Graph Fourier MMD for Signals on Graphs

***Abstract*—While numerous methods have been proposed for computing distances between probability distributions in Euclidean space, relatively little attention has been given to computing such distances for distributions on graphs. However, there has been a marked increase in data that either lies on graph (such as protein interaction networks) or can be modeled as a graph (single cell data), particularly in the biomedical sciences. Thus, it becomes important to find ways to compare signals defined on such graphs. Here, we propose Graph Fourier MMD (GFMMD), a novel distance between distributions and signals on graphs. GFMMD is defined via an optimal witness function that is both smooth on the graph and maximizes difference in expectation between the pair of distributions on the graph. We find an analytical solution to this optimization problem as well as an embedding of distributions that results from this method. We also prove several properties of this method including scale invariance and applicability to disconnected graphs. We showcase it on graph benchmark datasets as well on single cell RNA-sequencing data analysis. In the latter, we use the GFMMD-based gene embeddings to find meaningful gene clusters. We also propose a novel type of score for gene selection called *gene localization score* which helps select genes for cellular state space characterization.**

I. Introduction

Here, we address the question of how to organize and compare signals on graphs in such a way that accounts for geometric structure on their underlying space. In particular, given a weighted graph $\mathcal{G} = (V, \mathcal{E}, w)$ and a set of functions $\{f_i\}_i$ on the vertices: $f_i : \mathcal{V} \to \mathbb{R}$, how can we structure and analyze these signals? We will first consider the case when f_i is a probability mass function and extend the framework to arbitrary signals. This has a very natural applications to many modern datasets.

We present a new distance that belongs to the family of integral probability metrics [17] called maximal mean discrepancy or MMD. Integral probability metrics are distances between probability distributions that are characterized by a witness function that maximizes the discrepancy between distributions in expectation. MMDs have further structure in the witness function, requiring that they come from a Reproducing Kernel Hilbert Space. Our notion of MMD, that we call *Graph Fourier MMD* (GFMMD), is a distance between signals on a data graph that is found by analytically solving for an optimal witness function. GFMMD borrows notions from optimal transport, but does not require a distance metric, and thus generalizes to any undirected graph with nonnegative affinities. Furthermore, through the use of Chebyshev polynomials [13], GFMMD can be computed rapidly, and has a closed-form solution. We demonstrate its potential on toy datasets as well as single cell data, where we use it to identify gene modules. In the single cell setting, we focus on the application of embedding a set of genes on a graph of cells, as created from single cell RNA-sequencing data, and also in measuring whether the expression of a gene is localized (i.e., characteristic of a subpopulation of cells) or global like a house-keeping gene.

Our main contributions are as follows: 1) We define Graph Fourier MMD as a distance between signals on arbitrary graphs, and prove that it is both an integrable probability metric and maximum mean discrepancy. 2) We derive an exact analytical solution for GFMMD which can be approximated in $\mathcal{O}(n(\log n + m^2))$ time to calculate all pairwise-distances between distributions, where n is the number of vertices of the graph and m is the number of signals. 3) We derive a feature map for GFMMD that allows for efficient embeddings and dimensionality reduction. 4) We provide an efficient Chebyshev approximation method for computing GFMMD among a set of signals. 5) We showcase application of GFMMD to single cell RNA-sequencing data. In short, Graph Fourier MMD is a simple, interpretable, and above all else effective method of comparing abstract distributions.

A. Preliminaries

a) Integral probability metrics: IPMs [15], [17] constitute a family of distances between probability distributions. They are often used when dealing with empirical samples (datasets) sampled from a continuous space. In contrast, the alternative class of ϕ-divergences (such as KL-divergence) is often less useful as a measure between empirical samples with poor behavior when the domains do not overlap. In contrast to ϕ-divergences, integral probability metrics are defined over a metric space, this allows for a reasonable distance between distributions with non-overlapping support.

Definition 1. *Given a metric space $(\mathcal{X}, d)$, a family $\mathcal{F}$ of measurable, bounded functions on $\mathcal{X}$, and two measures P and Q on $\mathcal{X}$, the IPM between P and Q is defined as*

$$\gamma_{\mathcal{F}}(P, Q) \triangleq \sup_{f \in \mathcal{F}} \mathbb{E}_P(f) - \mathbb{E}_Q(f).$$

Here, $\mathcal{F}$ is a family of "witness function" since it emphasizes the differences between P and Q, choosing a certain $\mathcal{F}$ determines the IPMs. For certain classes of $\mathcal{F}$, the resulting distance is called an *kernel Maximum Mean Discrepancy* (MMD) [9]. If $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) of functions on $\mathcal{X}$ (equipped with norm $\|(\cdot)\|_{\mathcal{H}}$), then the IPM corresponding to $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ is an MMD. Numerous distances between distributions are IPMs, given a suitable choice of $\mathcal{F}$. For example, the Wasserstein distance is an IPM where $\mathcal{F}$ corresponds to the family of Lipschitz functions.

b) Wasserstein Distance: The Earth Mover's Distance (EMD), also known as the 1-Wasserstein distance, is a distance between probability distributions designed to measure the least amount of "work" it takes to move mass from one distribution to another. Formally, we are given two distributions P and Q on a measure space $(\Omega, \mathcal{F}, \mu)$ and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Most commonly, Ω might be a Riemann manifold, $\mathbb{R}^d$, or in our case, a finite graph. We define the space of couplings of P and Q, denoted $\Pi(P, Q)$ to be the set of joint probability distributions whose marginals are equal to P and Q.

