

---

# GENERALIZED SPECTRAL CLUSTERING VIA GROMOV-WASSERSTEIN LEARNING

---

**Samir Chowdhury**  
Stanford University

**Tom Needham**  
Florida State University

March 4, 2021

## ABSTRACT

We establish a bridge between spectral clustering and Gromov-Wasserstein Learning (GWL), a recent optimal transport-based approach to graph partitioning. This connection both explains and improves upon the state-of-the-art performance of GWL. The Gromov-Wasserstein framework provides probabilistic correspondences between nodes of source and target graphs via a quadratic programming relaxation of the node matching problem. Our results utilize and connect the observations that the GW geometric structure remains valid for any rank-2 tensor, in particular the adjacency, distance, and various kernel matrices on graphs, and that the heat kernel outperforms the adjacency matrix in producing stable and informative node correspondences. Using the heat kernel in the GWL framework provides new multiscale graph comparisons without compromising theoretical guarantees, while immediately yielding improved empirical results. A key insight of the GWL framework toward graph partitioning was to compute GW correspondences from a source graph to a template graph with isolated, self-connected nodes. We show that when comparing against a two-node template graph using the heat kernel at the infinite time limit, the resulting partition agrees with the partition produced by the Fiedler vector. This in turn yields a new insight into the  $k$ -cut graph partitioning problem through the lens of optimal transport. Our experiments on a range of real-world networks achieve comparable results to, and in many cases outperform, the state-of-the-art achieved by GWL.

## 1 INTRODUCTION

The Gromov-Wasserstein (GW) problem is a nonconvex quadratic program whose solution yields the *GW distance*, a variant of Wasserstein distance from classical optimal transport (OT) which is able to compare distributions defined on different metric spaces. This is accomplished by replacing classical Wasserstein loss with a loss function defined in terms of relational information coming from metric data. Because it is able to compare distributions defined on a priori incomparable spaces, GW distance is increasingly finding applications for learning problems on irregular domains such as graphs (Hendrikson, 2016; Tsitsulin et al., 2018; Vayer et al., 2019a; Xu et al., 2019b; Xu, 2020). In this context, a graph can be considered as a metric space by endowing it with geodesic distance. A soft matching between nodes of two different graphs is obtained by choosing distributions on each graph's nodes (e.g., uniform distributions) and computing the GW optimal transport plan between them.

Applications of GW distance have been bolstered by the observation that the GW problem does not fundamentally require a *metric* to operate (Peyré et al., 2016); i.e., the definition of the GW loss function extends to other forms of relational data (Chowdhury and Mémoli, 2019). Xu et al. (2019a) used this observation to produce the state-of-the-art Scalable Gromov-Wasserstein Learning framework for graph matching and partitioning, which fundamentally uses the the adjacency matrix (as opposed to the shortest path distance matrix) of a graph. There are many ways to derive relational data from a graph beyond its distance and adjacency matrices, such as its various graph Laplacians and their corresponding heat kernels. A limitation of the current literature is a lack of tools for ascertaining if the adjacency matrix (or any other rank-2 tensor derived from a graph) is optimal in some sense beyond empirical benchmarks. Thus the potential flexibility of the GW framework for graph analysis remains largely unexplored.

Here we study the GW graph OT problem by representing graphs via heat kernels rather than adjacency matrices. This amounts to finding soft correspondences between the nodes of two graphs by comparing spectral, rather than adjacency, data. We refer to this as the SpecGWL framework, reserving GWL for the adjacency-based framework of Xu et al. (2019a).

Numerical experiments<sup>1</sup> demonstrate that SpecGWL outperforms GWL in graph partitioning tasks. Moreover, a main goal of this paper is to introduce tools for studying the GW problem in a more rigorous manner. We use a Markov Chain sampling technique to explore the energy landscape of the GW loss function for adjacency and heat kernel graph representations, showing empirically that SpecGWL loss has fewer spurious local minima and a 10x acceleration in convergence of gradient descent over GWL loss. We also introduce a visualization technique which allows one to ascertain the quality of soft node matchings obtained by any GW method. Examples of this technique intuitively demonstrate the idea that heat kernel-based matchings more faithfully preserve global graph structure than adjacency-based matchings. Finally, we establish theoretical results on the sparsity of optimal couplings in the SpecGWL framework and on the precise relation of SpecGWL graph partitioning to classical spectral clustering. The latter result creates a novel connection between spectral clustering and optimal transport.

