Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning and Data Set Parameterization

Stéphane Lafon[1] and Ann B. Lee[2]

[1]Google Inc., stephane.lafon@gmail.com.

[2]Department of Statistics, Carnegie Mellon University, annlee@stat.cmu.edu.

Abstract

We provide evidence that non-linear dimensionality reduction, clustering and data set parameterization can be solved within one and the same framework. The main idea is to define a system of coordinates with an explicit metric that reflects the connectivity of a given data set and that is robust to noise. Our construction, which is based on a Markov random walk on the data, offers a general scheme of simultaneously reorganizing and subsampling graphs and arbitrarily shaped data sets in high dimensions using intrinsic geometry.

We show that clustering in embedding spaces is equivalent to compressing operators. The objective of data partitioning and clustering is to coarse-grain the random walk on the data while at the same time preserving a diffusion operator for the intrinsic geometry or connectivity of the data set up to some accuracy. We show that the quantization distortion in diffusion space bounds the error of compression of the operator, thus giving a rigorous justification for k-means clustering in diffusion space and a precise measure of the performance of general clustering algorithms.

Index Terms

Machine learning, Text analysis, Knowledge retrieval, Quantization, Graph-theoretic methods, Compression (coding), Clustering, Clustering similarity measures, Information visualization, Markov processes, Graph algorithms

I. Introduction

When dealing with data in high dimensions, one is often faced with the problem of how to reduce the complexity of a data set while preserving information that is important for, for example, understanding the data structure itself or for performing later tasks such as clustering, classification and regression. Dimensionality or complexity reduction is an ill-posed problem until one clearly defines what one is ready to lose. In this work, we attempt to find both a parameterization and an explicit metric that reflects

the *intrinsic* geometry of a given data set. With intrinsic geometry, we here mean a set of rules that describe the relationship between the objects in the data set without reference to structures outside of it; in our case, we define intrinsic geometry by the connectivity of the data points in a diffusion process. One application of this work, is manifold learning where we have a manifold, say a 2D "Swiss roll", embedded in a higher-dimensional space — but more generally, the problems of data parameterization, dimensionality reduction and clustering extend beyond manifold learning to general graphs of objects that are linked by edges with weights.

There is a large body of literature regarding the use of the spectral properties (eigenvectors and eigenvalues) of a pairwise similarity matrix for geometric data analysis. These methods can roughly be divided into two main categories: spectral graph cuts [1], [2], [3] and eigenmaps [4], [5], [6], [7]. The two methodologies were originally developed for different types of applications: segmentation and partitioning of graphs versus locality-preserving embeddings of data sets, respectively. Below we briefly review previous work and how it relates to the diffusion framework.

Suppose that $\Omega = \{x_1, ..., x_n\}$ is a data set of points, and assume that these points form the nodes of a weighted graph with weight function $w(x, y)$. In the graph-theoretic approach [8] to data partitioning, one seeks to divide the set of vertices into disjoint sets, where by some measure, the similarity among the vertices in a set is high, and the similarity across different sets is low. Different algorithm use different matrices but, in general, these spectral grouping methods are based on an analysis of the dominant eigenvectors of a suitably normalized weight matrix (see e.g. [1] for a review). If the weight function $w(x, y)$ satisfies certain conditions (symmetry and pointwise positivity), then one can interpret the pairwise similarities as edge flows in a Markov random walk on the graph. In this probabilistic formulation, the transition probability of going from point x to y in one step is

$$p(x, y) = \frac{w(x, y)}{\sum_{z \in \Omega} w(x, z)} .$$

The Normalized Cut problem provides a justification and some intuition for the use of the first non-trivial eigenfunction of the random walk's transition matrix [2]; the authors Shi and Malik also mention using higher-order eigenfunctions but do not provide a theoretical justification for such an analysis. More recently, Meila and Shi [3] have shown that the transition matrix P has piecewise constant eigenvectors relative to a partition $S = (S_1, S_2, \ldots, S_k)$ when the underlying Markov chain is lumpable with respect to S, i.e. when one is able to group vertices together due to similarities of their transition probabilities to the subsets S_j. The authors also define a "Modified Ncut" algorithm which, for the special case of lumpable Markov chains, finds all k segments by k-means of the eigenvectors of P.

Despite recent progress in the field of spectral graph theory, there are still many open questions. In particular: What is the intuition behind spectral clustering when eigenvectors are not piece-wise constant

(and Markov chains are not lumpable)? Naturally occuring data sets only display, at best, approximate lumpability; the issue then is whether we can say something more precise about the performance of various clustering algorithms. Furthermore, for general data sets, which eigenvectors of the Markov matrix should be considered and what is the relative importance of these? Below, we answer these questions by unifying ideas in spectral clustering, operator compression and data set parameterization.

The problem of spectral clustering is very closely related to the problem of finding low-dimensional locality-preserving embeddings of data sets. For example, suppose that we wish to find an embedding of Ω in $\mathbb{R}^p$ according to

$$x \mapsto f(x) = (f_1(x), \ldots, f_p(x))$$

that preserves the local neighborhood information. Several algorithms, such as LLE [4], Laplacian eigenmaps [6], Hessian eigenmaps [7], LTSA [5] and diffusion maps [9], [10], all aim at minimizing distortions of the form $Q(f) = \sum_i Q_i(f)$ where $Q_i(f)$ is a symmetric, positive semi-definite quadratic form that measures local variations of f around x_i. The p-dimensional embedding problem can, in these cases, be rewritten as an eigenvalue problem where the first p eigenvectors $(f_1, ..., f_p)$ provide the optimal embedding coordinates. The close relationship between spectral clustering and locality-preserving dimension reduction has, in particular, been pointed out by Belkin and Niyogi. In [6], the authors show that the Laplacian of a graph (whose eigenvectors are used in spectral cuts) is the discrete analogue of the Laplace-Beltrami operator on manifolds, and the eigenfunctions of the latter operator have properties desired for embeddings. However, as in the case of spectral clustering, the question of the number of eigenvectors in existing eigenmap methods is still open. Furthermore, as the distance metric in the embedding spaces is not explicitly defined, it is not clear *how* one should cluster and partition data. The usual approach is: First pick a dimension k, then calculate the first k non-trivial eigenvectors and weight these equally in clustering and other subsequent data analysis.

The contribution of this paper is two-fold: First, we provide a unified framework for spectral data analysis based on the idea of diffusion and put previous work in a new perspective. Our starting point is an *explicit* metric that reflects the connectivity of the data set. This so called "diffusion metric" can be explained in terms of transition probabilities of a Markov chain that evolves forward in time and is, unlike the geodesic distance, or the shortest path of a graph, very robust to noise. Similar distance measures have previously been suggested in clustering and data classification, see for example [11]. However, the use of such probabilistic distance measures in data parameterization is completely new. This paper unifies various ideas in eigenmaps, spectral cuts and Markov random walk learning (see Table I for a list of different methods). We show that, in the diffusion framework, the defined distance measure is induced by a non-linear embedding in Euclidean space where the embedding coordinates are weighted eigenvectors

Methods for clustering and non-linear dim. reduction	data set parameterization?	explicit metric in embedding space?
Spectral graph methods [1], [2], [3]	not directly addressed	no
Eigenmaps [4], [5], [6], [7]	yes	no
Isomap [13]	yes	yes
Markov random walk learning [11]	no	yes
Diffusion maps	yes	yes

TABLE I

A SIMPLIFIED TABLE OF DIFFERENT METHODS FOR CLUSTERING AND NON-LINEAR DIMENSIONALITY REDUCTION

of the graph Laplacian. Furthermore, the time parameter in the Markov chain defines the scale of the analysis, which in turn, determines the dimensionality reduction or the number of eigenvectors in the embedding.

