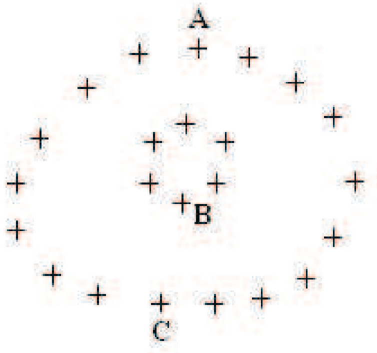
# MAT 585: Diffusion Maps

Amit Singer

March 3, 2015

In this lecture we introduce the diffusion maps framework for dimensionality reduction and data analysis. Recall the motivating problem for clustering data points that belong to two separated rings:



We intuitively see here two clusters corresponding to the two rings. Although the Euclidean distance between A and C is larger than the Euclidean distance between B and C, we would like A and C to belong to the same cluster. We therefore need to introduce a new metric, other than the Euclidean metric, that will assign a small distance between A and C and a larger distance between B and C.

**Notation:** $n$ denotes the number of data points and $x_1, x_2, \ldots, x_n$ stand for the data points themselves. The data points can be points in a Euclidean space, e.g., the points in $\mathbb{R}^2$ shown above, or abstract data objects such as web pages, or Facebook users for which it is not obvious how to represent them as vectors of real numbers. When considering data points in a Euclidean space, we denote the dimension of the ambient space by $p$, that is, $x_1, \ldots, x_n \in \mathbb{R}^p$. Typically $p$ is large, for example, 10 mega-pixels photos can be viewed as vectors in $\mathbb{R}^p$ with $p = 10^7$.

**Weighted undirected graph.** We describe the affinity or similarity between $x_i$ and $x_j$ by a real number $w_{ij}$. We require that the similarity is symmetric, that is, $w_{ij} = w_{ji}$ and also non-negative $w_{ij} \geq 0$. All similarities are organized in an $n \times n$ symmetric matrix $W$ whose $ij$ entry is $w_{ij}$. The similarities are user defined. A popular way to define similarities for points in $\mathbb{R}^p$ is using a non-negative function $K : \mathbb{R}_+ \to \mathbb{R}$ and a parameter $\epsilon > 0$ as

$$w_{ij} = K\left(\frac{\|x_i - x_j\|_{\mathbb{R}^p}}{\sqrt{\epsilon}}\right).$$

The function $K$ satisfies $K(u) \geq 0$, is usually monotonic decreasing and assumed to decay to 0. Examples include the Gaussian function $K(u) = e^{-u^2/2}$ and the characteristic function $K = 1_{[0,1]}$. This means that the similarity between $x_i$ and $x_j$ is $O(1)$ if $\|x_i - x_j\|_{\mathbb{R}^p}$ is $O(\sqrt{\epsilon})$, and is negligible otherwise (that is, for $\|x_i - x_j\|_{\mathbb{R}^p} \gg \sqrt{\epsilon}$). For abstract data sets other ways are used to define similarities. For example, we may say that two web pages have high similarity if there is a hyperlink between them, or two Facebook users are similar if one is a friend of the other.

This description of the data is equivalent to a weighted undirected graph $G = (V, E, W)$, where $V$ are the vertices of the graph, $E$ is the edge set and $W$ are the weights. There are $n$ vertices, corresponding to the data points. We put edges between data points with strictly positive similarity between them and each edge $(i, j)$ is weighted by $w_{ij}$.

**Random walk over the data points.** We now construct a discrete random walk (Markov chain) that corresponds to the weighted graph. There are $n$ states corresponding to the vertices, and our random walker can jump in a single time step from one node to another only if they are linked by an edge. The probability to jump from $x_i$ to $x_j$ is proportional to $w_{ij}$. Denote $X(t)$ the location, or state, of the random walker at time $t$. Then,

$$\Pr\{X(t+1) = x_j | X(t) = x_i\} = \frac{w_{ij}}{d_i},$$

where

$$d_i = \sum_{k=1}^{n} w_{ik}$$

is the weighted degree of node $i$. Set $A$ to be the $n \times n$ Markov transition probability matrix with

$$a_{ij} = \frac{w_{ij}}{d_i}.$$

(we assume there are no isolated nodes, i.e., $d_i > 0$ for all $i = 1, \ldots, n$) $A$ is indeed a transition probability matrix, since $a_{ij} \geq 0$ and $\sum_{j=1}^n a_{ij} = 1$ for $i = 1, 2, \ldots, n$. We say that $A$ was obtained from $W$ by a row stochastic normalization.

