} \leq \frac{d_0 - \varepsilon_1}{d_0} \leq 1, \]

thus

\[ \log \left( 1 + \frac{1}{2} \cdot \frac{\text{(\#)}}{1 + 2\varepsilon_1} \right) \leq \log \left( 1 + \frac{1}{2} \cdot \frac{\text{(\#)}}{1 + 2\varepsilon_1} \right) < 0.5. \]

Together with \( \frac{1}{1+2\varepsilon_1} < 1 \), this means that the r.h.s. of (20) is less than 0.5:\( d_0 \).

As a result, under these assumptions, Proposition 2 and Proposition 3 apply. By (19),

\[ S(x, t) - S(x, 0) = \int_0^t \left\{ - \sum_{k \in I, j \in I} (\psi_j^T \dot{D} \psi_k) \psi_j(x) \psi_k(x) + 2 \sum_{k \in I, j \notin I} \psi_j^T (\dot{W} - \lambda_k \hat{D}) \psi_k(x) \psi_j(x) \right\} \, d\tau, \]

where in the integrand all the variables involving time take value at time \( \tau \).

Introducing the notation

\[ \tilde{S}(t) := \sup_{x \in V = B \cup C} S(x, t), \]

we are going to prove the following two claims: For any \( t \in [0, 1] \),

1. \( \forall x \in V = B \cup C \),

\[ |S(t, x) - S(0, x)| \leq \tilde{C} \int_0^t \tilde{S}(\tau) \, d\tau, \quad \tilde{C} := \left( 1 + \frac{4}{\Delta} \right) \frac{2C}{d_0}, \]

2. \( \tilde{S}(t) \leq e^{\tilde{C}t} \tilde{S}(0) \).

If true, then

\[ |S(t, x) - S(0, x)| \leq \tilde{C} \int_0^t \tilde{S}(\tau) \, d\tau \leq \tilde{C} \int_0^t \tilde{S}(0) e^{\tilde{C}t} \, d\tau = \tilde{S}(0)(e^{\tilde{C}t} - 1). \]
Meanwhile, by Proposition 1 (2), \( S(0) \leq \frac{1}{n\delta_0} K(1 + 2\varepsilon_1) \), thus
\[
|S(t, x) - S(0, x)| \leq (e^{\tilde{C}t} - 1) \frac{K(1 + 2\varepsilon_1)}{n\delta_0}, \quad \forall x \in \mathcal{V}.
\]
By Proposition 1 (1), the initial separation on \( B \) and \( C \) by \( S(x) \) is at least \( g_0 \), so the threshold claimed in the theorem exists as long as
\[
g_0 \geq 2(e^{\tilde{C}t} - 1) \frac{K(1 + 2\varepsilon_1)}{n\delta_0}.
\]
This is reduced to
\[
2(e^{\tilde{C}t} - 1) \leq \frac{(#)}{1 + 2\varepsilon_1},
\]
which is guaranteed by condition (ii).