Definition 2. *The 1-Wasserstein Distance between P and Q is defined to be:*

$$W(P, Q) \triangleq \min_{\pi \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)\pi(dx, dy).$$

The supremum joint distribution π would then be called the *optimal transport plan*. In the case that Ω is finite (say of size n), $\Pi(P, Q)$ could be thought of as the set of $n \times n$ matrices π for which $\pi\mathbf{1} = P, \mathbf{1}^T\pi = Q$. Then we could represent distances in a $n \times n$ matrix D, and the EMD is given by $\min_\pi \pi \cdot D$. Typical solutions to EMD in its *primal form* are found using linear programming. The Kantorivich-Rubinstein Theorem, however, provides a dual formulation in terms of smooth functions:

Theorem 1. *(Kantorovich-Rubinstein) The EMD is an IPM with $\mathcal{F}$ the space of 1-Lipschitz functions*

$$W(P, Q) = \sup_{\|f\|_{\leq 1}} \mathbb{E}_P(f) - \mathbb{E}_Q(f).$$

We refer to [6] for a proof of the previous theorem. Intuitively, we can think of suitable functions f as being varying slowly over $\mathcal{X}$. The 1-Lipschitz constraint prevents witness functions from behaving too erratically over the space. As we will see, duality provides a valuable intuition for using smooth functions to compare functionals in abstract spaces.

c) The Graph Laplacian: For a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ on n vertices, we have a number of associated matrices. The first of which is an adjacency / affinity matrix $\mathbf{A}$ for which, given vertices a and b, $\mathbf{A}(a, b) = w(a, b)$; for our purposes, we assume $w(a, b) \geq 0$. In the case when $\mathcal{V}$ belongs to a metric space $(\mathcal{X}, d)$, we have an associated distance matrix M for which $M(a, b) = d(a, b)$ for all $a, b \in \mathcal{V}$. Oftentimes, the affinity matrix $\mathbf{A}$ is generated by a nonlinear kernel function $k(\cdot)$ so that $\mathbf{A}(a, b) = k(M(a, b))$. For our purposes, if $\mathbf{A}$ is generated in this way, we will call $\mathcal{G}$ a *affinity graph*. There is also a diagonal degree matrix for which $\mathbf{D}(a, a) = \sum_{b \in \mathcal{V}} w(a, b)$. Finally, we define the combinatorial Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$. It can be shown that for any function on the vertices f, $f^T\mathbf{L}f = \sum_{(a,b)\in\mathcal{E}} w(a, b)(f(a) - f(b))^2$. From this, it's clear that $\mathbf{L}$ is positive semi-definite, and thus has a spectrum $\{(\lambda_1, \psi_1)...(\lambda_n, \psi_n)\}$, where $\lambda_1 \geq ... \geq \lambda_n$.

d) Effective Resistances: In [16], graphs are regarded as electrical circuits with edge weights providing capacities. In such a graph, the *effective resistance* $\mathrm{Re}(a, b)$ between vertices a and b is equal to $\mathrm{Re}(a, b) = \|\mathbf{L}^{-\frac{1}{2}}\delta_a - \mathbf{L}^{-\frac{1}{2}}\delta_b\|_2^2$, where δ_a, δ_b are the one-hot encodings of vertices a and b, respectively. Effective resistances provide valuable information about a graph. For instance, Spielman & Srivastava use effective resistances between adjacent vertices to sparsify a graph. We will show, if we view δ_a, δ_b as probability densities concentrated at a, b, Graph Fourier MMD provides an extension of resistances to arbitrary probability distributions on graphs. Effective resistences have been shown to be related to commute times, thus this provides a generalization of commute time when the initial and final position are not localized to a single node [1].

B. Related Work

The closest related work is that of [19], which constructs a family of spectral distances between graph signals based on weighted Fourier transforms. Algebraically, our distance $\mathcal{GFMMD}$ resembles a special case of these distances. Another similar distance is Diffusion EMD [18], which involves diffusion graph signals to different scales using a diffusion operator (similar to that of a diffusion map [3]) to create multiscale density estimates of the data. Then Diffusion EMD computes weighted L^1 distance between the multiscale density estimates of different signals. While this method is faster than most primal methods for EMD computation, it can be inaccurate unless the graph is significantly large.

In [11], [12], [8], the authors consider the EMD between distributions defined on a *distance* graph, that is the edge weights define the cost of moving mass from one node to another. The authors in [11], [12] provide a closed-form solution that relies on a graph shortest path distance. In this setting, there is no sparse approximation to diffusion distances in terms of graph shortest path. We consider a different problem where the edges of the graph are affinities. Among methods for MMD, the most common method has been a sampling based method that also forms a 2-sample Kernel test based on defining a kernel between empirical observations [9]. Note that semantically this takes distances between point clouds themselves by modeling them as a data graph with vertices as points. We define a method of taking signals which generalizes to an arbitrary graph, on a point cloud or otherwise, and demonstrate its effectiveness both when the graph lies in a metric space and when adjacencies are binary.

II. Methods

Given two probability distributions P and Q on an arbitrary graph, we are interested in taking a meaningful distance between them in a way that incorporates graph structure. Unlike the distance setting, there is no notion of Lipschitzness. However, there is still a notion of smoothness. Indeed, we define Graph Fourier MMD as the MMD induced by witness functions f which are *smooth over the graph*. That is, they have a low value in $f^T\mathbf{L}f$.

Definition 3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ be a finite graph with Laplacian $\mathbf{L}$ and P, Q be two bounded probability distributions on $\mathcal{V}$. The Graph Fourier MMD between P and Q is*

$$\mathcal{GFMMD}(P, Q) \triangleq \max_{f: f^T\mathbf{L}f \leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f).$$

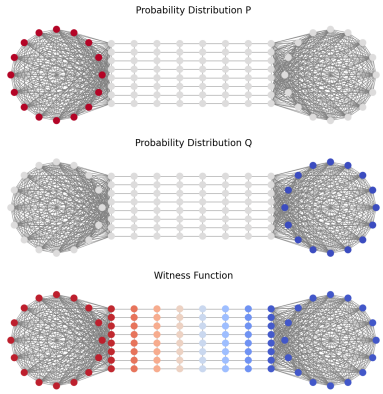

Fig. 1: The witness function f is positively activated over the left cluster, negatively activated over the right cluster, and varies smoothly along the middle segment. f attests to where P and Q differ, but smoothness requires that it do so while incorporating graph structure.

Note that this definition holds for any construction of a positive semi-definite Laplacian matrix $\mathbf{L}$. We can show that, under reasonable conditions, GFMMD is finite and simple to compute. First, we need to establish a property which indicates that P and Q do not differ on the scale of connected components.