The other contribution of this work is a novel approach to data partitioning and graph subsampling based on coarse-graining the dynamics of the Markov random walk on the data set. The goal is to subsample and reorganize the data set while retaining the spectral properties of the graph, and thus also the intrinsic geometry of the data set. We show that in order to maximize the quality of the eigenvector approximation, we need to minimize a distortion in the embedding space. Consequently, we are relating clustering in embedding spaces to lossy compression of operators — which is a key idea in this work. As a by-product, we are also obtaining a rigorous justification for k-means clustering in diffusion space. The latter method is, by construction, useful when dealing with data in high dimensions, and can (as in any kernel k-means algorithm [12]) be applied to arbitrarily shaped clusters and abstract graphs.

The organization of the paper is as follows. In Section II, we define diffusion distances and discuss their connection to the spectral properties and time evolution of a Markov chain random walk. In Section III, we construct a coarse-grained random walk for graph partitioning and subsampling. We relate the compression error to the distortion in the diffusion space. Moreover, we introduce diffusion k-means as a technique for distortion minimization. Finally, in Section IV, we give numerical examples that illustrate the ideas of a framework for simultaneous non-linear dimensionality reduction, clustering and subsampling of data using intrinsic geometry and propagation of local information through diffusion.

II. GEOMETRIC DIFFUSION AS A TOOL FOR HIGH-DIMENSIONAL DATA ANALYSIS

A. Diffusion distances

Our goal is to define a distance metric on an arbitrary set that reflects the connectivity of the points within the set. Suppose that one is dealing with a data set in the form of a graph. When identifying clusters, or groups of points, in this graph, one needs to measure the amount of interaction, as described

by the graph structure, between pairs of points. Following this idea, two points should be considered to be close if they are connected by many short paths in the graph. As a consequence, points within regions of high density (defined as groups of nodes with a high degree in the graph), will have a high connectivity. The connectivity is furthermore decided by the strengths of the weights in the graph. Below, we review the diffusion framework that first appeared in [10], and put it into the context of eigenmaps, dimensionality reduction and Markov random walk learning on graphs.

Let $G = (\Omega, W)$ be a finite graph with n nodes, where the weight matrix $W = \{w(x, y)\}_{x,y\in\Omega}$ satisfies the following conditions:

- symmetry: $W = W^T$, and
- pointwise positivity: $w(x, y) \geq 0$ for all $x, y \in \Omega$,

The way we define the weights should be completely application-driven, the only requirement being that $w(x, y)$ should represent the degree of similarity or affinity (as defined by the application) of x and y. In particular, we expect $w(x, x)$ to be a positive number. For instance, if we are dealing with data points on a manifold, we can start with a Gaussian kernel $w_\varepsilon = \exp\left(-||x - y||^2/\varepsilon\right)$, and then normalize it in order to adjust the influence of geometry versus the distribution of points on the manifold. Different normalization schemes and their connection to the Laplace-Beltrami operator on manifolds in the large sample limit $n \to \infty$ and $\varepsilon \to 0$ are discussed in [9].

The graph G with weights W represents our knowledge of the local geometry of the set. Next we define a Markov random walk on this graph. To this end, we introduce the degree $d(x)$ of node x as

$$d(x) = \sum_{z\in\Omega} w(x, z)\,.$$

If one defines P to be the $n \times n$ matrix whose entries are given by

$$p_1(x, y) = \frac{w(x, y)}{d(x)}\,,$$

then $p_1(x, y)$ can be interpreted as the probability of transition from x to y in 1 time step. By construction, this quantity reflects the first-order neighborhood structure of the graph. A new idea introduced in the diffusion maps framework, is to capture information on larger neighborhoods by taking powers of the matrix P, or equivalently, to run the random walk forward in time. If P^t is the t^{th} iterate of P, then the entry $p_t(x, y)$ represents the probability of going from x to y in t time steps. Increasing t, corresponds to propagating the local influence of each node with its neighbors. In other words, the quantity P^t reflects the intrinsic geometry of the data set defined via the connectivity of the graph in a diffusion process, and the time t of the diffusion plays the role of a scale parameter in the analysis.

If the graph is connected, we have that [8]:

$$\lim_{t\to+\infty} p_t(x, y) = \phi_0(y)\,, \tag{1}$$

where ϕ_0 is the unique stationary distribution

$$\phi_0(x) = \frac{d(x)}{\sum_{z\in\Omega} d(z)}\,.$$

This quantity is proportional to the degree of x in the graph, which is one measure of the density of points. The Markov chain is furthermore reversible, *i.e.*, it verifies the following detailed balance condition

$$\phi_0(x)p_1(x,y) = \phi_0(y)p_1(y,x)\,. \tag{2}$$

We are mainly concerned with the following idea: For a fixed but finite value $t > 0$, we want to define a metric between points of Ω which is such that two points x and z will be close if the corresponding conditional distributions $p_t(x,.)$ and $p_t(z,.)$ are close. A similar idea appears in [11], where the authors consider the L^1 norm $||p_t(x,.) - p_t(z,.)||$. Alternatively, one can use the Kullback-Leibler divergence or any other distance between $p_t(x,.)$ and $p_t(z,.)$. However, as shown below, the L^2 metric between the conditional distributions has the advantage that it allows one to relate distances to the spectral properties of the random walk — and thereby, as we will see in the next section, *connect Markov random walk learning on graphs with data parameterization via eigenmaps*. As in [14], we will define the "diffusion distance" D_t between x and y as the weighted L^2 distance

$$D_t^2(x,z) = \|p_t(x,\cdot) - p_t(z,\cdot)\|^2_{1/\phi_0} = \sum_{y\in\,\Omega} \frac{(p_t(x,y) - p_t(z,y))^2}{\phi_0(y)}\,, \tag{3}$$

where the "weights" $\frac{1}{\phi_0(x)}$ penalize discrepancies on domains of low density more than those of high density.

This notion of proximity of points in the graph reflects the intrinsic geometry of the set in terms of connectivity of the data points in a diffusion process. The diffusion distance between two points will be small if they are connected by many paths in the graph. This metric is thus a key quantity in the design of inference algorithms that are based on the preponderance of evidences for a given hypothesis. For example, suppose one wants to infer class labels for data points based on a small number of labeled examples. Then one can easily propagate the label information from a labeled example x to the new point y following (i) the shortest path, or (ii) all paths connecting x to y. The second solution (which is employed in the diffusion framework and in [11]) is usually more appropriate, as it takes into account all "evidences" relating x to y. Furthermore, since diffusion-based distances add up the contribution from several paths, they are also (unlike the shortest path) robust to noise; the latter point is illustrated via an example in Section IV-B.