To prove Claim (1): By (39), \( \forall x \in \mathcal{V} \),
\[
|S(x, t) - S(x, 0)| \leq \int_0^t I(x, \tau) + II(x, \tau) + III(x, \tau) d\tau,
\]
where
\[
I(x, \tau) = \left| \sum_{k \in I, j \in I} (\psi_j^T \hat{D}\psi_k) \psi_k(x) \psi_j(x) \right|
\]
\[
II(x, \tau) = 2 \sum_{k \in I, j \notin I} \psi_j^T \hat{W} \psi_k \frac{\lambda_k - \lambda_j}{\lambda_k - \lambda_j} \psi_k(x) \psi_j(x)
\]
\[
III(x, \tau) = 2 \sum_{k \in I, j \notin I} \frac{\lambda_j \psi_j^T \hat{D}\psi_k}{\lambda_k - \lambda_j} \psi_k(x) \psi_j(x).
\]
For \( I(x, \tau) \), note that
\[
I(x, \tau) \leq \sum_{k \in I, j \in I} \sum_{y \in \mathcal{V}} |\hat{D}(y)||\psi_j(y)||\psi_k(y)||\psi_k(x)||\psi_j(x)|
\]
\[
= \sum_{y \in \mathcal{V}} |\hat{D}(y)|(\sum_{k \in I} |\psi_k(y)||\psi_k(x)|)^2
\]
\[
\leq \sum_{y \in \mathcal{V}} |\hat{D}(y)|(\sum_{k \in I} |\psi_k(y)|^2)(\sum_{i \in I} |\psi_i(x)|^2)
\]
\[
= \sum_{y \in \mathcal{V}} |\hat{D}(y)||S(y)||S(x)|.
\]
Utilizing the estimate that (similar to (36))
\[
S(x) = \sum_{j \in I} \psi_j(x)^2 \leq \sum_{j = 1}^n \psi_j(x)^2 = \frac{1}{d(x)} \leq \frac{1}{d_0}, \quad \forall x \in \mathcal{V},
\]
then
\[
I(x, \tau) \leq S(x) \sum_{y \in \mathcal{V}} |\hat{D}(y)| \frac{1}{d_0} = S(x) \frac{1}{d_0} \sum_{y \in \mathcal{V}} \sum_{z \in \mathcal{V}} |E(y, z)| = S(x) \frac{2C}{d_0}.
\]
By the definition of \( \tilde{S} \) in (40), this gives
\[
I(x, \tau) \leq \tilde{S}(\tau) \frac{2C}{d_0}.
\]
For $\Pi(x, \tau)$, by Proposition \([\ref{prop:prop_2}]\) in the denominator $|\lambda_k - \lambda_j| \geq \Delta$, then

\[
\Pi(x, \tau) \leq 2 \sum_{k \in \mathcal{I}, j \notin \mathcal{I}} \frac{|\psi_j^T \hat{W} \psi_k|}{|\lambda_k - \lambda_j|} \psi_k(x) || \psi_j(x) |
\]

\[
\leq \frac{2}{\Delta} \sum_{k \in \mathcal{I}, j \notin \mathcal{I}} |\psi_j^T E \psi_k| || \psi_j(x) || \psi_j(x) |
\quad (\hat{W} = E)
\]

\[
\leq \frac{2}{\Delta} \sum_{k \in \mathcal{I}, j \notin \mathcal{I}} \sum_{y \in \mathcal{V}, z \in \mathcal{V}} E(y, z) |\psi_j(y)| |\psi_j(z)| |\psi_k(x)| || \psi_j(x) || \psi_j(x) |
\]

\[
= \frac{2}{\Delta} \sum_{y \in \mathcal{V}, z \in \mathcal{V}} E(y, z) \sum_{k \in \mathcal{I}} |\psi_k(z)| |\psi_k(x)| \sum_{j \notin \mathcal{I}} |\psi_j(y)| || \psi_j(x) || \psi_j(x) |
\]

\[
\leq \frac{2}{\Delta} \sum_{y \in \mathcal{V}, z \in \mathcal{V}} E(y, z) \sqrt{S(z)} \sqrt{S(x)} \frac{1}{\sqrt{d(y)}} \frac{1}{\sqrt{d(x)}}
\quad \text{(definition of } S, \sum_{j=1}^{n} \psi_j(x)^2 = 1 \text{)}
\]

\[
\leq \frac{2}{\Delta} S(\tau) \frac{1}{d_0} \sum_{y \in \mathcal{V}, z \in \mathcal{V}} E(y, z) \quad (S(x), S(z) \leq S(\tau), \text{and } d(y), d(x) \geq d_0)
\]

which means that

\[
\Pi(x, \tau) \leq \frac{2}{\Delta} \frac{S(\tau)}{d_0} 2C^2.
\quad (45)
\]