Definition 4. *Two probability distributions P and Q are said to have equal component mass if, for all connected components S of $\mathcal{G}$, $P(v \in S) = Q(v \in S)$.*

Theorem 2. *Let P and Q be bounded probability distributions defined on $\mathcal{V}$. If P and Q have equal component mass, then $\mathcal{GFMMD}(P,Q) = \|\mathbf{L}^{-\frac{1}{2}}(P-Q)\|_2$. And otherwise, $\mathcal{GFMMD}(P,Q) = +\infty$.*

Proof. Suppose first that P and Q do not have the equal mass property. Then there exists a connected component S for which,

$$\sum_{v \in S} P(v) < \sum_{v \in S} Q(v)$$

In particular, we can write $\sum_{v \in S} P(v) = \sum_{v \in S} Q(v) - c$ for some $c > 0$. Now, let f_α be a signal such that $f_\alpha(v) = \alpha$ if $v \in S$ and $f_\alpha(v) = 0$ otherwise. Then certainly, $f_\alpha^T \mathbf{L} f_\alpha = 0$, since it is known that indicator functions for connected components arer in the null space of $\mathbf{L}$. And so $f_\alpha^T \mathbf{L} f_\alpha \leq 1$, yet,

$$\begin{aligned}
\mathbb{E}_P(f_\alpha) - \mathbb{E}_Q(f_\alpha) &= \sum_{v \in \mathcal{V}} P(v) f_\alpha(v) - \sum_{v \in \mathcal{V}} Q(v) f_\alpha(v) \\
&= \sum_{v \in S} \alpha P(v) - \sum_{v \in S} \alpha Q(v) \\
&= \alpha c
\end{aligned}$$

S GFMMD is defined as $\sup_{f^T \mathbf{L} f \leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f)$, we have $\mathcal{GFMMD}(P,Q) \geq \alpha c$. Taking $\alpha \to \infty$, we have $\mathcal{GFMMD}(P,Q) = +\infty$.

Now suppose that P and Q do have the equal mass property. If we let $\mathbb{I}_{S_1} ... \mathbb{I}_{S_m}$ be indicator functions for connected components $S_1 ... S_m$, the equal mass property insists that $P^T \mathbb{I}_{S_i} = Q^T \mathbb{I}_{S_i}$ for all i. And thus, $(P-Q)^T \mathbb{I}_{S_i} = 0$. Since it is known that these indicator functions form a basis for the kernel of $\mathbf{L}$, it follows that $P - Q \in \ker(\mathbf{L})^\perp$. Now, any function f such that $f^T \mathbf{L} f$ can be broken up into $f = f_1 + f_2$, where $f_1 \in \ker(\mathbf{L})$ and $f_2 \in \ker(\mathbf{L})$. Finally, observe that we can view P and Q as probability vectors indexed over $\mathcal{V}$. And so,

$$\begin{aligned}
\mathbb{E}_P(f) - \mathbb{E}_Q(f) &= P^T f - Q^T f = (P-Q)^T (f_1 + f_2) \\
&= (P-Q)^T f_1 + (P-Q)^T f_2 \\
&= (P-Q)^T f_2
\end{aligned}$$

Furthermore, $f^T \mathbf{L} f = f_2^T \mathbf{L} f_2$. Combined, these observations tell us that we may assume, without loss of generality, that $f \in \ker(\mathbf{L})^\perp$. And thus,

$$\mathcal{GFMMD}(P,Q) = \sup_{\substack{f^T \mathbf{L} f \leq 1 \\ f \in \ker(\mathbf{L})^\perp}} (P-Q)^T f$$

Now, for any such f, we can define $y = \mathbf{L}^{1/2} f$. And thus, $f^T \mathbf{L} f = f^T \mathbf{L}^{1/2} \mathbf{L}^{1/2} f = \|y\|_2^2$. Furthermore, since $f \in \ker(\mathbf{L})^\perp$, $f = \mathbf{L}^{-\frac{1}{2}} y$ as well. Thus,

$$\mathcal{GFMMD}(P,Q) = \sup_{y : \|y\|_2^2 \leq 1} (P-Q)^T \mathbf{L}^{-\frac{1}{2}} y$$

Which clearly, by Cauchy Schwarz, is simply equal to $\|(P-Q)^T \mathbf{L}^{-\frac{1}{2}}\|_2^2 = \|\mathbf{L}^{-\frac{1}{2}}(P-Q)\|^2$, as desired. □

The effect of equal component mass is highly intuitive: we would expect "infinite effort" to move a probability distribution between disconnected sets of vertices. GFMMD possesses a set of convenient properties. Namely, we have a representation in terms of an explicit feature map $\mathbf{L}^{-\frac{1}{2}}$. So to compute pairwise distances, it is sufficient to apply the feature map and then take Euclidean distances. Also, the distance value is a true distance (particularly an MMD).

Lemma 3. *(i) $\mathcal{GFMMD}(\cdot,\cdot)$ defines a valid distance on the probability distributions acting on $\mathcal{V}$. Furthermore, (ii) $\mathcal{GFMMD}(P,Q)$ is a Maximum Mean Discrepancy with explicit feature map $\mathbf{L}^{-\frac{1}{2}}$ in the finite case.*

Proof. Let $\mathcal{P}$ denote the set of probability distributions on $\mathcal{V}$. Nonnegativity and symmetry of $\mathcal{GFMMD}$ are simple to verify directly from the definition. We now show the triangle inequality. Fix $P, Q, R \in \mathcal{P}$. Letting $\mathcal{F} = \{f : f^T \mathbf{L} f \leq 1\}$,

$$\begin{aligned}
\mathcal{GFMMD}(P,Q) &= \sup_{f \in \mathcal{F}} \mathbb{E}_P(f) - \mathbb{E}_Q(f) \\
&= \sup_{f \in \mathcal{F}} \mathbb{E}_P(f) - \mathbb{E}_R(f) + \mathbb{E}_R(f) - \mathbb{E}_Q(f) \\
&\leq \sup_{f \in \mathcal{F}} \mathbb{E}_P(f) - \mathbb{E}_R(f) + \sup_{f \in \mathcal{F}} \mathbb{E}_R(f) - \mathbb{E}_Q(f) \\
&= \mathcal{GFMMD}(P,R) + \mathcal{GFMMD}(Q,R)
\end{aligned}$$

We conclude (i) by showingd that $\mathcal{GFMMD}(P,Q) = 0$ if and only if $P = Q$. The first direction is trivial, since for all $f \in \mathcal{F}$, $\mathbb{E}_P(f) - \mathbb{E}_Q(f) = 0$. For the other direction, suppose

that $\mathcal{GFMMD}(P,Q) = 0$. Then by theorem 2, P and Q necessarily have equal component mass. From this, it follows that for all components S, $\mathbb{I}_S^T P = \mathbb{I}_S^T Q$, so $P = Q$ over their projections onto $\ker(\mathbf{L})$. But since $\mathcal{GFMMD}(P,Q) = \|\mathbf{L}^{-\frac{1}{2}}P - \mathbf{L}^{-\frac{1}{2}}Q\|_2 = 0$, $P = Q$ over $\ker(\mathbf{L})^\perp$. Thus, $P = Q$.