Now, let's do some linear algebra. Clearly, we can rewrite the matrix $A$ as

$$A = D^{-1}W, \tag{1}$$

where $D$ is an $n \times n$ diagonal matrix with $D_{ii} = d_i$. Although $A$ is not symmetric, it is similar to the symmetric matrix $S$ given by

$$S = D^{-1/2}WD^{-1/2} \tag{2}$$

through

$$A = D^{-1/2}(D^{-1/2}WD^{-1/2})D^{1/2} = D^{-1/2}SD^{1/2}. \tag{3}$$

Since $S$ is symmetric, it has a complete set of orthonormal eigenvectors and real eigenvalues. Suppose

$$Sv_l = \lambda_l v_l, \quad \text{for } l = 1, 2, \ldots, n.$$

We organize the orthonormal eigenvectors as columns of an $n \times n$ matrix $V$ and set $\Lambda$ to be a diagonal matrix with the eigenvalues on its diagonal. Then,

$$SV = V\Lambda,$$

or

$$S = V\Lambda V^T, \tag{4}$$

where we used orthonormality:

$$VV^T = V^TV = I.$$

The spectral decomposition of $S$ is therefore given by

$$S = \sum_{l=1}^n \lambda_l v_l v_l^T.$$

From (3) and (4) we see that

$$A = D^{-1/2}SD^{1/2} = D^{-1/2}V\Lambda V^T D^{1/2}.$$

We define

$$\Phi = D^{-1/2}V,$$

and

$$\Psi = D^{1/2}V,$$

so

$$A = \Phi\Lambda\Psi^T. \tag{5}$$

Notice that

$$\Phi^T\Psi = V^T D^{-1/2} D^{1/2} V = V^T V = I. \tag{6}$$

This means that the columns of $\Phi$, denoted $\phi_1, \ldots, \phi_n$ and the columns of $\Psi$, denoted $\psi_1, \ldots, \psi_n$ form a bi-orthogonal system. That is,

$$\phi_l = D^{-1/2}v_l,$$

$$\psi_l = D^{1/2}v_l,$$

and

$$\langle \phi_l, \psi_m \rangle = \delta_{lm}.$$

From (5) and (6) it follows that

$$A\Phi = \Phi\Lambda$$

and

$$\Psi^T A = \Lambda\Psi^T.$$

In other words, $\phi_l$ are the right eigenvectors of $A$

$$A\phi_l = \lambda_l\phi_l,$$

$\psi_l$ are the left eigenvectors of $A$

$$\psi_l^T A = \lambda_l\psi_l^T,$$

and $A$ has the decomposition

$$A = \sum_{l=1}^{n} \lambda_l \phi_l \psi_l^T.$$

We notice that there is one eigenvector/eigenvalue pair which is easy to obtain regardless of the specific form of $A$. Indeed, the all-ones vector $\mathbf{1} = (1\ 1\ \cdots\ 1)^T$ satisfies

$$A\mathbf{1} = \mathbf{1},$$

which means that $\mathbf{1}$ is an eigenvector of $A$ and the associated eigenvalue is $\lambda = 1$. The corresponding left eigenvector is $\psi_1 = (d_1, d_2, \ldots, d_n)^T$. Next, we show that all eigenvalues of $A$ are in the interval $[-1, 1]$.

**Fact:** All the eigenvalues of $A$ satisfy $|\lambda| \leq 1$.

**Proof:** Let $\lambda$ be an eigenvalue for $A$ associated to the eigenvector $\phi$. Consider $i_0 = \text{argmax}_{1\leq i\leq n} |\phi(i)|$. Then,

$$\lambda\phi(i_0) = \sum_{j=1}^{n} a_{i_0 j}\phi(j)$$

and

$$|\lambda| \leq \sum_{j=1}^{n} a_{i_0 j} \left| \frac{\phi(j)}{\phi(i_0)} \right| \leq \sum_{j=1}^{n} a_{i_0 j} = 1.$$