For III($x$, $\tau$), since $|\lambda_k| \leq 1$, and $\hat{D}(y) = \sum_{z \in \mathcal{V}} E(y, z)$, one can show that

\[
III(x, \tau) \leq \frac{2}{\Delta} \frac{S(\tau)}{d_0} 2C
\quad (46)
\]

using a similar argument as in bounding $\Pi(x, \tau)$. Putting \([\ref{eq:44}]\) \([\ref{eq:45}]\) \([\ref{eq:46}]\) together, one has that

\[
|S(x, t) - S(x, 0)| \leq \int_0^t \tilde{S}(\tau) \frac{2C}{d_0} (1 + \frac{4}{\Delta}) d\tau = \int_0^t \tilde{S}(\tau) \tilde{C} d\tau,
\]

namely Claim (1).

To prove Claim (2): Note that

\[
\tilde{S}(t) - \tilde{S}(0) \leq \sup_{x \in \mathcal{V}} (S(x, t) - S(x, 0))
\]

(suppose that $\tilde{S}(t) = S(x_0, t)$ for some $x_0$, then $S(x_0, t) - \tilde{S}(0) \leq S(x_0, t) - S(x_0, 0)$). Since Claim (1) holds uniformly for $x$, this implies that

\[
\tilde{S}(t) - \tilde{S}(0) \leq \tilde{C} \int_0^t \tilde{S}(\tau) d\tau,
\]

and the claim then follows by Gronwall’s inequality.

\[
\square
\]

6 Further Comments

Viewed as diffusion distance. With $f(\lambda) = \lambda^p$, $S(x)$ can be interpreted as the (squared) diffusion distance between node $x$ and the origin at diffusion time $\frac{p}{2} \frac{X}{\Delta}$. The diffusion distance can be interpreted as
a “geometrical” distance between two nodes when the affinity graph is built from data points lying on a manifold embedded in the ambient space, and the distance is intrinsic to the manifold geometry and invariant to the specific embedding. Since \( \lambda_k^p \to 0 \) when \( p \) is large (except for \( \lambda_k = 1 \)), the origin point is the limiting point of the diffusion map embedding. Thus for \( x \) in a sub-clusters in \( \mathcal{C} \), the weighted norm \( S(x) \) with positive \( p \) can be viewed as a measurement of the extent of metastability (the “depth” of the well) of the potential well associated with the sub-cluster. In view of the diffusion distance, under the setting of this paper, nodes in \( \mathcal{B} \) are very similar to one another, and in comparison nodes in \( \mathcal{C} \) are distinct from those in \( \mathcal{B} \) (as well as from other sub-clusters in \( \mathcal{C} \), which is not reflected in \( S \)). A similar weighted form has also been studied in [28] for graph-based outlier detection. In the primary application considered in this paper, the leading eigenvalues are all close to 1, which means that the weighted form (22) is not very different unless \( p \) is large. On the other hand setting \( p \) to be large may suppress the high-index eigenvectors by small weights while they are actually the informative ones to indicate \( \mathcal{C} \). Due to these reasons, we mainly consider \( f = 1 \) in the current paper, though the analysis directly extends.