For (ii), we see that $\mathcal{GFMMD}$ clearly takes the form of an MMD, so there is nothing to verify. And long as $\mathcal{GFMMD}(P,Q) < +\infty$, the distance is simply the Euclidean distance between $\varphi(P)$ and $\varphi(Q)$ using the feature map $\varphi : x \mapsto \mathbf{L}^{-\frac{1}{2}}x$. □

A. Relationship with f-Spectral Distances

The f spectral distance between two signals P and Q per [19] is defined as $\sum_i f(\lambda_i)(\hat{P}(i) - \hat{Q}(i))^2$, where f is some monotone increasing or decreasing function, and $\hat{P}, \hat{Q}$ denotes the Fourier transform of P and Q. Likewise, $\mathcal{GFMMD}(P,Q) = \sum_{\lambda_i \neq 0} \frac{1}{\lambda_i}(\hat{P}(i) - \hat{Q}(i))^2$. Algebraically, $\mathcal{GFMMD}$ resembles the f-spectral distance for $f(x) = 1/x$, although there is different treatment of components in $\ker(\mathbf{L})$, which arises from the fact that we only consider probability distributions.

B. Relationship with Resistive Embeddings

Graph Fourier MMD provides a natural extension of resistive embeddings through application of the same feature map. In fact, resistive embeddings can be viewed as a special case of Graph Fourier MMD for dirac distributions. On the other hand, theorem 2 offers a variational characteristic of resistive embeddings, which could be used for lower bounds. Finally, it can be shown that Graph Fourier MMD can be related to the the expected effective resistance of $X \sim P$ and $Y \sim Q$. Not only does this provide an interpretation of GFMMD, but it illustrates that spectral sparsification algorithms such as [16] preserve GFMMD.

Theorem 4. *If $X \sim P$ and $Y \sim Q$, not necessarily independent, then $\mathcal{GFMMD}(P,Q)^2 \leq \mathbb{E}_{X,Y}[Re(X,Y)]$.*

C. Graph Fourier MMD for Signal Localization

By fixing a probability distribution P and considering its distance to the uniform distribution $U = \frac{1}{n}\mathbf{1}$, we obtain a measure of how much P concentrates on the graph. For a high GFMMD to the uniform distribution, we say a signal is *localized*, and otherwise we say it is *dispersed*. Interestingly, localization simply corresponds to length in feature space.

Definition 5. *The* Localization Score *$s(P)$ of a signal P is defined as $\mathcal{GFMMD}(P,U) = \|\mathbf{L}^{-\frac{1}{2}}(P-U)\|_2 = \|\mathbf{L}^{-\frac{1}{2}}P\|_2$.*

D. Computational Complexity and Speedup

Computation of the pseudoinverse or $\mathbf{L}^{-\frac{1}{2}}$ is roughly $O(n^3)$. An alternate approach would be to simply calculate the solution to $\mathbf{L}y = P - Q$ via conjugate gradient descent. We elect to use Chebyshev polynomials to approximate $\mathbf{L}^{-\frac{1}{2}}P$ and $\mathbf{L}^{-\frac{1}{2}}Q$, then take Euclidean distances. This is a large improvement in the particular case where the graph $\mathcal{G}$ is sparse i.e. $|E| = O(n \log n)$. In such cases, we present an $O(n \log n)$ algorithm for the computation of GFMMD which is substantially faster than naive implementations based on a Chebyshev polynomial approximation of the filter in Algorithm 1 as well as a KNN kernel. The steps for an arbitrary graph are the same, but with $\mathbf{W}$ provided.

This gives a total fine grained time complexity of $\mathcal{O}((k_1 + t)n \log n + k_2 nm \log m)$ and a space complexity of $O(n \log n + mn)$ space. Here, t is the order of the Chebyshev polynomial, k_1 is the threshold for number of nearest-neighbors in constructing $\mathcal{G}$, and k_2 is the number of nearest-distributions we'd like to calculate. More simply, for fixed Chebyshev order, and number of neighbors, the time to estimate distances between all distributions is $\mathcal{O}(n \log n + nm^2)$.

Algorithm 1 GFMMD in Metric Space

Input: A set of n points $X \subseteq \mathbb{R}^d$, m probability distributions $f_i : X \to \mathbb{R}$ in an $n \times m$ matrix F, and a kernel function $k : X \times X \to \mathbb{R}$
Output: An $m \times n$ embedding matrix $\boldsymbol{E}$ in which $\|\boldsymbol{E}_i - \boldsymbol{E}_j\| = \mathcal{GFMMD}(f_i, f_j)$. M and a distance matrix in which $\boldsymbol{M}_{ij} = \mathcal{GFMMD}(f_i, f_j)$

Create a thresholded K-Nearest Neighbor graph $\mathcal{G}$ over X with $\mathcal{O}(n \log n)$ edges
$\boldsymbol{W}_{ij} \leftarrow k(X_i, X_j)$ for all $(i,j) \in \mathcal{E}$
$\mathbf{L} \leftarrow \mathbf{D} - \mathbf{W}$
$\boldsymbol{E}_i \leftarrow \mathbf{L}^{-\frac{1}{2}} f_i$ by exact or Chebyshev approximation of the filter $h(\lambda) = \lambda^{-\frac{1}{2}}$
$\boldsymbol{M}_{ij} \leftarrow \|\boldsymbol{E}_i - \boldsymbol{E}_j\|$ for all $i, j \in [n]$
return $\boldsymbol{M}, \boldsymbol{E}$

III. Experimental Results & Applications

In our experiments, signals are always nonnegative and normalized to be interpreted as probability distributions. However, the metric induced by $\mathbf{L}^{-\frac{1}{2}}$ *always* induces a valid seminorm on graph signals (but in particular, a norm on probability distributions).