**Definition of Diffusion Map.** We are now ready to define the diffusion mapping. Let $t > 0$ and suppose the eigenvalues are sorted in decreasing order of magnitude: $1 = \lambda_1 \geq |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$. The diffusion map $\Phi_t : V \mapsto \mathbb{R}^n$ is defined as

$$\Phi_t(x_i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix}. \tag{7}$$

The diffusion map assigns to each data point $n$ coordinates, given by the evaluation of the eigenvectors at that point (and scaled by the eigenvalues to the $t$ power). As such, for large data sets there is not much of dimensionality reduction happening, since $n$ can be quite large. We notice that the first coordinate is redundant, since $\phi_1$ is proportional to the all-ones vector, all data points share the same first coordinate. We can therefore neglect this coordinate as it does not help us to differentiate between data points. So we can equivalently consider the diffusion mapping $\Phi_t : V \mapsto \mathbb{R}^{n-1}$:

$$\Phi_t(x_i) = \begin{pmatrix} \lambda_2^t \phi_2(i) \\ \lambda_3^t \phi_3(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix}. \tag{8}$$

In order to substantially reduce the dimensionality, we define the truncated diffusion mapping $\Phi_t^\delta : V \mapsto \mathbb{R}^{m-1}$, where $m$ is a function of $t$ and $\delta$ is defined as the largest integer for which $|\lambda_m|^t > \delta$ and $0 < \delta < 1$, so $|\lambda_{m+1}|^t \leq \delta$

$$\Phi_t^\delta(x_i) = \begin{pmatrix} \lambda_2^t \phi_2(i) \\ \lambda_3^t \phi_3(i) \\ \vdots \\ \lambda_m^t \phi_m(i) \end{pmatrix}. \tag{9}$$

## Examples

In order to get some intuition about the diffusion map, we first consider a few simple examples.

**The ring graph.** The ring graph on $n$ vertices has $n$ edges: $(1,2), (2,3), \ldots, (n-1, n)$ and $(n, 1)$. The edges are of the form $(i, i+1)$ where addition is modulo $n$. Suppose that all edges are associated with the same weight. Then, the corresponding $n \times n$ Markov transition matrix $A$ has the form:

$$A = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \cdots & 0 & \frac{1}{2} & 0 \end{pmatrix} \tag{10}$$

We notice that $A$ is a cyclic matrix, that is, all rows are cyclic shifts of the first row. As a result, the eigenvectors of this matrix is the Fourier basis.

**Fact:** The vectors $v_l$ $(l = 1, \ldots, n)$ given by

$$v_l(r) = e^{2\pi i l r/n}$$

are eigenvectors of $A$, that is, $Av_l = \lambda_l v_l$.

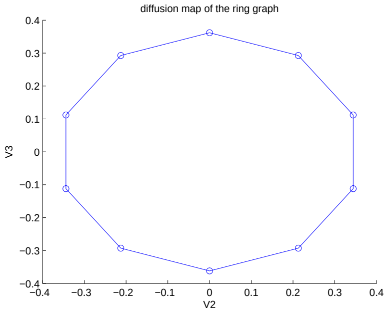
**Proof:**

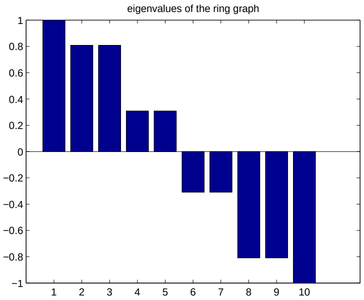
$$(Av_l)(r) = \frac{1}{2} e^{2\pi i l (r+1)/n} + \frac{1}{2} e^{2\pi i l (r-1)/n} = \cos\left(\frac{2\pi l}{n}\right) v_l(r).$$

It is easy to check that $v_1, \ldots, v_n$ are linearly independent since they are orthogonal vectors. The eigenvalues are $\lambda_l = \cos\left(\frac{2\pi l}{n}\right)$. We see that all eigenvalues are between $-1$ and $1$ as expected, the eigenvalue $\lambda = 1$ is simple (corresponding to $l = 0$, or equivalently, $l = n$), the eigenvalue $-1$ appears whenever $n$ is even, and all other eigenvalues (beside 1 and -1) have multiplicity 2 (corresponding to $l$ and $-l$). The eigenvectors $v_l$ are complex valued, but due to the multiplicity 2 we can generate the eigenvectors $\frac{1}{2}(v_l + v_{-l}) = \cos(\frac{2\pi lr}{n})$ and $\frac{1}{2i}(v_l - v_{-l}) = \sin(\frac{2\pi lr}{n})$. In particular, the truncated diffusion mapping of a point $x_r$ using the two eigenvectors corresponding to $l = 1$ gives

$$x_r \mapsto \begin{pmatrix} \lambda_1^t \cos(\frac{2\pi r}{n}) \\ \lambda_1^t \sin(\frac{2\pi r}{n}) \end{pmatrix}, \quad \text{for } r = 1, 2, \ldots, n.$$

This mapping traces a circle in $\mathbb{R}^2$, in agreement with the ring structure of this graph.