**Indexing eigenvectors by support regions.** The phenomenon being studied here also suggests that sorting by the magnitude of eigenvalues may not be the most informative way to index the eigenvectors, a problem recently addressed in [44]. Our paper here studies the special case where eigenvectors can be grouped by where they are mainly supported on. In the pseudo-dynamic \( \mathbf{1} \), the eigenvectors begin with being exactly supported on either \( \mathcal{C} \) or \( \mathcal{B} \) at \( t = 0 \), and as time develops this pattern is nearly preserved as long as the \( \mathcal{C} - \mathcal{B} \) inter-block connections are not too strong. The distinct support regions of eigenvectors appears to be irrelevant to the magnitude of the eigenvalues nor the existence of spectral gaps. This suggests that grouping eigenvectors by their localization regions maybe a better way to arrange them in such cases. However, one still needs to be careful with the instability of eigenvectors: As shown in the numerical example, when two eigenvalues get close in the pseudo-dynamic, the associated pair of eigenvectors “swap” their values. (The swapping may be analyzed by the differential equation \( \mathbf{15} \). assuming that among all the pairs of neighboring eigenvalues only one pair \( (\lambda_k - \lambda_j) \) is approaching zero, then the dynamic of \( \psi_k \) evolution is dominated by that pair, which approximates a “rotation” among the indices \( j \) and \( k \).) Our analysis in the current paper handles this by the summation in \( S \) over the index group \( I \), which makes \( S \) invariant to such “swaps” as long as \( j \) and \( k \) both belong to \( I \). Generally, since eigen-crossings only happen at isolated times in the deformation dynamics \( \mathbf{37} \), these special times can be excluded. Then one can say that the eigenvectors continue to almost localize on one of the two blocks most of the time.

**Eigenvector selection.** Related work have devised different methods to perform eigenvector selection to identify anomalies \( \mathbf{14, 15} \). The spectral embedding norm can also be used for eigenvector selection. \( \mathbf{15} \) demonstrated that it can be used to identify pixels which define clusters and find the embedding coordinates that best separate them from the background. An example in the anomaly detection case is given in the Appendix.

**References**

[1] B. Nadler and M. Galun, “Fundamental limitations of spectral clustering,” in Advances in Neural Information Processing Systems (NIPS) 19, 2006, pp. 1017–1024.

[2] F. Zhao, L. Jiao, H. Liu, X. Gao, and M. Gong, “Spectral clustering with eigenvector selection based on entropy ranking,” Neurocomputing, vol. 73, no. 10-12, pp. 1704–1717, 2010.

[3] T. Xiang and S. Gong, “Spectral clustering with eigenvector selection,” Pattern Recognition, vol. 41, no. 3, pp. 1012–1029, 2008.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient object detection: A benchmark,” IEEE transactions on image processing, vol. 24, no. 12, pp. 5706–5722, 2015.
[5] B. Peng, L. Zhang, and D. Zhang, “A survey of graph theoretical approaches to image segmentation,” *Pattern Recognition*, vol. 46, no. 3, pp. 1020 – 1038, 2013.

[6] Y. Weiss, “Segmentation using eigenvectors: a unifying view,” in *Proc. ICCV 1999*, vol. 2, 1999, pp. 975–982.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, 2002.

[8] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[9] L. Zelnik-Manor and P. Perona, “Self-tuning spectral clustering,” in *NIPS 17*, 2005, pp. 1601–1608.

[10] B. Nadler, S. Lafon, and R. Coifman, “Diffusion Maps - A Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms,” in *Principal Manifolds for Data Visualization and Dimension Reduction*, 2007, pp. 242–264.

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

[12] T. Idé and H. Kashima, “Eigenspace-based anomaly detection in computer systems,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 440–449.

[13] K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li, “Auto insurance fraud detection using unsupervised spectral ranking for anomaly,” *The Journal of Finance and Data Science*, vol. 2, no. 1, pp. 58–75, 2016.

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

[15] L. Wu, X. Wu, A. Lu, and Z.-H. Zhou, “A spectral approach to detecting subtle anomalies in graphs,” *Journal of Intelligent Information Systems*, vol. 41, no. 2, pp. 313–337, 2013.

[16] G. Mishne and I. Cohen, “Multiscale anomaly detection using diffusion maps,” *IEEE J. Sel. Topics Signal Process.*, vol. 7, pp. 111 – 123, Feb. 2013.

[17] Y. Aizenbud, A. Bermanis, and A. Averbuch, “PCA-based out-of-sample extension for dimensionality reduction,” *arXiv preprint arXiv:1511.00831*, 2015.