B. Dimensionality reduction and parameterization of data by diffusion maps

As mentioned, an advantage of the above definition of the diffusion distance is the connection to the spectral theory of the random walk. As is well known, the transition matrix P that we have constructed

has a set of left and right eigenvectors and a set of eigenvalues $|\lambda_0| \geq |\lambda_1| \geq ... \geq |\lambda_{n-1}|$:

$$\phi_j^T P = \lambda_j \phi_j^T \text{ and } P\psi_j = \lambda_j \psi_j \,,$$

where it can be verified that $\lambda_0 = 1$, $\psi_0 \equiv 1$ and that $\phi_k^T \psi_l = \delta_{kl}$. In fact, left and right eigenvectors are dual, and can be regarded as signed measures and test functions, respectively. These two sets of vectors are related according to

$$\psi_l(x) = \frac{\phi_l(x)}{\phi_0(x)} \text{ for all } x \in \Omega \,. \tag{4}$$

For ease of notation, we normalize the left eigenvectors of P with respect to $1/\phi_0$:

$$\|\phi_l\|^2_{1/\phi_0} = \sum_x \frac{\phi_l^2(x)}{\phi_0(x)} = 1 \,, \tag{5}$$

and the right eigenvectors with respect to the weight ϕ_0:

$$\|\psi_l\|^2_{\phi_0} = \sum_x \psi_l^2(x)\phi_0(x) = 1 \,. \tag{6}$$

If $p_t(x, y)$ is the kernel of the t^{th} iterate P^t, we will then have the following biorthogonal spectral decomposition:

$$p_t(x, y) = \sum_{j \geq 0} \lambda_j^t \psi_j(x)\phi_j(y) \,. \tag{7}$$

The above identity corresponds to a weighted principal component analysis of P^t. The first k terms provide the best rank-k approximation of P^t, where "best" is defined according to the following weighted metric for matrices:

$$\|A\|^2 = \sum_x \sum_y \phi_0(x) a(x, y)^2 \frac{1}{\phi_0(y)} \,.$$

Here is our main point: If we insert Equation 7 into Equation 3, we will have that

$$D_t^2(x, z) = \sum_{j=1}^{n-1} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2 \,.$$

Since $\psi_0 \equiv 1$ is a constant vector, it does not enter in the sum above. Furthermore, because of the decay of the eigenvalues [1], we only need a few terms in the sum for a certain accuracy. To be precise, let $q(t)$ be the largest index j such that $|\lambda_j|^t > \delta |\lambda_1|^t$. The diffusion distance can then be approximated to relative precision δ using the first $q(t)$ non-trivial eigenvectors and eigenvalues according to

$$D_t^2(x, z) \simeq \sum_{j=1}^{q(t)} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2 \,.$$

[1] The speed of the decay depends on the graph structure. For example, for the special case of a fully connected graph, the first eigenvalue will be 1 and the remaining eigenvalues will be equal to 0. The other extreme case is a graph that is totally disconnected with all eigenvalues equal to 1.

Now observe that the identity above can be interpreted as the Euclidean distance in $\mathbb{R}^{q(t)}$ if we use the right eigenvectors weighted with λ_j^t as coordinates on the data. In other words, this means that if we introduce the diffusion map

$$\Psi_t : x \longmapsto \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_{q(t)}^t \psi_{q(t)}(x) \end{pmatrix}, \tag{8}$$

then clearly,

$$D_t^2(x,z) \simeq \sum_{j=1}^{q(t)} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2 = \|\Psi_t(x) - \Psi_t(z)\|^2 . \tag{9}$$

Note that the factors λ_j^t in the definition of Ψ_t are crucial for this statement to hold.

The mapping $\Psi_t : \Omega \to \mathbb{R}^{q(t)}$ provides a parameterization of the data set Ω, or equivalently, a realization of the graph G as a cloud of points in a lower-dimensional space $\mathbb{R}^{q(t)}$, where the re-scaled eigenvectors are the coordinates. The dimensionality reduction and the weighting of the relevant eigenvectors are dictated by both the time t of the random walk and the spectral fall-off of the eigenvalues.

Equation 9 means that Ψ_t embeds the entire data set in $\mathbb{R}^{q(t)}$ in such a way that the Euclidean distance is an approximation of the diffusion distance. Our approach is thus different from other eigenmap methods: Our starting points is an *explicitly* defined distance metric on the data set or graph. This distance is also the quantity we wish to preserve during a non-linear dimensionality reduction.

III. Graph partitioning and subsampling

In what follows, we describe a novel scheme for subsampling data sets that — as above — preserves the intrinsic geometry defined by the connectivity of the data points in a graph. The idea is to construct a coarse-grained version of the original random walk on a new graph $\widetilde{G}$ with similar spectral properties. This new Markov chain is obtained by grouping points into clusters and appropriately averaging the transition probabilities between these clusters. We show that in order to retain most of the spectral properties of the original random walk, the choice of clusters in critical. More precisely, the quantization distortion in diffusion space bounds the error of the approximation of the diffusion operator.

One application is dimensionality reduction and clustering of arbitrarily shaped data sets using geometry; see Section IV for some simple examples. However, more generally, the construction also offers a systematic way of subsampling operators [15] and arbitrary graphs using geometry.

A. Construction of a coarse-grained random walk

Start by considering an arbitrary partition $\{S_i\}_{1\leq i\leq k}$ of the set of nodes Ω. Our aim is to aggregate the points in each set in order to coarse-grain both the state set Ω and the time evolution of the random walk. To do so, we regard each set S_i as corresponding to the nodes of a k-node graph $\widetilde{G}$, whose weight function is defined as

$$\widetilde{w}(S_i, S_j) = \sum_{x\in S_i}\sum_{y\in S_j} \phi_0(x)p_t(x,y)\,,$$

where the sum involves all the transition probabilities between points $x \in S_i$ and $y \in S_j$ (see Figure 1).

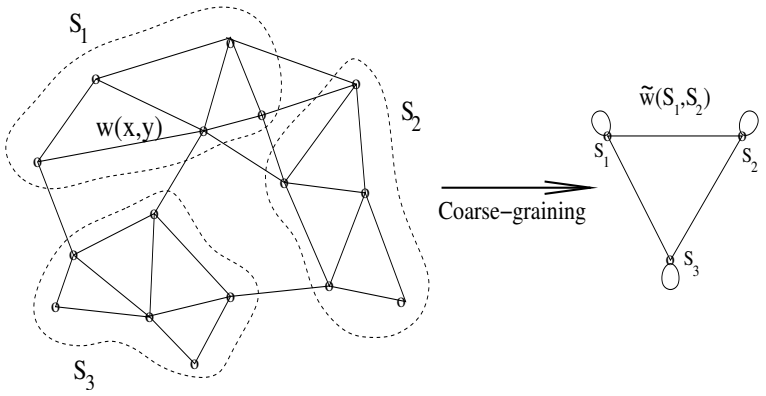

Fig. 1. Example of a coarse-graining of a graph: For a given partition $\Omega = S_1 \cup S_2 \cup S_3$ of the set of nodes in a graph G, we define a coarse-grained graph $\widetilde{G}$ by aggregating all nodes belonging to a subset S_i into a meta-node. By appropriately averaging the transition probabilities between points $x \in S_i$ and $y \in S_j$, for $i, j = 1, 2, 3$, we then compute new weights $\widetilde{w}(S_i, S_j)$ and a new Markov chain with transition probabilities $\widetilde{p}(S_i, S_j)$.