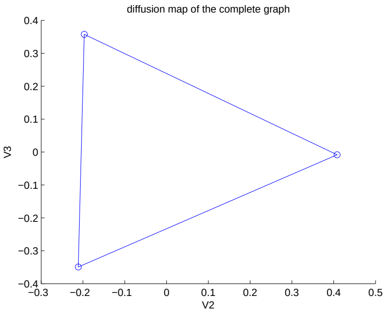
**The complete graph.** The complete graph on $n$ nodes has all possible $\binom{n}{2}$ edges. Suppose all edges are assigned the same weight. The matrix $A$ is given by

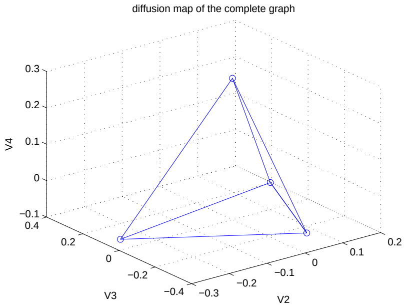
$$A = \begin{pmatrix} 0 & \frac{1}{n-1} & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \frac{1}{n-1} & 0 & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \frac{1}{n-1} & \cdots & \frac{1}{n-1} & 0 & \frac{1}{n-1} \\ \frac{1}{n-1} & \cdots & \frac{1}{n-1} & \frac{1}{n-1} & 0 \end{pmatrix}$$

Again, we see that $A$ is a cyclic matrix so we can conclude the form of its eigenvectors as before, but here we can make a simple observation about the matrix $A$:

$$A = \frac{1}{n-1}\mathbf{1}\mathbf{1}^T - \frac{1}{n-1}I.$$

The matrix $\mathbf{1}\mathbf{1}^T$ is the all-ones matrix and is of rank 1. We conclude that the all-ones vector $\mathbf{1}$ is an eigenvector whose eigenvalue is 1, and all vectors in the $n-1$-dimensional orthogonal subspace are eigenvectors with eigenvalue $\lambda = -\frac{1}{n-1}$. Since diffusion map disregards the all-ones vector, and all other eigenvalues are the same, it would need to make use of all remaining $(n-1)$ eigenvectors. This means that no dimensionality reduction can be achieved for the complete graph: all $n-1$ coordinates are needed, and the embedding is a simplex (e.g., for $n=3$ we get a triangle in $\mathbb{R}^2$, for $n=4$ a tetrahedron in $\mathbb{R}^3$, etc.). In the embedding, all nodes are equidistant from one another, in agreement with the fact that each node is connected to all other nodes and there is no notion of ordering or structure.

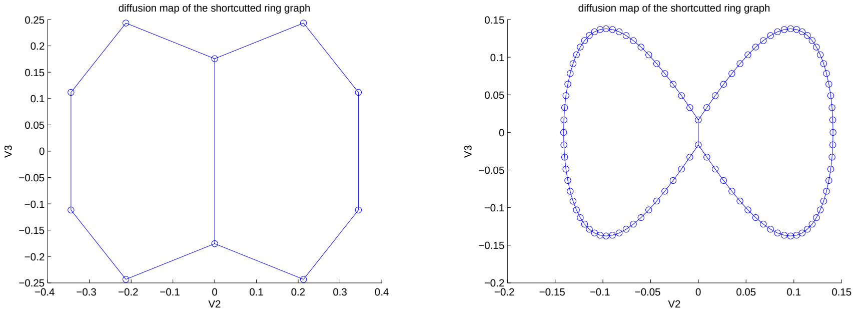




**Ring graph with a shortcut.** Suppose $n$ is even and add one edge to the ring graph between nodes 1 and $\frac{n}{2}+1$. We assign equal weights to all edges as before, so the jump probabilities from nodes 1 and $\frac{n}{2}+1$ are $(1/3, 1/3, 1/3)$ while the jump probabilities from all other nodes are $(1/2, 1/2)$. The diffusion map cannot be calculated analytically as before, so we use MATLAB to find the embedding in $\mathbb{R}^2$ (see Figure).