[18] G. Mishne, U. Shaham, A. Cloninger, and I. Cohen, “Diffusion nets,” *Applied and Computational Harmonic Analysis*, 2017. [Online]. Available: [http://www.sciencedirect.com/science/article/pii/S1063520317300955](http://www.sciencedirect.com/science/article/pii/S1063520317300955)

[19] N. Rabin and A. Averbuch, “Detection of anomaly trends in dynamically evolving systems,” in *AAAI Fall Symposium Series*, 2010.

[20] A. Singer, Y. Shkolnisky, and B. Nadler, “Diffusion interpretation of nonlocal neighborhood filters for signal denoising,” *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 118–139, Jan. 2009.

[21] G. Schiebinger, M. J. Wainwright, B. Yu et al., “The geometry of kernelized spectral clustering,” *The Annals of Statistics*, vol. 43, no. 2, pp. 819–846, 2015.

[22] A. Damle, V. Minden, and L. Ying, “Robust and efficient multi-way spectral clustering,” *arXiv preprint arXiv:1609.08251*, 2016.
[23] C. R. de Oliveira, “Eigenfunction entropy as a criterion of localization,” *Physics Letters A*, vol. 296, no. 4-5, pp. 165–170, 2002.

[24] C. J. Dsilva, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, “Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study,” *Applied and Computational Harmonic Analysis*, vol. 44, no. 3, pp. 759 – 773, 2018.

[25] U. Von Luxburg, M. Belkin, and O. Bousquet, “Consistency of spectral clustering,” *The Annals of Statistics*, pp. 555–586, 2008.

[26] M. Cucuringu and M. W. Mahoney, “Localization on low-order eigenvectors of data matrices,” *arXiv preprint arXiv:1109.1355*, 2011.

[27] G. Mishne and I. Cohen, “Iterative diffusion-based anomaly detection,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1682–1686.

[28] X. Cheng, G. Mishne, and S. Steinerberger, “The geometry of nodal sets and outlier detection,” *Journal of Number Theory*, vol. 185, pp. 48 – 64, 2018. [Online]. Available: [http://www.sciencedirect.com/science/article/pii/S0022314X17303621](http://www.sciencedirect.com/science/article/pii/S0022314X17303621)

[29] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[30] M. Goldstein and S. Uchida, “A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data,” *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[31] L. Akoglu, H. Tong, and D. Koutra, “Graph based anomaly detection and description: a survey,” *Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[32] A. Cloninger and W. Czaja, “Eigenvector localization on data-dependent graphs,” in *Sampling Theory and Applications (SampTA), 2015 International Conference on*. IEEE, 2015, pp. 608–612.

[33] D. C. Hoaglin and R. E. Welsch, “The hat matrix in regression and anova,” *The American Statistician*, vol. 32, no. 1, pp. 17–22, 1978.

[34] R. R. Coifman and S. Lafon, “Diffusion maps,” *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[35] M. Zworski, “Semiclassical analysis, volume 138 of graduate studies in mathematics,” *American Mathematical Society, Providence, RI*, vol. 15, 2012.

[36] S. Zelditch, “Local and global analysis of eigenfunctions on riemannian manifolds, handbook of geometric analysis. no. 1, 545658, adv. lect. math.(alm), 7,” *Int. Press, Somerville, MA*, 2008.

[37] T. Kato, *Perturbation theory for linear operators*. Springer Science & Business Media, 2013, vol. 132.

[38] L. Tian, J. Akerboom, E. R. Schreiter, and L. L. Looger, “Neural activity imaging with genetically encoded calcium indicators,” *Prog Brain Res*, vol. 196, pp. 79–94, 2012. [Online]. Available: [http://www.ncbi.nlm.nih.gov/pubmed/22341322](http://www.ncbi.nlm.nih.gov/pubmed/22341322)

[39] S. Reynolds, T. Abrahamsson, R. Schuck, P. Jesper Sjöström, S. R. Schultz, and P. L. Dragotti, “ABLE: An activity-based level set segmentation algorithm for two-photon calcium imaging data,” *eNeuro*, 2017.
Appendix