From the reversibility condition of Equation 2, it can be verified that this graph is symmetric, i.e. that $\widetilde{w}(S_i, S_j) = \widetilde{w}(S_j, S_i)$. By setting

$$\widetilde{\phi}_0(S_i) = \sum_{x\in S_i} \phi_0(x)\,,$$

one can define a reversible Markov chain on this graph with stationary distribution $\widetilde{\phi}_0 \in \mathbb{R}^k$ and transition probabilities

$$\widetilde{p}(S_i, S_j) = \frac{\widetilde{w}(S_i, S_j)}{\sum_k \widetilde{w}(S_i, S_k)} = \sum_{x\in S_i}\sum_{y\in S_j} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} p_t(x,y)\,.$$

Let $\widetilde{P}$ be the $k \times k$ transition matrix on the coarse-grained graph. More generally, for $0 \leq l \leq n-1$, we define in a similar way coarse-grained versions of ϕ_l by summing over the nodes in a partition:

$$\widetilde{\phi}_l(S_i) = \sum_{x\in S_i} \phi_l(x)\,. \tag{10}$$

As in Equation 4, we define coarse-grained versions of ψ_l according to the duality condition

$$\widetilde{\psi}_l(S_i) = \frac{\widetilde{\phi}_l(S_i)}{\widetilde{\phi}_0(S_i)}\,, \tag{11}$$

which is equivalent to taking a weighted average of ψ_l over S_i:

$$\widetilde{\psi}_l(S_i) = \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \psi_l(x) \ . \tag{12}$$

The coarse-grained kernel $\widetilde{p}(S_i, S_j)$ contains all the information in the data regarding the connectivity of the new nodes in the graph $\widetilde{G}$. The extent to which the above vectors constitute approximations of the left and right eigenvectors of $\widetilde{P}$ depends on the particular choice of the partition $\{S_i\}$. We investigate this issue more precisely in the next section.

B. Approximation error. Definition of geometric centroids

In a similar manner to Equation 5 and Equation 6, we define the norm on coarse-grained signed measures $\widetilde{\phi}_l$ to be

$$\|\widetilde{\phi}_l\|^2_{1/\widetilde{\phi}_0} = \sum_i \frac{\widetilde{\phi}_l^2(S_i)}{\widetilde{\phi}_0(S_i)} \,,$$

and on the coarse-grained test functions $\widetilde{\psi}_l$ to be

$$\|\widetilde{\psi}_l\|^2_{\widetilde{\phi}_0} = \sum_i \widetilde{\psi}_l^2(S_i) \widetilde{\phi}_0(S_i) \,.$$

We now introduce the definition of a geometric centroid, or a representative point, of each partition S_i:

Definition 1 (geometric centroid): Let $1 \leq i \leq k$. The geometric centroid $c(S_i)$ of subset S_i of Ω is defined as the weighted sum

$$c(S_i) = \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \Psi_t(x) \,.$$

The following result shows that for small values of l, $\widetilde{\phi}_l$ and $\widetilde{\psi}_l$ are approximate left and right eigenvectors of $\widetilde{P}$ with eigenvalue λ_l^t.

Theorem 2: We have for $0 \leq l \leq n-1$,

$$\widetilde{\phi}_l^T \widetilde{P} = \lambda_l^t \widetilde{\phi}_l^T + e_l \text{ and } \widetilde{P} \widetilde{\psi}_l = \lambda_l^t \widetilde{\psi}_l + f_l \,.$$

where

$$\|e_l\|^2_{1/\widetilde{\phi}_0} \leq 2\mathcal{D} \text{ and } \|f_l\|^2_{\widetilde{\phi}_0} \leq 2\mathcal{D} \,,$$

and

$$\mathcal{D} = \sum_i \sum_{x \in S_i} \phi_0(x) \|\Psi_t(x) - c(S_i))\|^2$$

This means that if $|\lambda_l|^t \gg \sqrt{\mathcal{D}}$ then $\widetilde{\phi}_l$ and $\widetilde{\psi}_l$ are approximate left and right eigenvectors of $\widetilde{P}$ with approximate eigenvalue λ_l^t. The proof of this theorem can be found in Appendix .

The previous result also shows that in order to maximize the quality of approximation, we need to minimize the following distortion in diffusion space:

$$\begin{aligned} \mathcal{D} &= \sum_i \sum_{x \in S_i} \phi_0(x) \|\Psi_t(x) - c(S_i))\|^2 \\ &\approx \mathbb{E}_i \left\{ \mathbb{E}_{X|i} \left\{ \|\Psi_t(X) - c(S_i))\|^2 | X \in S_i \right\} \right\} , \end{aligned} \tag{13}$$

which can also be written in terms of a weighted sum of pairwise distances according to

$$\begin{aligned} \mathcal{D} &= \frac{1}{2} \sum_i \widetilde{\phi}_0(S_i) \sum_{z \in S_i} \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \frac{\phi_0(z)}{\widetilde{\phi}_0(S_i)} \|\Psi_t(x) - \Psi_t(z)\|^2 \\ &\approx \frac{1}{2} \mathbb{E}_i \left\{ \mathbb{E}_{X,Z|i} \left\{ \|\Psi_t(X) - \Psi_t(Z))\|^2 | X, Z \in S_i \right\} \right\} . \end{aligned} \tag{14}$$

C. An algorithm for distortion minimization

Finally, we make a connection to kernel k-means and the algorithmic aspects of the minimization. The form of $\mathcal{D}$ given in Equation 13 is classical in information theory, and its minimization is equivalent to solving the problem of quantizing the diffusion space with k codewords based on the mass distribution of the sample set $\Psi_t(\Omega)$. This optimization issue is often addressed via the k-means algorithm [16] which guarantees convergence towards a local minimum:

1) Step 0: initialize the partition $\{S_i^{(0)}\}_{1 \leq i \leq k}$ at random in the diffusion space,
2) For $p > 0$, update the partition according to

$$S_i^{(p)} = \{x \text{ such that } i = \arg\min_j \|\Psi_t(x) - c(S_j^{(p-1)})\|^2\} ,$$

   where $1 \leq i \leq k$, and $c(S_j^{(p-1)})$ is the geometric centroid of $S_j^{(p-1)}$,
3) Repeat point 2 until convergence.

A drawback of this approach is that each center of mass $\{c(S_i)\}$ may not belong to the set $\Psi_t(E)$ itself. This can be a problem in some applications where such combinations have no meaning, such as in the case of gene data. In order to obtain representatives $\{c_i\}$ of the clusters that belong to the original set E, we introduce the following definition of diffusion centers:

Definition 3 (diffusion center): The diffusion center $u(S)$ of a subset S of Ω is any solution of

$$\arg\min_{x \in \Omega} \|\Psi_t(x) - c(S)\|^2 .$$

This notion does not define a unique diffusion center, but it is sufficient for our purpose of minimizing the distortion. Note that $u(S)$ is a generalization of the idea of center of mass to graphs.

Now, if $\{S_i\}$ is the output of the k-means algorithm, then we can assign to each point in S_i the representative center $u(S_i)$. In that sense, the graph $\widetilde{G}$ is a subsampled version of G that, for a given value of k, retains the spectral properties of the graph. The analysis above provides a rigorous justification for

k-means clustering in diffusion spaces, and furthermore links our work to both spectral graph partitioning (where often only the first non-trivial eigenvector of the graph Laplacian is taken into account) and eigenmaps (where one uses spectral coordinates for data parameterization).

IV. Numerical examples

A. *Importance of learning the nonlinear geometry of data in clustering*

In many applications, real data sets exhibit highly nonlinear structures. In such cases, linear methods such as Principal Components will not be very efficient for representing the data. With the diffusion coordinates, however, it is possible to learn the intrinsic geometry of data set, and then project the data points into a non-linear coordinate space with a diffusion metric. In this diffusion space, one can use classical geometric algorithms (such as separating hyperplane-based methods, k-means algorithms, etc.) for unsupervised as well as supervised learning.

To illustrate this idea, we study the famous Swiss roll. This data set is intrinsically a surface embedded in 3 dimensions. In this original coordinate system, global extrinsic distances, such as the Euclidean distance, are often meaningless as they do not incorporate any information on the structure or shape of the data set. For instance, if we run the k-means algorithm for clustering with $k = 4$, the obtained clusters do not reflect the natural geometry of the set. As shown in Figure 2, there is some "leakage" between different parts of the spiral due to the convexity of the k-means clusters in the ambient space.