We see that compared to the circular embedding of the ring graph, nodes 1 and $\frac{n}{2}+1$ get closer together. In fact, when increasing the number of nodes from $n=10$ to $n=100$, this pair of nodes get even closer. Notice that in the embedding, the length associated to all edges is more or less the same, expressing the equal weights assigned to all edges. Increasing the number of nodes shortens this length and brings the end nodes of the shortcut edge closer together. (Can you make this hand-waving mathematically rigorous?)

**Two disjoint ring graphs.** We finish with our motivating example of two disjoint rings graphs, that is, there are no edges linking the two rings. Suppose that one ring has $n_1$ nodes, while the other has $n_2$ nodes. The



(a) $n = 10$

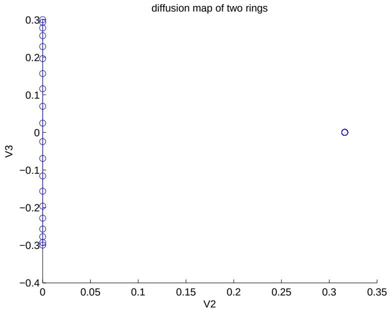
(b) $n = 100$

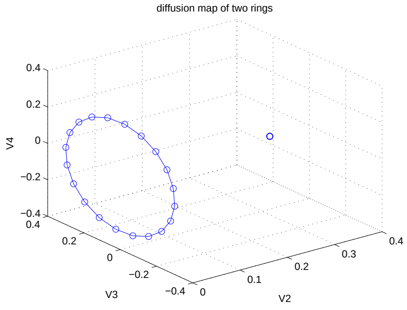
matrix $A$ takes the form of a block matrix

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

where $A_1$ is $n_1 \times n_1$ and $A_2$ is $n_2 \times n_2$ (and 0 stands for block matrices of zeros). As such, the eigenvectors of $A$ take the form $\begin{pmatrix} v_1 \\ 0 \end{pmatrix}$ where $v_1$ is some eigenvector of $A_1$ and 0 is a zero vector of length $n_2$, or $\begin{pmatrix} 0 \\ v_2 \end{pmatrix}$ where $v_2$ is some eigenvector of $A_2$ and 0 is a zero vector of length $n_1$. It follows that the eigenvalue $\lambda = 1$ has multiplicity 2, corresponding to the eigenvectors $(1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$ and $(0, 0, \ldots, 0, 1, 1, \ldots, 1)^T$. In diffusion maps, we usually declare the all-ones vector to be the first eigenvector $\phi_1$, so choosing the second eigenvector orthogonal to $\mathbf{1}$ in this case gives (a normalized version of) $(n_2, n_2, \ldots, n_2, -n_1, -n_1, \ldots, -n_1)^T$. We see that this vector cluster all nodes of the first ring to a single point (whose coordinate is $n_2$) and all nodes of the second ring to a single point (whose coordinate is $-n_1$). Thus, the first coordinate of the diffusion mapping automatically picks the dominant intrinsic structure. The third and fourth eigenvectors correspond to the ring graph that has more nodes, thus revealing the secondary structure of this graph. The structure of the smaller ring is expressed in the remaining eigenvectors, but cannot be visualized using the diffusion map embedding in 3D.

**Diffusion Distance.** Recall that we view the data points $x_1, x_2, \ldots, x_n$ as nodes of a weighted undirected graph $G = (V, E)$, where the weight $w_{ij}$ is a measure of the similarity between $x_i$ and $x_j$. We further defined a





Markov probability transition matrix $A$ as $A = D^{-1}W$. We used the right eigenvectors of $A$ and the corresponding eigenvalues to define the diffusion map in the following way: if $A\phi_l = \lambda_l \phi_l$, then the diffusion map $\Phi_t : V \mapsto \mathbb{R}^n$ is defined as

$$\Phi_t(x_i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix}.$$

The diffusion map is an embedding of a data set in an Euclidean space which is equipped with an inner product:

$$\langle \Phi_t(x_i), \Phi_t(x_j) \rangle_{\mathbb{R}^n} = \sum_{l=1}^{n} \lambda_l^{2t} \phi_l(i) \phi_l(j). \qquad (11)$$

We see that the Gram matrix $B$ of inner products, i.e., $B_{ij} = \langle \Phi_t(x_i), \Phi_t(x_j) \rangle_{\mathbb{R}^n}$ is given by

$$B = \Phi \Lambda^{2t} \Phi^T. \qquad (12)$$

The matrix $B$ is positive semidefinite (PSD), i.e., it is a symmetric matrix and all its eigenvalues are non-negative, or alternatively, it defines a non-negative quadratic form. The converse is also true, every PSD matrix is a Gram matrix of points in Euclidean space whose dimension is the rank of the matrix.

