APPENDIX

A. Proof of Stated Results

We first prove the main result for calculating Graph Fourier MMD:

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

$$\mathbb{E}_P(f_\alpha) - \mathbb{E}_Q(f_\alpha) = \sum_{v\in\mathcal{V}} P(v)f_\alpha(v) - \sum_{v\in\mathcal{V}} Q(v)f_\alpha(v)$$

$$= \sum_{v\in S} \alpha P(v) - \sum_{v\in S} \alpha Q(v) = \alpha c$$

And thus, since GFMMD is defined as $\sup_{f:f^T\mathbf{L}f\leq 1}$, we have GFMMMD$(P,Q) \geq \alpha c$. Taking $\alpha \to \infty$, we have GFMMMD$(P,Q) = +\infty$.

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,

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

$$= (P-Q)^T f_1 + (P-Q)^T f_2 = (P-Q)^T f_2$$

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,

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

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 $f \in \ker(\mathbf{L})^\perp$, $f = \mathbf{L}^{-1/2}y$ as well. Meaning, by a change of variables,

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

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

Lemma 4. *(i) GFMMD$(\cdot,\cdot)$ defines a valid distance on the probability distributions acting on $\mathcal{V}$. Furthermore, (ii) GFMMD$_\mathcal{G}(P,Q)$ is a Maximum Mean Discrepancy with explicit feature map $\mathbf{L}^{-\frac{1}{2}}$.*

Proof. For (i), note that $\mathbf{L}^{-1/2}(P-Q)$ is linear. By the usual nonnegativity of lengths, GFMMD$(P,Q) = \|\mathbf{L}^{-\frac{1}{2}}P - \mathbf{L}^{-\frac{1}{2}}Q\| \geq 0$, so GFMMD$(\cdot,\cdot)$ is nonnegative. Furthermore, note that GFMMD$(P,Q) = 0$ if and only if $\mathbf{L}^{-\frac{1}{2}}P = \mathbf{L}^{-\frac{1}{2}}Q$. But since P and Q sum to 1, $\mathbf{L}^{-\frac{1}{2}}$ as injectively on the set of functions orthogonal to its kernel, so $P = Q$. Thus, GFMMD$(P,Q) \geq 0$, with equality if and only if $P = Q$. Finally, the triangle inequality holds, since for arbitrary probability densities P, Q, R, GFMMD$(P,Q) = \|\mathbf{L}^{-\frac{1}{2}}P - \mathbf{L}^{-\frac{1}{2}}Q\| \leq \|\mathbf{L}^{-\frac{1}{2}}P - \mathbf{L}^{-\frac{1}{2}}R\| + \|\mathbf{L}^{-\frac{1}{2}}Q - \mathbf{L}^{-\frac{1}{2}}R\| = \text{GFMMD}(P,R) + \text{GFMMD}(Q,R)$ follows from the usual triangle inequality in ℓ_2. Thus, GFMMD is a valid distance acting on probability distributions. For (ii), by definition, an MMD γ between P and Q takes the form $\gamma(P,Q) = \sup_{\|f\|_\mathcal{H}\leq 1} \mathbb{E}_P(f) - \mathbb{E}_Q(f)$, where $\mathcal{H}$ is some Hilbert Space and $\{f : \|f\|_\mathcal{H} \leq 1\}$ corresponds to the unit ball. If we define a Hilbert space on ℓ_2 with $\langle x,y\rangle_\mathcal{H} = x\mathbf{L}^- y$, it follows that $\|f\|_\mathcal{H} \leq 1$ corresponds to $f^T\mathbf{L}f \leq T$. Thus, GFMMD$(\cdot,\cdot)$ possesses the form of a valid MMD. Therefore, this distance is also an IPM. □