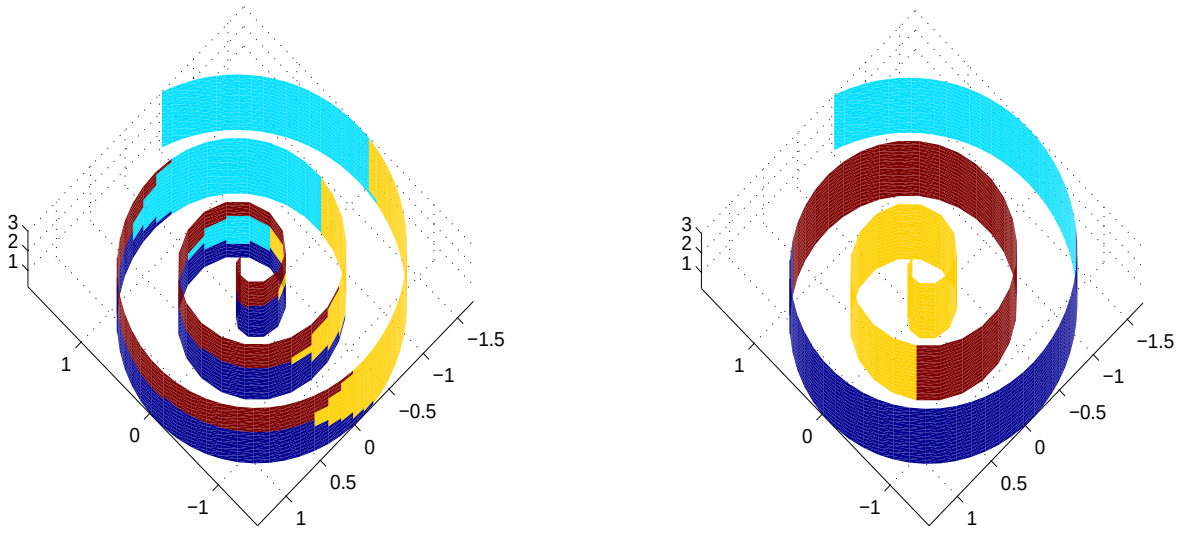

Fig. 2. The Swiss roll, and its quantization by k-means ($k = 4$) in the original coordinate system (left) and in the diffusion space (right).

As a comparison, we also show in Figure 2 the result of running the k-means algorithm in diffusion space. In the latter case, we obtain meaningful clusters that respect the intrinsic geometry of the data set.

B. *Robustness of the diffusion distance*

One of the most attractive features of the diffusion distance is its robustness to noise and small perturbations of the data. In short, its stability follows from the fact that it reflects the connectivity of the points in the graph. We illustrate this idea by studying the case of data points approximately lying on a spiral in the two-dimensional plane. The goal of this experiment is to show that the diffusion distance is a robust metric on the data, and in order to do so, we compare it to the shortest path (or geodesic) distance that is employed in schemes such as ISOMAP [13].

We generate 1000 instances of a noisy spiral in the plane, each corresponding to a different realization of the random noise perturbation (see Figure 3). From each instance, we construct a graph by connecting all pairs of points at a distance less than a given threshold τ, which is kept constant over the different realizations of the spiral. The corresponding adjacency matrix W contains only zeros or ones, depending on the absence or presence of an edge, respectively. In order to measure the robustness of the diffusion distance, we repeatedly compute the diffusion distance between two points of reference A and B in all 1000 noisy spirals. We also compute the geodesic distance between these two points using Dijkstra's algorithm.

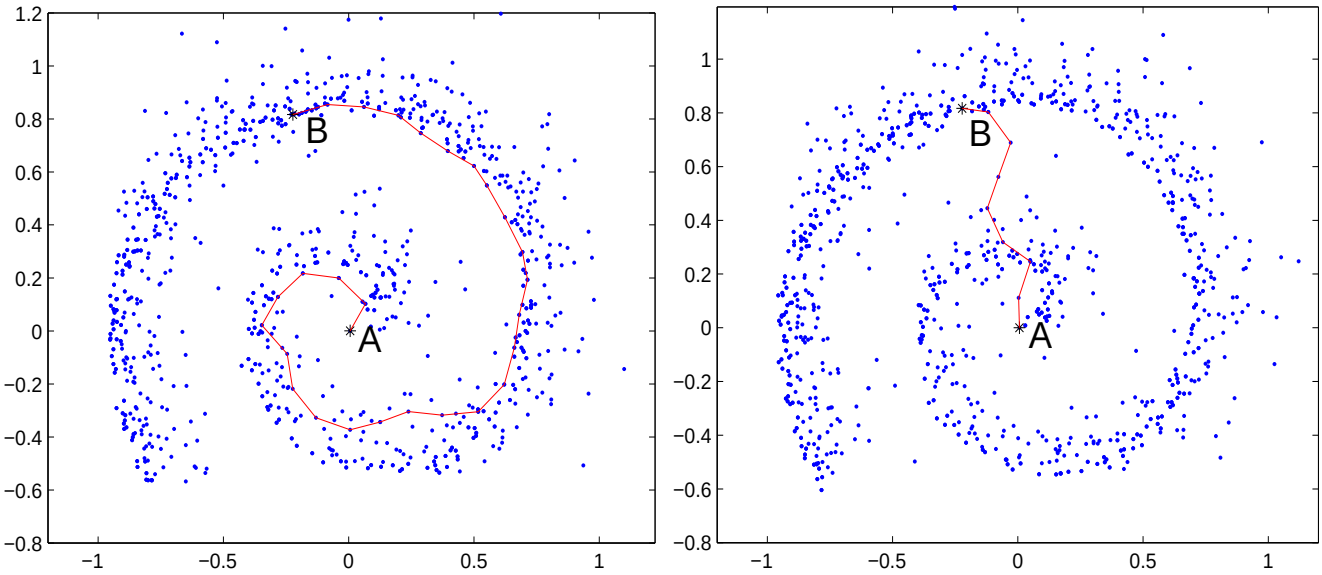

Fig. 3. Two realizations of a noisy spiral with points of references A and B. Ideally, the shortest path between A and B should follow the branch of the spiral (left). However, some realizations of the noise may give rise to shortcuts, thereby dramatically reducing the length of the shortest path (right).

As shown in Figure 3, depending on the presence of shortcuts arising from points appearing between the branches of the spiral, the geodesic distance (or shortest path length) between A and B may vary by large amounts from one realization of the noise to another. The histogram of all geodesic distances

measurements between A and B over the 1000 trials is shown on Figure 4. The distribution of the geodesic distance appears poorly localized, as its standard deviation equals 42% of its mean. This indicates that the geodesic distance is extremely sensitive to noise and thus unreliable as a measure of distance.