We would now like to understand the meaning of the Euclidean distances $\|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^n}$ to which we refer as the *diffusion distances* denoted $D_t(x_i, x_j)$:

$$D_t(x_i, x_j) = \|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^n}.$$

Specifically, we will show the following probabilistic interpretation of the diffusion distance: The diffusion distance between $x_i$ and $x_j$ is a weighted $\ell^2$

distance between the probability clouds after $t$ time steps of random walks that start at $x_i$ and $x_j$.

From the interpretation of the matrix $A$ as a Markov transition probability matrix

$$a_{ij} = \Pr\{X(t+1) = x_j | X(t) = x_i\}$$

it follows that

$$a_{ij}^t = \Pr\{X(t) = x_j | X(0) = x_i\}.$$

That is, the elements of $A^t$ give us the probability to get from one state to another in $t$ time steps. For example,

$$\begin{aligned} a_{ij}^2 &= \sum_{k=1}^{n} a_{ik} a_{kj} \\ &= \sum_{k=1}^{n} \Pr\{X(1) = x_k | X(0) = x_i\} \Pr\{X(2) = x_j | X(1) = x_k\} \\ &= \Pr\{X(2) = x_j | X(0) = x_i\}. \end{aligned}$$

We refer to the $i$'th row of the matrix $A^t$, denoted $A_{i,\cdot}^t$ as the probability cloud of a random walk that starts at $x_i$ after $t$ steps. We can express $A^t$ using the decomposition of $A$. Indeed, from

$$A = \Phi \Lambda \Psi^T$$

and the fact that $\Psi^T \Phi = I$ (i.e., $\{\phi_l\}_{l=1}^n$ and $\{\psi_l\}_{l=1}^n$ form a bi-orthogonal system) we get

$$A^2 = \Phi \Lambda \Psi^T \Phi \Lambda \Psi^T = \Phi \Lambda^2 \Psi^T,$$

and generally,

$$A^t = \Phi \Lambda^t \Psi^T.$$

Written componentwise, this is equivalent to

$$a_{ik}^t = \sum_{l=1}^{n} \lambda_l^t \phi_l(i) \psi_l(k).$$

We now calculate the weighted $\ell^2$ distance between the probability clouds $A_{i,\cdot}^t$ and $A_{j,\cdot}^t$. For the weights we choose $1/d_k$, that is, inversely proportional to the vertex degrees, so nodes with higher degrees have a lesser weight in this weighted $\ell^2$ space:

$$\begin{aligned}
\|A^t_{i,\cdot} - A^t_{j,\cdot}\|^2_{\ell^2(\mathbb{R}^n,1/d)} &= \sum_{k=1}^n (a^t_{ik} - a^t_{jk})^2 \frac{1}{d_k} \\
&= \sum_{k=1}^n \left[\sum_{l=1}^n \lambda_l^t \phi_l(i)\psi_l(k) - \lambda_l^t \phi_l(j)\psi_l(k)\right]^2 \frac{1}{d_k} \\
&= \sum_{k=1}^n \sum_{l,r=1}^n \lambda_l^t \lambda_r^t (\phi_l(i) - \phi_l(j))(\phi_r(i) - \phi_r(j)) \frac{\psi_l(k)\psi_r(k)}{d_k} \\
&= \sum_{l,r=1}^n \lambda_l^t \lambda_r^t (\phi_l(i) - \phi_l(j))(\phi_r(i) - \phi_r(j)) \sum_{k=1}^n \frac{\psi_l(k)\psi_r(k)}{d_k} \\
&= \sum_{l,r=1}^n \lambda_l^t \lambda_r^t (\phi_l(i) - \phi_l(j))(\phi_r(i) - \phi_r(j)) \delta_{lr} \\
&= \sum_{l=1}^n \lambda_l^{2t} (\phi_l(i) - \phi_l(j))^2 \\
&= D_t^2(x_i, x_j).
\end{aligned}$$

Thus, we proved

$$D_t(x_i, x_j) = \|A^t_{i,\cdot} - A^t_{j,\cdot}\|_{\ell^2(\mathbb{R}^n,1/d)}. \tag{13}$$