A Selection of eigenvectors in anomaly detection

Eigenvector selection can be used to better visualize and characterize an anomaly in the data, by finding a subspace in which it is separated from the normal data. Let \( x_{\text{max}} = \arg \max_x S(x) \) be the pixel with the maximal embedding norm. In Figures 12-13 we select the three eigenvectors with maximum absolute value on \( x_{\text{max}} \) (\( |\Psi_k(x_{\text{max}})| \) is plotted on the left). In the middle plot we color the pixels in the image according to pseudo-RGB values assigned to the three selected coordinates (right plot). Note that for Fig. 13 the selected coordinates are quite deep in the spectrum: \( k = 59, 47, 48 \).

![Figure 12: Selected eigenvectors of the graph Laplacian computed from the side-scan sonar image in Fig. 5(a). (Left) Plot of \( |\Psi_k(x_{\text{max}})| \) where \( x_{\text{max}} = \arg \max_x S(x) \), and x-axis is the index \( k \). (Middle) Each pixel colored according to RGB colors assigned to the three embedding coordinates, (Right) with maximum absolute value on \( x_{\text{max}} \).](image-url)
Figure 13: Same plot as Figure 12 for the side-scan sonar image in Fig. 5(c).

B Derivation of (14) and (15)

Recall that \( \{\lambda_k\}_{k=1}^n \) are the \( n \) real-valued eigenvalues of \( P \), and \( \psi_k \) is the eigenvector associated with \( \lambda_k \). By definition,
\[
W \dot{\psi}_k = \lambda_k D \psi_k, \quad \psi_k^T D \dot{\psi}_l = \delta_{kl},
\]
where \( W, D, \lambda_k \) and \( \psi_k \) all depend on \( t \) (the dependence is omitted in the notation) and are differentiable over time. Taking derivative w.r.t \( t \) of both sides,
\[
\begin{align*}
\dot{W} \psi_k + W \dot{\psi}_k &= \dot{\lambda}_k D \psi_k + \lambda_k \dot{D} \psi_k + \lambda_k D \dot{\psi}_k \\
0 &= \dot{\psi}_k^T D \dot{\psi}_l + \psi_k^T \dot{D} \psi_l + \psi_k^T D \dot{\psi}_l
\end{align*}
\]
(47)

Applying \( \psi_k^T \) to both sides of (47) gives
\[
\psi_k^T \dot{W} \psi_k = \dot{\lambda}_k \psi_k^T D \psi_k + \lambda_k \psi_k^T \dot{D} \psi_k = \dot{\lambda}_k + \lambda_k \psi_k^T \dot{D} \psi_k
\]
where in the first equality we use \( \psi_k^T D \psi_l = \delta_{kl} \).

To prove (15), since \( \{\psi_k\}_k \) form a \( D \)-orthonormal basis of \( \mathbb{R}^n \), let
\[
\dot{\psi}_k = \sum_{j=1}^n b_j \psi_j, \quad b_j = \psi_k^T D \psi_j.
\]
Apply \( \psi_j^T, j \neq k \), to both sides of (47) gives
\[
\psi_j^T \dot{W} \psi_k = (\lambda_k - \lambda_j) \psi_j^T D \psi_k + \lambda_k \psi_j^T \dot{D} \psi_k,
\]
thus
\[
\beta_j = \frac{\psi_j^T (\dot{W} - \lambda_k D) \psi_k}{\lambda_k - \lambda_j}, \quad j \neq k
\]
whenever \( \lambda_k - \lambda_j \neq 0 \). Letting \( l = k \) in (48) gives
\[
2 \psi_k^T D \dot{\psi}_k + \psi_k^T \dot{D} \psi_k = 0,
\]
thus
\[
\beta_k = -\frac{1}{2} \psi_k^T \dot{D} \psi_k.
\]
This proves (15).