**Related literature.** Gromov-Wasserstein distance—as used in current ML settings—was introduced by Mémoli (2007) as a convex relaxation scheme for the Gromov-Hausdorff distance between different metric spaces. Further theoretical development was carried out by Mémoli (2011a); Sturm (2012), and a related formulation was introduced by Sturm (2006) to study the convergence of sequences of metric measure spaces. The idea of using heat kernels for GW matching goes back to the *Spectral Gromov-Wasserstein distance* introduced by Mémoli (2011b) in the context of Riemannian manifolds to further develop the notion of *Shape-DNA* introduced by Reuter et al. (2006). A related theoretical construction also appeared in Kasue and Kumura (1994). The Riemannian heat kernel has been celebrated for its *multiscale* and *informative* properties (Sun et al., 2009)—the former refers to the observation that the heat kernel defines a family of Gaussian filters that get progressively shorter and wider as  $t \rightarrow \infty$ , and the latter refers to the classical lemma of Varadhan showing that at the small time limit, the log-heat kernel approximates the geodesic distance on a Riemannian manifold. A surprising result due to Sun et al. (2009) in this direction is that under mild conditions, the collection of traces of the heat kernel forms an isometry invariant of a manifold despite giving up most of the information contained in the heat kernel. Heat kernel traces have recently been used for graph comparison by Tsitsulin et al. (2018, 2020), who showed that the desirable properties of the heat kernel for Riemannian manifolds have natural and informative analogues in the setting of graphs. Heat kernels have been incorporated into GW-based graph matching in recent work of Barbe et al. (2020), where they are used to augment the matching process in the Fused GW framework of Vayer et al. (2018). Our work is distinguished from that of Barbe et al. (2020) in that we use heat kernels directly to define a GW loss function for graph matching, whereas Barbe et al. (2020) employs heat kernels to improve feature space matchings for attributed graphs. There is a deep literature on spectral graph comparison, and a few other references include works of Patro and Kingsford (2012); Bronstein and Glashoff (2013); Hu et al. (2014); Nassar et al. (2018); Dong and Sawin (2020).

Starting from its early applications in computer vision (Mémoli, 2007, 2011a; Schmitzer and Schnörr, 2013), GW distance has been applied to alignment of word embedding spaces (Alvarez-Melis and Jaakkola, 2018), learning of generative models across different domains (Bunne et al., 2019), graph factorization (Xu et al., 2019b), and learning autoencoders (Xu et al., 2020). Related theoretical directions include the *sliced GW* of Vayer et al. (2019b) and the *Gromov-Monge* problem studied by Mémoli and Needham (2018). Sturm (2012) studied the structure of geodesics and gradient flows in GW geometry, and recently these techniques were utilized by Chowdhury and Needham (2020) to create a Riemannian framework for performing averaging and tangent PCA across different graph-structured domains. Unbalanced formulations of the GW problem permitting input probability measures of different total sums have been studied by Chapel et al. (2020); De Ponti and Mondino (2020); Séjourné et al. (2020).

## 2 SPECTRAL GW DISTANCES

**Gromov-Wasserstein Distance.** The following can be stated for Borel probability measures on Polish spaces; however, we are interested in the finite setting where measure-theoretic complications do not arise, so we freely use matrix-vector notation. Given probability distributions  $p, q$  on finite sets  $X, Y$ , a coupling of  $p$  and  $q$  is a joint probability measure  $C$  on  $X \times Y$  with marginals  $p$  and  $q$ ; i.e.,  $C = (C_{ij}) \in \mathbb{R}^{|X| \times |Y|}$  satisfies equality constraints  $C\mathbf{1}^{|Y| \times 1} = p$  and  $C^T\mathbf{1}^{|X| \times 1} = q$  ( $1^{m \times n}$  denoting the  $m \times n$  matrix of all ones), and entrywise inequality constraints  $0 \leq C \leq 1$ . The set of all couplings of  $p$  and  $q$  is denoted  $\mathcal{C}(p, q)$ —this is a convex polytope in  $\mathbb{R}^{|X| \times |Y|}$ . Given functions  $f^X : X \times X \rightarrow \mathbb{R}$ ,  $f^Y : Y \times Y \rightarrow \mathbb{R}$ , written as square matrices  $F^X, F^Y$ , the *GW problem* solves:

$$\min_{C \in \mathcal{C}(p, q)} \sum_{i,k} \sum_{j,l} (F_{ik}^X - F_{jl}^Y)^2 C_{ij} C_{kl}. \quad (1)$$