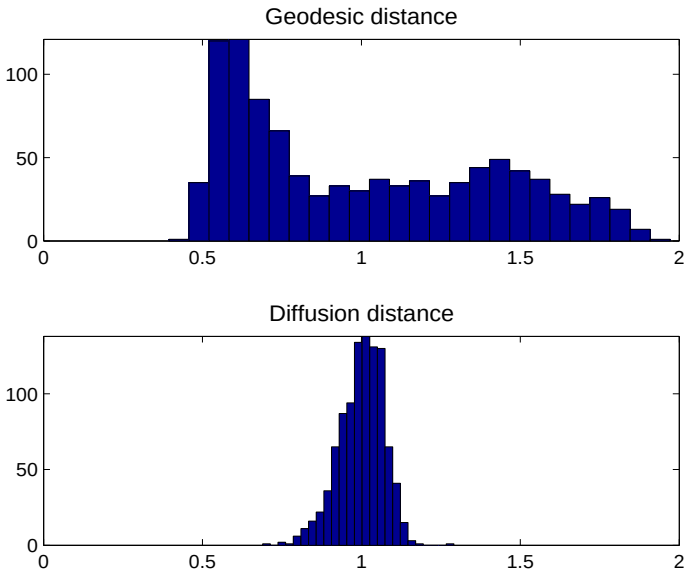

Fig. 4. Distribution of the geodesic (top) and diffusion (bottom) distances. Each distribution was rescaled in order to have a mean equal to 1.

The diffusion distance, however, is not sensitive to small random perturbations of the data set because, unlike the geodesic distance, it represents an average quantity. More specifically, it takes into account all paths of length less than or equal to t that connect A and B. As a consequence, shortcuts due to noise will have little weight in the computation, as the number of such paths is much smaller than the number of paths following the shape of the spiral. This is also what our experiment confirms: Figure 4 shows the distribution of the diffusion distances between A and B over the random trials. In this experiment, t was taken to be equal to 600. The corresponding histogram shows a very localized distribution, with a standard deviation equal to only 7% of its mean, which translates into robustness and consistency of the diffusion distance.

C. Organizing and clustering words via diffusion maps

Many of the ideas in this paper can be illustrated with an application to word-document clustering. We here show how we can measure the semantic association of words using diffusion distances, and how we can organize and form representative meta-words using diffusion maps and the k-means algorithm.

Our starting point is a collection of $p = 1161$ Science News articles. These articles belong to 8 different categories (see [17]). Our goal is to cluster words based on their distribution over the documents. From

the database, we extract the 20 most common words in each document, which corresponds to 3218 unique words total. Out of these words, we then select words with an intermediate document conditional entropy. The conditional entropy of a document X given a word y is defined as $H_{X|y} = -\sum_x p(x|y) \log p(x|y)$. Words with a very low entropy occur, by definition, in few documents and are often not good descriptors of the database, while high-entropy words such as "it", "if", "and", etc. can be equally uninformative. Thus, in our case, we choose a set of $N = 1004$ words with entropy $2 < H(X|y) < 4$. As in [17], we calculate the mutual information between document x and word y according to

$$m_{x,\,y} = \log \left(\frac{f_{x,\,y}}{\sum_\xi f_{\xi,\,y} \sum_\eta f_{\xi,\eta}} \right) \ ,$$

where $f_{x,y} = c_{x,y}/N$, and $c_{x,y}$ is the number of times word w appears in document x. In the analysis below, we describe word y in terms of the p-dimensional feature vector

$$e_y = [m_{1,\,y}, m_{2,\,y}, \ldots m_{p,\,y}] \ .$$

Our first task is to find a low-dimensional embedding of the words. We form the kernel

$$w(e_i, e_j) = \exp \left(-\frac{||e_i - e_j||^2}{\sigma^2} \right) \ ,$$

and normalize it, as described in Section II-A, to obtain the diffusion kernel $p_t(e_i, e_j)$. We then embed the data using the eigenvalues λ_k^t and the eigenvectors ψ_k of the kernel (see Equation 8). As mentioned, the effective dimensionality of the embedding is given by the spectral fall-off of the eigenvalues. For $\sigma = 18$ and $t = 4$, we have that $(\lambda_{10}/\lambda_1)^t < 0.1$, which means that we have effectively reduced the dimensionality of the original p-dimensional problem, where $p = 1161$, with a factor of about $1/100$. Figure 5 shows the first two coordinates in the diffusion map; Euclidean distances in the figure only approximately reflect diffusion distances since higher-order coordinates are not displayed. Note that the words have roughly been rearranged according to their semantics. Starting to the left, moving counter-clockwise, we have words that, respectively, express concepts in medicine, social sciences, computer science, physics, astronomy, earth sciences and anthropology.

Next, we show that the original 1004 words can be clustered and grouped into representative "meta-words" by minimizing the distortion in Equation 13. The k-means algorithm with $k = 100$ cluster leads to the results in Figure 5. Table II furthermore gives some examples of diffusion centers and words in a cluster. The diffusion centers or "meta-words" form a coarse-grained representation of the word graph and can, for example, be used as conceptual indices for document retrieval and document clustering. This will be discussed in later work.

Diffusion center	All other words in cluster
psychiatric	depression, psychiatrist, psychologist
talent	award, competition, finalist, intel, prize, scholarship, student, winner
laser	beam, nanometer, photon, pulse, quantum
velocity	detector, emit, infrared, ultraviolet
gravitational	bang, cosmo, gravity, hubble
orbiting	jupiter, orbit, solar
geologic	beneath, crust, depth, earthquake, ice, km, plate, seismic, trapped, volcanic
warming	climate, el, nino, pacific, weather
underwater	atlantic, coast, continent, floor, island, marine, seafloor, sediment
ecosystem	algae, drought, dry, ecologist, extinction, forest, gulf, lake, pollution, river
farmer	carolina, crop, fish, florida, insect, nutrient, pesticide, pollutant, soil, tree, tropical, wash, wood
virus	aids, allergy, hiv, resistant, vaccine, viral
cholesterol	aging, artery, fda, insulin, obesity, sugar, vitamin

TABLE II

EXAMPLES OF DIFFUSION CENTERS AND WORDS IN A CLUSTER

V. DISCUSSION

In this work, we provide evidence that clustering, graph partitioning and data set parameterization can be solved within one and the same framework. Our starting point is to find a meaningful representation of the data, and to *explicitly* define a distance metric on the data. Here we propose using a system of coordinates and a metric that reflects the connectivity of the data set. By doing so, we lay down a solid foundation for subsequent data analysis.

All the geometry of the data set is captured in a diffusion kernel. However, unlike SVM and so called "kernel methods" [18], [19], [20], we are working with the embedding coordinates explicitly. Our method is completely data driven: both the data representation and the kernel are computed directly on the data. The notion of a distance allows us to more precisely define our goals in clustering and dimensionality reduction. In addition, the diffusion framework makes it possible to directly connect grouping in embedding spaces to spectral graph clustering and data analysis by Markov chains [21], [11].

In a sense, we are extending Meila and Shi's work [3] from lumpable Markov chains and piece-wise constant eigenvectors to the general case of arbitrary Markov chains and arbitrary eigenvectors. The key idea is to work with embedding spaces directly and also to take powers of the transition matrix. The time parameter t sets the scale of the analysis. Note also that by using different values of t, we are able to perform a multiscale analysis of the data [22], [23].