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,

$$\Psi_k^T\mathbf{L}^{-\frac{1}{2}} = \Psi_k^T\Psi\Lambda^{-\frac{1}{2}}\Psi^T$$

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

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 3** *If $X \sim P$ and $Y \sim Q$, not necessarily independent, then 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 $\mathbf{L}^{-1/2}$, so that $\varphi(X), \varphi(Y)$ are random vectors. We have, GFMMD$(P,Q)^2 = \|\mathbf{L}^{-1/2}P - \mathbf{L}^{-1/2}Q\|^2 = \|\mathbb{E}_X[\varphi(X)] - \mathbb{E}_Y[\varphi(Y)]\|^2 = \|\mathbb{E}_Y[\mathbb{E}_X[\varphi(X)]] - \mathbb{E}_Y[\mathbb{E}_X[\varphi(Y)]]\|^2$. 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, 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) $= (\delta_a - \delta_b)^T\mathbf{L}^-(\delta_a - \delta_b) \leq 2/\lambda_2$. 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 $\mathbf{L}^{-1/2}$ 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 3 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

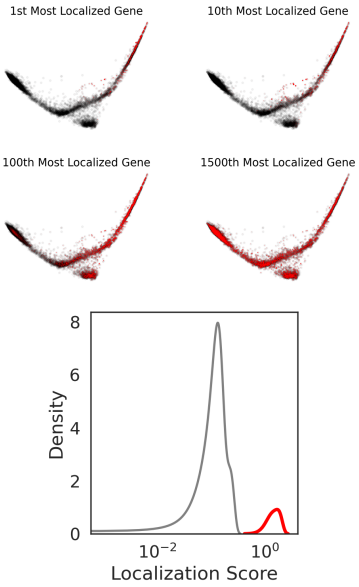

Fig. 2: We visualize the 1st, 10th, 100th, and 1500th most local genes on the cell graph. Indeed, we find the expected behavior. Density plots for localization scores, comparing housekeeping genes and naive CD8+ T cell signature. The naive gene signatures are given by the red curve and Housekeeping gene signatures by the gray.

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.

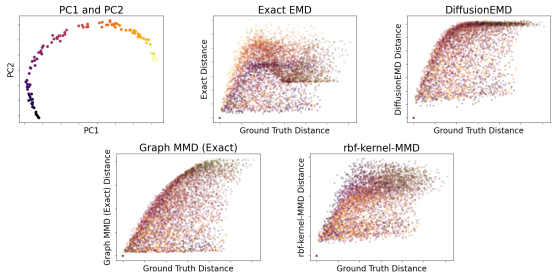

Fig. 3: Left: first two PCs of the embeddings E from Algorithm 1, colored by the coordinate of the corresponding center along the curved direction of the swiss roll. Right: Geodesic distance between centers vs. corresponding distance between distributions

C. Grid Graph

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:

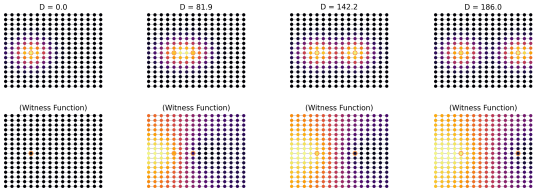

Fig. 4: Top row: the distributions P and Q, where the signal P stays fixed but the vertex at which Q is centered shifts to the right. Corresponding distances between distributions appear in the title, and the relevant centers of P and Q are highlighted. Bottom row: corresponding witness functions f to the difference between P and Q.

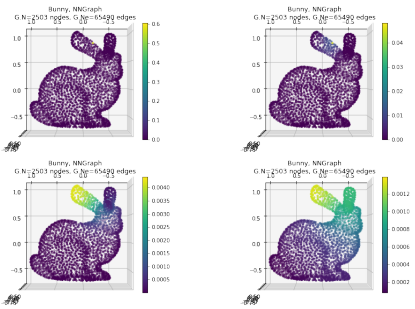

Fig. 5: The signal δ_{1400} diffused to levels 1,6^1, 6^2, and 6^3 using a heat filter.

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 4.

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.

D. Diffusing Signals on the 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:

E. 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.

F. 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:

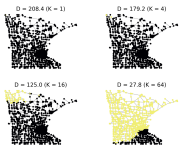

Fig. 6: The zeroth vertex's neighbors, then neighbors of neighbors, etc. for order $1, 4, 16$, and 64 neighbors. The corresponding distances to the uniform are given in the title.

G. 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}.

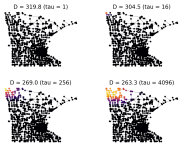

Fig. 7: Visualization of the diffusions of the signal δ_0 at times $2^0, 2^4, 2^8$, and 2^{12}. The corresponding distances to the uniform are given above.