The GW problem is a nonconvex quadratic program over a convex domain for which approximate solutions may be obtained via projected gradient descent (Peyré et al., 2016) or Sinkhorn iterations with an entropy (Solomon et al.,

<sup>1</sup><https://github.com/trneedham/Spectral-Gromov-Wasserstein>

2016) or KL divergence regularizer (Xu et al., 2019b). Computational implementations can be found in the Python Optimal Transport library of Flamary and Courty (2017).

Historically, the GW problem was defined by Mémoli (2007) in the setting of *metric measure (mm) spaces*, where its solution leads to a metric known as the Gromov-Wasserstein distance. A finite mm space  $(X, d, p)$  consists of a finite set of points  $X$ , a metric function  $d$  written as a matrix  $d = (d_{ik}) \in \mathbb{R}^{|X| \times |X|}$ , and a probability distribution  $p$  on  $X$  written as a vector  $p = (p_i) \in [0, 1]^{|X| \times 1}$ . Given mm spaces  $(X, d^X, p), (Y, d^Y, q)$ , the Gromov-Wasserstein distance is defined as:

$$d_{\text{GW}}(X, Y)^2 := \min_{C \in \mathcal{C}(p, q)} \sum_{i, k} \sum_{j, l} (d_{ik}^X - d_{jl}^Y)^2 C_{ij} C_{kl}. \quad (2)$$

Peyré et al. (2016) observed that solving the GW problem (1) with input matrices that are not strictly distances—in particular kernel matrices—still leads to a discrepancy that is informative when comparing matrices of different sizes and arising in incompatible domains. This idea was pushed further by the theoretical work of Chowdhury and Mémoli (2019), which showed that *any* square matrix representation of a graph, including the adjacency matrix, can be used in the GW problem (1) to obtain a bona fide distance (actually a pseudometric). This observation was used by Xu et al. (2019b,a) to create a unified *Gromov-Wasserstein Learning* (GWL) framework for unsupervised learning tasks on graphs—e.g. finding node correspondences between unlabeled graphs and graph partitioning—with state-of-the-art performance. We now formulate the GWL framework precisely.

**Gromov-Wasserstein Learning on Graphs.** Let  $G = (V, E)$  be a finite, unweighted, possibly directed graph. We refer to  $V$  as the set of nodes and  $E$  as the set of edges. Let  $A : V \times V \rightarrow \{0, 1\}$  and  $D : V \rightarrow \mathbb{Z}$  denote the *adjacency* and *degree* functions defined as  $A(v, w) := 1$  if  $(v, w) \in E$ , 0 otherwise, and  $D(v) := |\{w : (v, w) \in E\}|$ . Given an ordering on  $V$ , these functions can be represented as matrices in  $\mathbb{R}^{|V| \times |V|}$ , and we will switch between both the function and matrix forms. In addition to the adjacency function, which fully encodes a graph, there are numerous *derived representations* which capture information about a graph. An example is the geodesic distance function  $d$  which contains all the shortest path lengths in the graph. In particular, the graph geodesic distance representation was used by Hendrikson (2016) to solve graph matching problems via the original mm-space GW distance formulation (2) of Mémoli (2007).

One of the key ideas in the GWL framework is to use *adjacency matrices* in the GW problem. Specifically, let  $G$  and  $H$  be graphs with distributions  $p$  and  $q$  on their nodes and let  $A^G$  and  $A^H$  denote their adjacency matrices. The GWL framework considers the loss

$$\text{Adj} = \text{Adj}_{G, p, H, q} : \mathcal{C}(p, q) \rightarrow \mathbb{R}$$

defined by

$$\text{Adj}(C) = \sum_{i, k} \sum_{j, l} (A_{ik}^G - A_{jl}^H)^2 C_{ij} C_{kl}, \quad (3)$$