Our other contribution is a novel scheme for simultaneous dimensionality reduction, parameterization

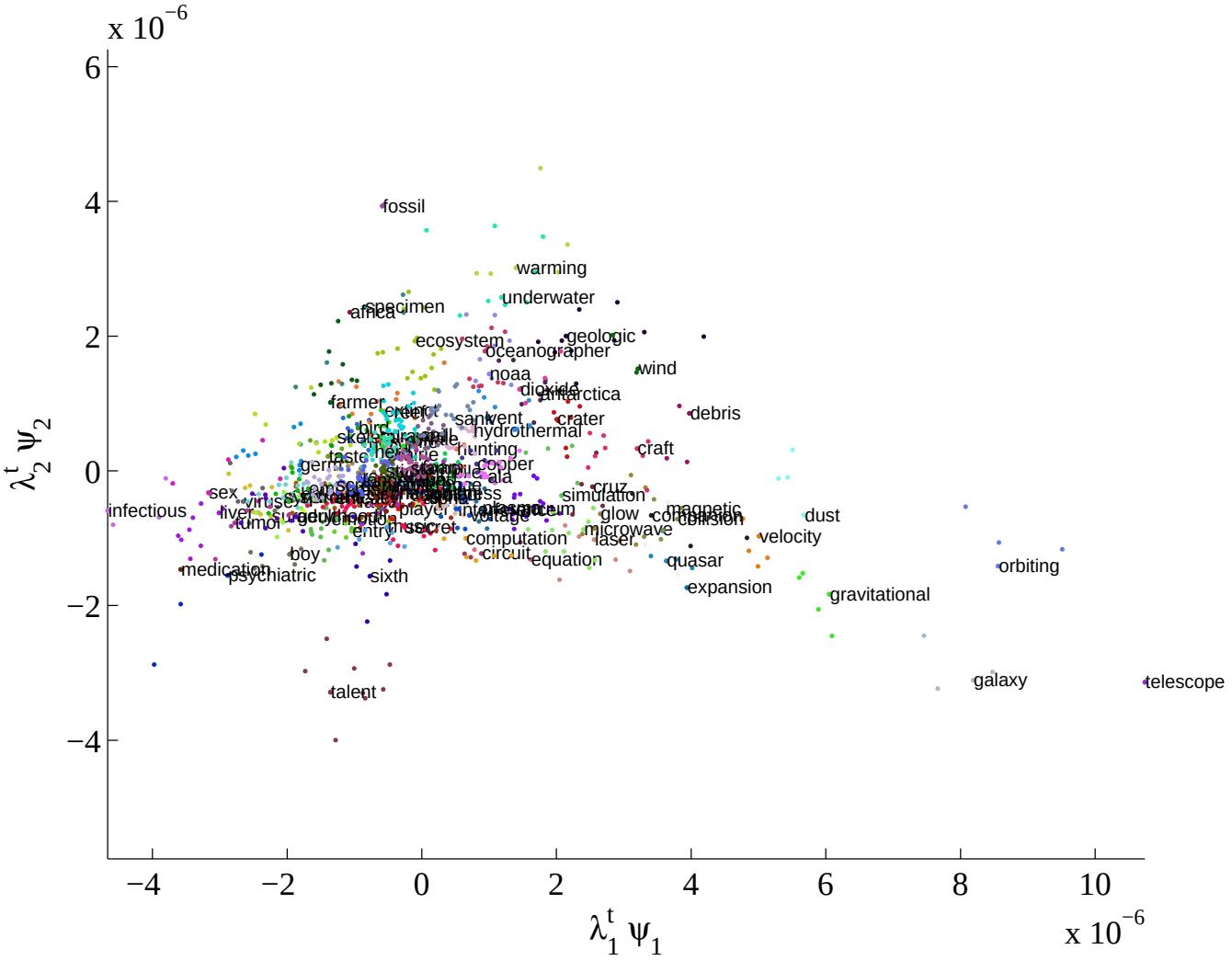

Fig. 5. Embedding and k-means clustering of 1004 words for $t = 4$ and $k = 100$. The colors correspond to the different word clusters, and the text labels the representative diffusion center or "meta-word" of each word cluster. Note that the words are automatically arranged according to their semantics.

and subsampling of data sets. We show that clustering in embedding spaces is equivalent to compressing operators. As mentioned, the diffusion operator defines the geometry of our data set. There are several ways of compressing a linear operator, depending on what properties one wishes to retain. For instance, in [22], the goal is to maintain sparseness of the representation while achieving the best compression rate. On the other hand, the objective in our work is to cluster or partition a given data set while at the same time preserving the operator (that captures the geometry of the data set) up to some accuracy. We show that, for a given partitioning scheme, the corresponding quantization distortion in diffusion space bounds the error of compression of the operator. This gives us a precise measure of the performance of clustering algorithms. To find the best clustering, one needs to minimize this distortion, and the k-means algorithm is one way to achieve this goal. Another aspect of our approach is that we are coarse-graining a Markov

chain defined on the data, thus offering a general scheme to subsample and parameterize graphs based on intrinsic geometry.

ACKNOWLEDGMENTS

We are grateful to R.R. Coifman for his insight and guidance. We would also like to thank M. Maggioni and B. Nadler for contributing in the development of the diffusion framework, and Y. Keller for providing comments on the manuscript.

APPENDIX

In this section, we provide a proof for Theorem 2, which we recall as

Theorem 4: We have for $0 \leq l \leq n-1$,

$$\widetilde{\phi}_l^T \widetilde{P} = \lambda_l^t \widetilde{\phi}_l^T + e_l \text{ and } \widetilde{P}\widetilde{\psi}_l = \lambda_l^t \widetilde{\psi}_l + f_l \,.$$

where

$$\|e_l\|^2_{1/\widetilde{\phi}_0} \leq 2\mathcal{D} \text{ and } \|f_l\|^2_{\widetilde{\phi}_0} \leq 2\mathcal{D} \,,$$

and

$$\mathcal{D} = \sum_i \sum_{x \in S_i} \phi_0(x) \|\Psi_t(x) - c(S_i))\|^2$$

This means that if $|\lambda_l|^t \gg \sqrt{\mathcal{D}}$ then $\widetilde{\phi}_l$ and $\widetilde{\psi}_l$ are approximate left and right eigenvectors of $\widetilde{P}$ with approximate eigenvalue λ_l^t.

Proof: We start by treating left eigenvectors: For all $z \in S_i$, we define

$$r_{ij}(z) = \widetilde{p}(S_i, S_j) - p_t(z, S_j) \,.$$

Then

$$\begin{aligned}
|r_{ij}(z)| &= \left| \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} (p_t(x, S_j) - p_t(z, S_j)) \right| \\
&\leq \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \sum_{y \in S_j} |p_t(x,y) - p_t(z,y)| \\
&\leq \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \left(\sum_{y \in S_j} \phi_0(y) \right)^{\frac{1}{2}} \left(\sum_{y \in S_j} \frac{1}{\phi_0(y)} |p_t(x,y) - p_t(z,y)|^2 \right)^{\frac{1}{2}} \quad \text{(Cauchy-Schwarz)} \\
&\leq \sqrt{\widetilde{\phi}_0(S_j)} \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \left(\sum_{y \in S_j} \frac{1}{\phi_0(y)} |p_t(x,y) - p_t(z,y)|^2 \right)^{\frac{1}{2}}
\end{aligned}$$

Another application of the Cauchy-Schwarz inequality yields

$$|r_{ij}(z)|^2 \leq \widetilde{\phi}_0(S_j) \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \sum_{y \in S_j} \frac{1}{\phi_0(y)} |p_t(x,y) - p_t(z,y)|^2 \tag{15}$$

Thus,

$$\begin{aligned} \sum_i \widetilde{\phi}_l(S_i) \widetilde{p}(S_i, S_j) &= \sum_i \sum_{z \in S_i} \phi_l(z) \widetilde{p}(S_i, S_j) \\ &= \sum_i \sum_{z \in S_i} \phi_l(z) (p_t(z, S_j) + r_{ij}(z)) \\ &= \lambda_l^t \widetilde{\phi}_l(S_j) + \sum_i \sum_{z \in S_i} \phi_l(z) r_{ij}(z) \end{aligned}$$