A. Identifying Distributions on the Swiss Roll

In this experiment, we generate random point clouds centered at points on the swiss roll. More specifically, we sample $n = 100$ points on the swiss roll $x_1, x_2, \ldots, x_n$ and around each of these points, generate a point cloud d_i of size $m = 100$ points from a multivariate normal distribution centered at x_i. The result is nm points in $\mathbb{R}^{10}$. For each $i, j \in [n]$, we have a known geodesic distance between x_i and x_j. Across the different measures, we can see how well the distance between the point clouds d_i and d_j compares to the geodesic distance between their corresponding centers x_i and x_j. We compare the induced probability distributions on point clouds using: 1) computation of earth Mover's Distance between point clouds in ambient space, 2) Sinkhorn algorithm [4], 3) Diffusion EMD [18], 4) Kernel MMD [9] between all pairs p_i, p_j via random sampling (20 points from each distribution with replacement), 5) Graph Fourier MMD between p_i, p_j, using

both the exact calculation and approximation via Chebyshev polynomials. We then take correlation between the estimated nearest distributions and geodesic distance between centers. Results are shown in Table I. As we see, Graph MMD outperforms all other methods in accuracy and speed. In the appendix, the corresponding feature maps are visualized alongside geodesic distance, where GFMMD visually outperforms other methods at extracting manifold nonlinearities.

TABLE I: Comparison of runtime and Spearman-ρ correlation to ground truth manifold distances between distributions with mean $\pm$ standard deviation over 10 seeds for 100 distributions of 100 points each on a swiss roll manifold. The exact Graph MMD is most performant but requires a eigen-decomposition. The Chebyshev approximated Graph MMD (Chebyshev, t) is extremely fast and almost as performant at even low orders t.

Method	Spearman-ρ	10-NN time (s)	All-pairs time(s)
DiffusionEMD	0.584 $\pm$ 0.017	2.171 $\pm$ 0.265	3.341 $\pm$ 0.333
Exact	0.253 $\pm$ 0.022	26.881 $\pm$ 1.104	26.881 $\pm$ 1.104
Sinkhorn	0.250 $\pm$ 0.022	54.346 $\pm$ 17.576	54.346 $\pm$ 17.576
rbf-kernel-MMD	0.509 $\pm$ 0.021	5.016 $\pm$ 0.237	5.016 $\pm$ 0.237
Graph MMD (Exact)	**0.613** $\pm$ **0.019**	139.453 $\pm$ 16.790	139.468 $\pm$ 16.794
Graph MMD (Cheby, 8)	0.606 $\pm$ 0.024	**0.619** $\pm$ **0.057**	**0.641** $\pm$ **0.056**
Graph MMD (Cheby, 64)	0.593 $\pm$ 0.021	1.155 $\pm$ 0.035	1.163 $\pm$ 0.035
Graph MMD (Cheby, 512)	0.612 $\pm$ 0.018	6.249 $\pm$ 2.896	6.258 $\pm$ 2.895
Graph MMD (Cheby, 4096)	0.612 $\pm$ 0.018	48.138 $\pm$ 1.184	48.159 $\pm$ 1.182

B. Single cell Analysis with GFMMD

To demonstrate the utility of Graph Fourier MMD for biological analysis, we leverage publicly available single-cell RNA sequencing dataset of CD8-positive T cells [21]. CD8-positive T cells are adaptive immune cells known to be critical for mediating immune response in infection, cancer, and other diseases. We apply Algorithm 1 with the adaptive Gaussian Kernel [14] between datapoints, to compute GFMMD between genes, where each gene (of $1,991$ genes) is regarded as a distribution in a nearest neighbor cell graph over $9,167$ cells. In Figure 2A, we visualize the gene embedding using both PCA and PHATE [14]. We find that clusters $0-9$ in from the gene embedding show characteristic expression on the cellular embedding in Figure 2B. In other words, the subplots in Figure 2B represent a PHATE map of the cells in this dataset, and when we highlight the expression of *gene clusters* on the cells we see that these clusters have localized expression on the cellular manifold. To interpret these gene clusters for biological significance, we analyzed the gene set enrichment of clusters 6 and 7 with Enrichr [2], which show high expression in opposite ends of the cellular manifold (see Figure 2. Enrichr shows that cluster 7 has strong enrichment for signatures of a naive T cell becoming activatied with mitosis and T cell activation signatures being significant. On the other hand, cluster 6 shows strong enrichment for an effector CD8 T cell, with signatures of cytotoxic activity and inflammatory signaling (interferon gamma). Thus, these genes can be used to characterize the cellular manifold as following a trajectory from naive to effector CD8 T cells. We compare these to gene clusters derived from DiffusionEMD, as well as to a more standard method of gene selection in biology: differential

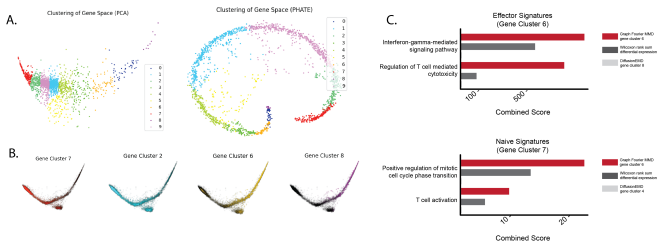

Fig. 2: A. Embeddings of genes of the dataset from [21] computed by the GFMMD Algorithm, visualized with PCA and PHATE [14], colored by results of K-means clustering. B. Embeddings of cells from [21] visualized with PHATE. Each plot is colored by the average expression of genes in the marked cluster over cells. C. Comparison of enrichment scores from Enrichr [2] on T-cell relevant annotations, between GFMMD-based gene sets from clusters 7, 6 and differential expression-based gene sets.