which we refer to as *adjacency loss*. The minimum of  $\text{Adj}(C)^{1/2}$  defines a distance between the pairs  $(G, p)$  and  $(H, q)$ —we refer to such pairs as *measure graphs* and assume for convenience that distributions are fully supported. If  $p$  and  $q$  are themselves derived from adjacency data then the minimizer  $C$  of (3) provides a natural soft correspondence between the nodes of  $G$  and  $H$ . For example, Xu et al. (2019a) consider the family of node distributions  $p = (p_1, \dots, p_n)^T$ , where

$$p_j = \frac{\bar{p}_j}{\sum_{k=1}^n \bar{p}_k}, \quad \bar{p}_j = (\deg(v_j) + a)^b, \quad (4)$$

where  $p_j = p(v_j)$  for  $v_j$  a vertex of  $G$ ,  $a \geq 0$  is used to enforce the full support condition and the exponent  $b \in [0, 1]$  allows interpolation between the *uniform distribution* and the *degree distribution*.

**Heat Kernels.** Our contributions begin by studying the structure imposed on the GW problem (1) by *spectral losses* that we describe next. Given an undirected graph  $G = (V, E)$ , let  $L^2(V)$  denote the linear space of functions  $f : V \rightarrow \mathbb{R}$ . The *Laplacian* of  $G$  is the operator  $L : L^2(V) \rightarrow L^2(V)$  defined by

$$L(\phi)(v) := D(v)\phi(v) - \sum_{(v, w) \in E} \phi(w).$$

After fixing an ordering on  $V$ , we can use matrix-vector notation to write  $\mathbb{R}^{|V| \times |V|} \ni L = D - A$  and  $\phi \in \mathbb{R}^{|V| \times 1}$ . The Laplacian is symmetric positive semidefinite, so the spectral theorem guarantees an eigendecomposition  $L = \Phi \Lambda \Phi^T$  with real, nonnegative eigenvalues  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$  arranged on the diagonal of  $\Lambda$ . The corresponding eigenvectors  $\phi_1, \phi_2, \dots, \phi_n$  are arranged as the columns of  $\Phi$ . There are several variants of the Laplacian with analogous properties, including the *normalized Laplacian* given by  $I - D^{-1/2} A D^{-1/2}$ . This is the version that we will typically use for experiments.

For a (strongly connected) directed graph  $G = (V, E)$ , we use the normalized Laplacian defined by Chung (2005) via the language of random walks. Consider the transition probability matrix  $P$  defined by writing  $P_{ij} = 1/D_i$  if  $(i, j) \in E$ , 0 otherwise. By Perron-Frobenius theory, there is a unique left eigenvector  $\psi$  with all entries positive such that  $\psi^T P = \psi^T$ . The directed graph Laplacian is defined as  $L := I - (\Psi^{1/2} P \Psi^{-1/2} + \Psi^{-1/2} P^T \Psi^{1/2})/2$ , where  $\Psi = \text{diag}(\psi)$ .

In what follows, we use  $L$  generically to refer to any of the Laplacians defined above. The *heat equation* on a graph  $G$  (either undirected or directed) is then given as  $du/dt = -Lu$ , where  $u \in L^2(V \times \mathbb{R}_{>0})$ . If  $u$  represents a time-dependent heat distribution on the nodes of  $G$ , the heat equation describes heat diffusion according to Newton's Law. The *heat kernel* is the fundamental solution to this heat equation, given in closed form as  $K^t = \exp(-tL) = \Phi \exp(-t\Lambda) \Phi^T$ .

**Spectral GW Distance.** Given measure graphs  $(G, p)$  and  $(H, q)$ , we consider the *spectral loss*

$$\text{Spec}^t = \text{Spec}_{G,p,H,q}^t : \mathcal{C}(p, q) \rightarrow \mathbb{R}$$

defined by

$$\text{Spec}^t(C) = \sum_{i,k} \sum_{j,l} (K_{ik}^{G,t} - K_{jl}^{H,t})^2 C_{ij} C_{kl} \quad (5)$$

for each  $t > 0$ , where  $K^{G,t}$  and  $K^{H,t}$  are the heat kernels of  $G$  and  $H$ , written in matrix form. We then obtain a one parameter family of pseudometrics

$$d_{\text{GW}}^{\text{spec}}[t]((G, p), (H, q)) := \min_{C \in \mathcal{C}(p, q)} \text{Spec}^t(C)^{1/2} \quad (6)$$