We therefore define $e_l \in \mathbb{R}^k$ by

$$e_l(S_j) = \sum_i \sum_{z \in S_i} \phi_l(z) r_{ij}(z) \,.$$

To prove the theorem, we need to bound

$$\|e_l\|^2_{1/\widetilde{\phi}_0} = \sum_j \frac{e_l(S_j)^2}{\widetilde{\phi}_0(S_j)} \,.$$

First, observe that by the Cauchy-Schwartz inequality,

$$e_l(S_j)^2 \leq \left(\sum_i \sum_{z \in S_i} \frac{\phi_l(z)^2}{\phi_0(z)} \right) \left(\sum_i \sum_{z \in S_i} r_{ij}(z)^2 \phi_0(z) \right) \,.$$

Now, since ϕ_l was normalized, this means that

$$e_l(S_j)^2 \leq \left(\sum_i \sum_{z \in S_i} r_{ij}(z)^2 \phi_0(z) \right) \,.$$

Invoking inequality (15), we conclude that

$$\begin{aligned} \|e_l\|^2_{1/\widetilde{\phi}_0} &\leq \sum_j \sum_i \sum_{z \in S_i} \phi_0(z) \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \sum_{y \in S_j} \frac{1}{\phi_0(y)} |p_t(x,y) - p_t(z,y)|^2 \\ &\leq \sum_i \sum_{z \in S_i} \phi_0(z) \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \sum_y \frac{1}{\phi_0(y)} |p_t(x,y) - p_t(z,y)|^2 \\ &\leq \sum_i \widetilde{\phi}_0(S_i) \sum_{z \in S_i} \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \frac{\phi_0(z)}{\widetilde{\phi}_0(S_i)} D_t^2(x,z) \\ &\leq \sum_i \widetilde{\phi}_0(S_i) \sum_{z \in S_i} \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \frac{\phi_0(z)}{\widetilde{\phi}_0(S_i)} \|\Psi_t(x) - \Psi_t(z)\|^2 \\ &\leq \sum_i \widetilde{\phi}_0(S_i) \sum_{z \in S_i} \sum_{x \in S_i} \frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)} \frac{\phi_0(z)}{\widetilde{\phi}_0(S_i)} \\ &\quad \times (\|\Psi_t(x) - c(S_i)\|^2 + \|\Psi_t(z) - c(S_i)\|^2 - 2\langle \Psi_t(x) - c(S_i), \Psi_t(z) - c(S_i) \rangle) \end{aligned}$$

By definition of $c(S_i)$,

$$\sum_{z\in S_i}\sum_{x\in S_i}\frac{\phi_0(x)}{\widetilde{\phi}_0(S_i)}\frac{\phi_0(z)}{\widetilde{\phi}_0(S_i)}\langle\Psi_t(x)-c(S_i),\Psi_t(z)-c(S_i)\rangle=0\,,$$

and therefore

$$\|e_l\|^2_{1/\widetilde{\phi}_0}\leq 2\sum_i\sum_{x\in S_i}\phi_0(x)\|\Psi_t(z)-c(S_i)\|^2\,.$$

As for right eigenvectors, the result follows from Equation 11 and the fact that the coarse-grained Markov chain is reversible with respect to $\widetilde{\phi}_0$. Indeed,

$$\begin{aligned}
\widetilde{P}\widetilde{\psi}_l(S_i) &= \sum_j\widetilde{p}(S_i,S_j)\widetilde{\psi}_l(S_j)\\
&= \sum_j\frac{\widetilde{p}(S_i,S_j)}{\phi_0(S_j)}\widetilde{\phi}_l(S_j)\text{ by Equation 11}\\
&= \sum_j\frac{\widetilde{p}(S_j,S_i)}{\phi_0(S_i)}\widetilde{\phi}_l(S_j)\text{ by reversibility}\\
&= \lambda_l^t\frac{\widetilde{\phi}_l(S_i)}{\widetilde{\phi}_0(S_i)}+\frac{e_l(S_i)}{\widetilde{\phi}_0(S_i)}\\
&= \lambda_l^t\widetilde{\psi}_l(S_i)+\frac{e_l(S_i)}{\widetilde{\phi}_0(S_i)}\text{ by Equation 11.}
\end{aligned}$$

If we set $f_l(S_i)=e_l(S_i)/\widetilde{\phi}_0(S_i)$, we conclude that

$$\|f_l\|^2_{\widetilde{\phi}_0}=\sum_i\frac{e_l(S_i)^2}{\widetilde{\phi}_0(S_i)^2}\widetilde{\phi}_0(S_i)=\|e_l\|^2_{1/\widetilde{\phi}_0}\leq 2\mathcal{D}\,.$$

■

REFERENCES

[1] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proceedings IEEE International Conference on Computer Vision*, volume 14, pages 975–982, 1999.

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Tran PAMI*, 22(8):888–905, 2000.

[3] M. Meila and J. Shi. A random walk's view of spectral segmentation. *AI and Statistics (AISTATS)*, 2001.

[4] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[5] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignement. Technical Report CSE-02-019, Department of computer science and engineering, Pennsylvania State University, 2002.

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

[7] D.L. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, May 2003.

[8] F. Chung. *Spectral graph theory*. Number 92. CBMS-AMS, May 1997.

[9] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*. To appear.

[10] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.

[11] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

[12] I.S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. *Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD)*, 2004.

[13] V. de Silva J.B. Tenenbaum and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[14] B. Nadler, S. Lafon, R.R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*. To appear.

[15] Private communication with R.R. Coifman.

[16] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–138, 1982.

[17] C.E. Priebe, D.J. Marchette, Y. Park, E.J. Wegman, J.L. Solka, D.A. Socolinsky, D. Karakos, K.W. Church, R. Guglielmi, R.R. Coifman, D. Lin, D.M. Healy, M.Q. Jacobs, and A. Tsao. Iterative denoising for cross-corpus discovery. In *Proceedings IEEE International Conference on Computer Vision pp. 975–982.*, 2004.

[18] B. Schölkopf, A.J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[19] V.N. Vapnik. *The nature of statistical learning theory*. 2nd edition.

[20] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2:121–167, 1998.

[21] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens (2004). A novel way of computing similarities between nodes of a graph, with application to collaborative recommendation. 2004. submitted to publication.

[22] R.R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*. To appear.

[23] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonics analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences*, 102(21):7432–7437, 2005.

Stéphane Lafon is a Software Engineer at Google. He received his B.Sc. degree in Computer Science from Ecole Polytechnique and his M.Sc. in Mathematics and Artificial Intelligence from Ecole Normale Supérieure de Cachan in France. He obtained his Ph.D. in Applied Mathematics at Yale University in 2004 and he was a Research Associate in the Applied Mathematics group during the year 2004-2005. He is currently with Google where his work focuses on the design, analysis and implementation of machine learning algorithms. His research interests are in data mining, machine learning and information retrieval.

Ann B. Lee is an Assistant Professor of Statistics at Carnegie Mellon University. She received her M.Sc. degree in Engineering Physics from Chalmers University of Technology in Sweden, and her Ph.D. degree in Physics from Brown University in 2002. She was a Research Associate in the Division of Applied Mathematics (Pattern Theory Group) at Brown University during the year 2001-2002, and a J.W. Gibbs Instructor and Assistant Professor of Applied Mathematics at Yale University from 2002-2005. In August 2005, she joined the Department of Statistics at Carnegie Mellon. Her research interests are in machine learning, statistical models in pattern analysis and vision, high-dimensional data analysis and multi-scale geometric methods.