expression of genes in different areas of cellular state space based on a Wilcoxon rank sum test between the two manually curated cell clusters from [21]. The gene clusters 4 and 8 from DiffusionEMD that were most enriched on the opposite ends of the manifold consisted of 6 genes and 11 genes, which resulted in no enrichment for the above signatures. These genes upregulated based on the Wilcoxon rank sum test give a much less clear picture of the cellular state space, with the same annotations scoring much lower.

a) Local Genes: A novel type of analysis enabled by GFMMD is a search for *localized* signals. Often, researchers in the single cell field search highly variable genes, but we posit that genes that have localized expression on cellular manifolds can be used to characterize salient cellular subtypes. We propose the use of the localization score of the genes, viewed as probability distributions over the cells. Based on this score, in Figure 3, we visualize first most local gene, 10th most local gene, and 20th most local. Here, we compare localization scores between housekeeping genes and the gene signature for naive CD8+ T cells. Housekeeping genes are expressed highly in many systems, but are not known to have a function that contributes strongly to cell-cell variation for T cells [7], [20], [5]. By contrast, cells enriched for the naive CD8+ T cell signature are a subset of T cells along the T cell differentiation axis. We show that the localization score is an order of magnitude higher for the naive gene signature versus the housekeeping signature Figure 3 (see the appendix), validating our intuition about localized genes.

IV. Conclusion

In this paper we have introduced Graph Fourier MMD, a framework for taking distances between signals on graphs and generating embeddings in which these distances hold. We have shown its intuitive performance in both the Riemannian and abstract graphical setting for known distributions, as well as its advantage in speed, and ability to capture global properties of the underlying data manifold compared to alternative methods like Earth Mover's Distance and Diffusion EMD. Its rapidity makes it particularly useful for high dimensional datasets, such as single cell data, where we have showed its ability to capture the natural trajectories of gene expression.

REFERENCES

[1] Ashok K Chandra, Prabhakar Raghavan, Walter L Ruzzo, and Roman Smolensky. The electrical resistance of a graph captures its commute and cover times. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 574–586, 1989. pages

[2] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, April 2013. pages

[3] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006. pages

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. pages

[5] Hendrik J M de Jonge, Rudolf S N Fehrmann, Eveline S J M de Bont, Robert M W Hofstra, Frans Gerbens, Willem A Kamps, Elisabeth G E de Vries, Ate G J van der Zee, Gerard J te Meerman, and Arja ter Elst. Evidence based selection of housekeeping genes. *PLoS One*, 2(9):e898, September 2007. pages

[6] Richard M Dudley. *Real analysis and probability*. CRC Press, 2018. pages

[7] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes are compact. *Trends Genet.*, 19(7):362–365, July 2003. pages

[8] Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018. pages

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. pages

[10] Kenneth M Hall. An r-dimensional quadratic placement algorithm. *Management science*, 17(3):219–229, 1970. pages

[11] Tam Le, Truyen Nguyen, Dinh Phung, and Viet Anh Nguyen. Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pages 9844–9868. PMLR, 2022. pages

[12] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019. pages

[13] John C Mason and David C Handscomb. *Chebyshev polynomials*. Chapman and Hall/CRC, 2002. pages

[14] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, December 2019. pages

[15] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. pages

[16] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 563–568, 2008. pages

[17] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. pages

[18] Alexander Y Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pages 10336–10346. PMLR, 2021. pages

[19] Saurabh Verma and Zhi-Li Zhang. Hunt for the unique, stable, sparse and fast feature learning on graphs. *Advances in Neural Information Processing Systems*, 30, 2017. pages

[20] Xiliang Wang, Yao He, Qiming Zhang, Xianwen Ren, and Zemin Zhang. Direct comparative analyses of 10X genomics chromium and smart-seq2. *Genomics Proteomics Bioinformatics*, 19(2):253–266, April 2021. pages

[21] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8(1):14049, January 2017. pages

APPENDIX

A. Proof of Stated Results

We can also show that there is a nice correspondence for PCA on the space of dirac-distributions $\{\delta_i\}_i$ on the vertices, upon applying the feature map offered by GFMMD. In fact, the best k-dimensional representation (by multidimensional scaling) of the vertices will coincide almost exactly with Hall's Spectral Graph Drawing[10], which uses the first k nontrivial eigenvectors to represent vertices using coordinates in $\mathbb{R}^k$. This is made formal by Theorem 5.

Theorem 5. *If $X = \{\delta_i\}_{i\in\mathcal{V}}$ is a family of Kronecker-delta functions centered at each vertex of $\mathcal{G}$, then the k-dimensional embedding which best preserves the distances between signals in X is equivalent up to rescaling to Hall's Spectral Graph Drawing of the Graph $\mathcal{G}$ in k-dimensions.*

Proof. Note that X, the data matrix of Kronecker Deltas, is equal to $\mathbf{I}$, the n-dimensional identity. So $\sqrt{T}\mathbf{L}^{-\frac{1}{2}}X = \sqrt{T}\mathbf{L}^{-\frac{1}{2}}$, hence the best k-dimensional embedding of $\sqrt{T}\mathbf{L}^{-\frac{1}{2}}$ (respecting the L^2 norm between columns) will be equivalent to Principal Component Analysis (P.C.A.). Since $\mathbf{L}^{-\frac{1}{2}}\mathbf{1} = 0$, $\mathbf{L}^{-\frac{1}{2}}$'s columns are mean-centered, so its covariance matrix of $\sqrt{T}\mathbf{L}^{-\frac{1}{2}}$ is $\frac{T}{n}\mathbf{L}^{-\frac{1}{2}^T}\mathbf{L}^{-\frac{1}{2}} = \frac{T}{n}\mathbf{L}^{-}$.