on the space of measure graphs. As in the GWL framework, choosing node distributions  $p$  and  $q$  from the family (4) yields minimizing couplings  $C$  which give meaningful correspondences between the nodes of  $G$  and  $H$ . As  $t$  varies, one obtains multiscale couplings between nodes, with small  $t$  encoding local and large  $t$  encoding global structure (this is made precise in Section 3). Intuitively, the heat kernel replaces each node of a graph with a Gaussian filter with width controlled by  $t$ , progressively “smoothing” the graph. Different smoothing levels emphasize multiscale geometric and topological features that drive coupling optimization; Figure 6 in the Supplementary Materials shows clusters in coupling space corresponding to matchings at different scales. The “best” choice of  $t$  emphasizes the feature scale which is most important for a given task.

One can further define an overall pseudometric by, say, considering the function

$$f_{G,p,H,q}(t) = d_{\text{GW}}^{\text{spec}}[t]((G, p), (H, q))$$

and defining  $d_{\text{GW}}^{\text{spec}} := \|\xi \cdot f_{G,p,H,q}\|$  for an appropriately chosen norm on functions  $\mathbb{R} \rightarrow \mathbb{R}$  and normalizing function  $\xi$  preventing blow up of the norm; for example, taking  $\xi(t) = \exp(-(t + t^{-1}))$  and the  $\ell_\infty$  norm here is analogous to the spectral GW distance between Riemannian manifolds defined by Mémoli (2011b). This is an interesting direction of future research, but we found it most useful in our applications to instead treat  $t$  as a scale parameter that is tuned via cross-validation.

**Properties of SpecGWL** For measure graphs  $(G, p)$  and  $(H, q)$ , fix  $t > 0$  and write  $J := K^{G,t}, K := K^{H,t}$ . After expanding the square and invoking the marginalization constraints (Solomon et al. (2016) provide an explicit derivation), one sees that minimizing (5) is equivalent to maximizing  $\langle JC, CK \rangle$  subject to  $C \in \mathcal{C}(p, q)$ , where  $\langle \cdot, \cdot \rangle$  denotes the Frobenius inner product. Because the graph heat kernel is symmetric positive definite, we take Cholesky decompositions  $J = U^T U, K = V^T V$  to write

$$\begin{aligned} \langle JC, CK \rangle &= \text{tr}((JC)^T CK) = \text{tr}(C^T U^T U C V^T V) \\ &= \text{tr}(V C^T U^T U C V^T) = \|UCV^T\|^2. \end{aligned}$$

The map  $C \mapsto \|UCV^T\|^2$  is convex. We record this as the following lemma, which appeared previously as (Alvarez-Melis et al., 2019, Lemma 4.3).

**Lemma 1.** *For each  $t > 0$ , spectral loss (5) is minimized over the convex polytope  $\mathcal{C}(p, q)$  by a maximizer of the convex function  $C \mapsto \langle K^{G,t} C, CK^{H,t} \rangle$ .*

Optimization problems of this type are not tractable to solve deterministically, but we demonstrate experimentally that approximation via gradient descent enjoys faster convergence and fewer spurious local minima than adjacency loss (3). It has been empirically observed by Xu et al. (2019b) that (local) minimizers of adjacency loss tend to become sparse, and this sparsity is crucial in the gradient descent-based algorithm of Chowdhury and Needham (2020) for computing averages of networks. A key advantage of our spectral setting (5) over (2) or (3) is that this empirical observation of sparsity admits a formal proof.

**Theorem 2.** *Let  $(G, p)$  and  $(H, q)$  be measure graphs on  $\sim n$  nodes. Then for any  $t > 0$ , there is a minimizer of spectral loss with  $o(n)$  nonzero entries.*

**Complexity of SpecGWL** Assuming graphs of comparable size  $n$ , the eigendecomposition incurs a time complexity of  $O(n^3)$  and memory complexity of  $\Theta(n^2)$ . Computing the GW loss using gradient descent involves computing  $\nabla \langle JC, CK \rangle = J^T CK + JCK^T$ , which incurs a time complexity of  $O(n^3 \log(n))$  (Kolouri et al., 2017) and memory complexity of  $O(n^2)$ . Regularized methods with Sinkhorn iterations still require paying a cost for matrix multiplication (Peyré et al., 2016). However, accelerations have already been proposed: Tsitsulin et al. (2018) suggest methods for approximating heat kernels in  $O(n^2)$  operations and Xu et al. (2019a) propose a recursive divide-and-conquer approach to reduce the complexity of the GW comparison to  $O(n^2 \log n)$ . Note that because heat kernel matrices are dense, we lose the advantages of sparse matrix operations.