Since its columns and rows are already mean centered. And thus P.C.A. will select the eigenvectors of $\mathbf{L}^-$ corresponding to the kth largest eigenvalues. Note that these are precisely given by $\psi_1, \psi_2..\psi_k$ with associated eigenvalues in $\mathbf{L}^-$ given by $\lambda_1^{-1}\dots\lambda_k^{-1}$. Letting $\Lambda_k = \text{diag}(\lambda_1^{-1/2}\dots\lambda_k^{-1/2})$ and $\Psi_k = \begin{pmatrix}\psi_1 & \dots \psi_k\end{pmatrix}$, P.C.A. would embed $\sqrt{T}\mathbf{L}^{-\frac{1}{2}}$ as,

$$\begin{aligned}\Psi_k^T\mathbf{L}^{-\frac{1}{2}} &= \Psi_k^T\Psi\Lambda^{-\frac{1}{2}}\Psi^T \\ &= \begin{pmatrix}\mathbf{I}_k & \mathbf{0}_{n-k}\end{pmatrix}\Lambda^{-\frac{1}{2}}\Psi^T \\ &= \begin{pmatrix}\mathbf{I}_k\Lambda_k & \mathbf{0}_{n-k}\end{pmatrix}\Psi^T = \Lambda_k\Psi_k^T.\end{aligned}$$

So our embedding of distributions would be given by $\Lambda_k\Psi_k^T$. On the other hand, Hall's Spectral Graph Drawing would embed the graph $\mathcal{G}$ simply as Ψ_k^T, since it chooses the first k nontrivial eigenvectors of $\mathbf{L}$. Thus, coordinates in each embedding are the same up to the rescaling by eigenvalues. □

1) Effective Resistances & Couplings: **Theorem 4** *If $X \sim P$ and $Y \sim Q$, not necessarily independent, then $\mathcal{GFMMD}(P,Q)^2 \leq \mathbb{E}_{X,Y}[Re(X,Y)]$*

Proof. The bias variance decomposition in dimension n states that for a random vector Z and point $a \in \mathbb{R}^n$, $\mathbb{E}\|Z - a\|^2 = \|\mathbb{E}Z - a\|^2 + Var(Z)$. Let $\varphi(a)$ denote column a of $\mathbf{L}^{-\frac{1}{2}}$, so that $\varphi(X), \varphi(Y)$ are random vectors. We have, $\mathcal{GFMMD}(P,Q)^2 = \|\mathbf{L}^{-\frac{1}{2}}P - \mathbf{L}^{-\frac{1}{2}}Q\|^2 = \|\mathbb{E}_X[\varphi(X)] - \mathbb{E}_Y[\varphi(Y)]\|^2 = \|\mathbb{E}_Y[\mathbb{E}_X[\varphi(X)]] - \mathbb{E}_Y[\mathbb{E}_X[\varphi(Y)]]\|^2$. By Fubini's Theorem for expectations, this is equal to, $\|\mathbb{E}_{X,Y}[\varphi(X)] - \mathbb{E}_{X,Y}[\varphi(Y)]\|^2 = \|\mathbb{E}_{X,Y}[\varphi(X) - \varphi(Y)]\|^2$. By the Bias-Variance Decomposition, $\|\mathbb{E}_{X,Y}[\varphi(X) - \varphi(Y)]\|^2 = \mathbb{E}_{X,Y}\|\varphi(X) - \varphi(Y)\|^2 - \text{Var}_{X,Y}[\varphi(X) - \varphi(Y)] \leq \mathbb{E}_{X,Y}\text{Re}(X,Y)$. We recognize that $\|\varphi(X) - \varphi(Y)\|^2 = \text{Re}(X,Y)$. □

Corollary 5.1. *Suppose P & Q agree on a set of size $\mathcal{A}$, and suppose the union of their supports is $\mathcal{S}$. Then, GFMMD$(P,Q) \leq \sqrt{(1-p)M} \leq \sqrt{(1-p)/2\lambda_2}$, where $p = \sum_{a\in\mathcal{A}} P(a)$, $M = \sup\{Re(X,Y) : X,Y \in S \setminus A\}$, and λ_2 is the Fiedler value for the graph.*

Proof. Let Z be a Bernoulli random variable with success probability p. First, choose $X_0 \sim P, Y_0 \sim Q$. Construct $X = X_0\mathbb{I}\{Z = 0\} + Z\mathbb{I}\{Z = 1\}$ and $Y = Y_0\mathbb{I}\{Z = 0\} + Z\mathbb{I}\{Z = 1\}$. Thus, $\mathcal{GFMMD}(P,Q)^2 \leq \mathbb{E}_{X,Y}\text{Re}(X,Y) \leq E_{X,Y}[\text{Re}(X,Y)|Z = 1]\mathbb{P}(Z = 1) + (1-p)\mathbb{E}_{X,Y}[\text{Re}(X,Y)|Z = 0]\mathbb{P}(Z = 0) \leq (1-p)\mathbb{E}_{X,Y}[\sup\{\text{Re}(X,Y) : X,Y \in S \setminus A] = (1-p)M$. Furthermore, we can provide an upper bound for M. The Courant-Fisher theorem tells us that for nonzero x, $x^T\mathbf{L}x \leq \frac{1}{\lambda_2}\|x\|^2$, as $1/\lambda_2$ is the maximal eigenvector of $\mathbf{L}^{-1}$. Thus, letting $a \neq b$ be arbitrary vertices, we have that Re(a,b) $= (\delta_a - \delta_b)^T\mathbf{L}^\dagger(\delta_a - \delta_b) \leq 2/\lambda_2$. In particular, maximizing over all $a, b \in \mathcal{S} \setminus \mathcal{A}$, $M \leq 2/\lambda_2$. □

B. Additional Figures for Experiments

1) Swiss Roll Experiment: The first of these figures is the first two principal components of the feature map $\mathbf{L}^{-\frac{1}{2}}$ applied to the distributions, which demonstrates the ability of GFMMD to capture nonlinear directions in a linear space in the presence of strong noise. On the left of Figure 4 is EMD, where the oscillatory pattern illustrates its ineffectiveness at calculating distances between distributions on graphs, since Euclidean distance between points on the swiss roll has periodic behavior in curvature. Diffusion EMD and Kernel MMD are effective at taking distances between points initially, but fail to discern between higher and higher distances. Graph Fourier MMD, on the other hand, has a far more clear linear correlation, which levels off much slower.

2) Single Cell Localization: Below, we have visualizations of the spread of the most localized signals over the graph. Here, PHATE is used to produce two dimensional embeddings of cells in Euclidean space, and color intensity is used as an indicator for gene expression.

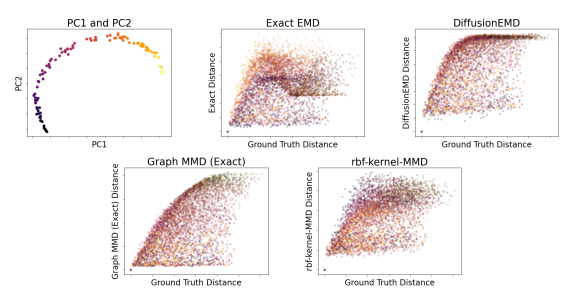

Fig. 4: Left: first two PCs of the embeddings E from Algorithm 1., colored by the coordinate of the corresponding center along the curved direction of the swiss roll. Right: Geodesic distance between centers vs. corresponding distance between distributions

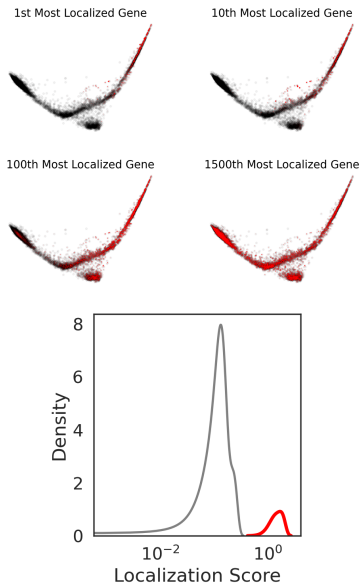

Fig. 3: We visualize the 1st, 10th, 100th, and 1500th most local genes on the cell graph. Indeed, we find the expected behavior. Density plots for localization scores, comparing housekeeping genes and naive CD8+ T cell signature. The naive gene signatures are given by the red curve and Housekeeping gene signatures by the gray.

C. Additional Toy Experiments

a) Grid Graph: First, we consider a 16×16 grid graph (vertices given by $\{(i,j)\}_{1 \leq i,j \leq 16}$. We can construct a signal P by placing a Dirac $\delta_{(8,4)}$ on the vertex (8,4) and then diffusing it with a heat filter (using time $\tau = 16$). Q is generated likewise, but by applying a heat filter to $\delta_{(8,4+2j)}$ and diffusing for each $j = 0, 1, 2, 3$. The result are two modes: P on the left, and Q moving along the right. The distributions are visualized in the top row, and the witness function to their difference in the bottom row of figure 5.

And of course, the corresponding distances between P and the Q's (per the order presented above) are increasing in the distances between the appropriate centers.

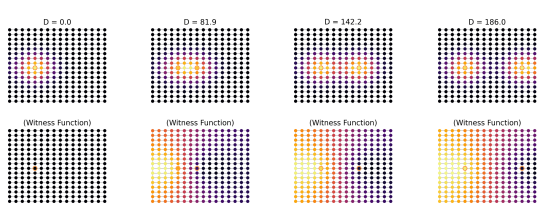

Fig. 5: Top row: the distributions P and Q, where the signal P stays fixed but the vertex at which Q is centered shifts to the right. Corresponding distances between distributions appear in the title, and the relevant centers of P and Q are highlighted. Bottom row: corresponding witness functions f to the difference between P and Q.

b) Bunny Graph: One very simple sanity check of a measure of spread is to verify that the more we diffuse a Dirac, the lower the distance to the uniform. Indeed, if we begin with the Bunny graph (from pygsp's built in library) and diffuse the Dirac δ_{1400} (1400 was chosen for visual appeal) for scales $\tau = 2^0, 2^4, 2^8$, and 2^12 (using a heat filter), we find that the corresponding measures of spread are 40.5, 26.9, 21.5, and 9.76. The signals are visualized below:

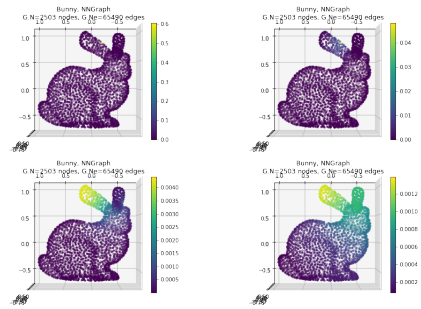

Fig. 6: The signal δ_{1400} diffused to levels 1,6^1, 6^2, and 6^3 using a heat filter.

D. Bimodal Signals

We can take the earlier signals from the grid graph (each pair of P and Q for translations of Q) and combine them into a new signal $\frac{1}{2}(P + Q)$. This forms a family of bimodal signals for which the two modes spread. Accordingly, in the example above, the distance to the uniform is given by 11.14, 8.66, 6.13, and 6.09.

E. Localization on the Minnesota Graph

1) Example: Minnesota Graph (Binarized): A final sanity check for a measure of closeness to the uniform would be to begin with a density which puts all its mass on one vertex. Then, put equal mass on that vertex and its neighbors, then the neighbors of neighbors, etc. More specifically, let $N_k(i,j) = \{\exists k' \in [k] : A^{k'} > 0\}$, or $N_k(i,j) = \mathbf{1}\{\text{there is a path of length} \leq k \text{ from } i \text{ to } j\}$. Then we can consider multiplying this by a Dirac, say δ_0 to get a family of signals. Using $k = 1, 4^1, 4^2, 4^3$, we have a family of distributions proportional to $N_1\delta_0, N_2\delta_0, N_3\delta_0$, and $N_4\delta_0$. Again, we can visualize the activated vertices in yellow:

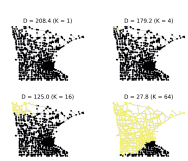

Fig. 7: The zeroth vertex's neighbors, then neighbors of neighbors, etc. for order $1, 4, 16$, and 64 neighbors. The corresponding distances to the uniform are given in the title.

F. Example: Minnesota Graph (Smooth Waves)

A similar example we can consider is a similar class of signals which "spread" across the graph, but rather than activating neighbors, simply diffusing the signal from a given start vertex. Here, we choose the same start vertex, and run heat diffusion at times $\tau = 2^0, 2^4, 2^8$, and 2^{12}.

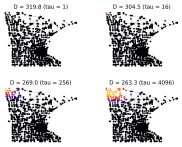

Fig. 8: Visualization of the diffusions of the signal δ_0 at times $2^0, 2^4, 2^8$, and 2^{12}. The corresponding distances to the uniform are given above.