### 3 GRAPH PARTITIONING

**Graph Partitioning Method.** Graph partitioning is a crucial unsupervised learning task used for community detection in social and biological networks (Girvan and Newman, 2002). The goal is to partition the vertices of a graph into some number of clusters  $m$  in accordance with the *maximum modularity principle*—edges within clusters are dense, while edges between clusters are sparse. Xu et al. (2019a) proposed a GW-based approach to graph partitioning where an  $m$ -way partition of a measured graph  $(G, p)$  is obtained by minimizing the following variant of adjacency loss (3):

$$\mathcal{C}(p, q) \ni C \mapsto \sum_{i,k} \sum_{j,l} (A_{ik} - Q_{jl})^2 C_{ij} C_{kl}, \quad (7)$$

where  $A$  is the adjacency matrix of  $G$ ,  $Q = \text{diag}(q)$  and  $q$  is a distribution estimated by sorting the weights of  $p$ , sampling  $m$  values via linear interpolation and renormalizing. Intuitively,  $Q$  is the weighted adjacency matrix of a graph on  $m$  nodes with only self-loops—an *ideally clustered template* graph. The heuristic for choosing the distribution  $q$  in this manner is that if within-cluster nodes of the graph  $G$  have similar degrees then this method allows a node in  $Q$  to accept all of the mass from this cluster. A minimizer  $C$  of (7) defines an  $m$ -way partition of  $G$ : each node  $v_i$  of  $G$  is assigned a label in  $\{1, \dots, m\}$  according to the column index of the maximum weight in row  $i$  of  $C$ .

Soft-matching nodes of the target graph to an ideally clustered template is intuitively appealing and it was shown by Xu et al. (2019b) that this method achieves state-of-the-art performance. We propose a variant of the algorithm: letting  $K^t$  denote the heat kernel for  $G$  at some  $t > 0$ , we minimize

$$\mathcal{C}(p, q) \ni C \mapsto \sum_{i,k} \sum_{j,l} (K_{ik}^t - Q_{jl})^2 C_{ij} C_{kl}, \quad (8)$$

with  $Q$  defined as above. For each  $t > 0$ , we obtain an optimal coupling which is used to partition  $G$  as described above. Experimental results in Section 4 show that this change to the algorithm gives a significant performance boost over the adjacency-based version.

**Connection to Spectral Clustering.** Let  $G$  be an undirected, connected graph with graph Laplacian  $L$ . The connectivity of  $G$  implies that  $L$  has exactly one zero eigenvalue with constant eigenvector. Assume for simplicity that the multiplicity of the smallest positive eigenvalue of  $L$  is one. A fundamental concept in spectral graph theory is that the corresponding eigenvector—the *Fiedler vector* of  $G$ —gives a 2-way partitioning of  $G$  with good theoretical properties: nodes of  $G$  are partitioned according to the sign of their entry in the Fiedler vector (Fiedler, 1973). We refer to this as the *Fiedler partition* of  $G$ , and find that it arises as a special case of spectral GW partitioning:

**Theorem 3.** *Let  $G$  be a connected graph whose first positive eigenvalue has multiplicity one, endowed with the uniform node probability distribution  $p$ . For sufficiently large  $t$ , the 2-way partition of  $G$  derived from a minimizer of (8) agrees with the Fiedler partitioning.*

The theorem demonstrates a novel connection between optimal transport and classical spectral clustering and shows that partitioning graphs via spectral GW matching (8) is a generalization of these classical methods. For the small- $t$  regime, note that the heat kernel of  $G$  has Taylor expansion  $K^t = I_n + tL + O(t^2)$ , where  $I_n$  is the  $n \times n$  identity matrix. Thus for low values of  $t$ , spectral GW partitioning is driven by matchings of graph Laplacians, which contain local adjacency information.

### 4 EXPERIMENTS

We present several numerical experiments demonstrating the boost in performance obtained by using heat kernels in the GW problem rather than adjacency matrices. In experiments with undirected graphs, we used the normalized graph Laplacian to construct heat kernels—results were qualitatively similar using heat kernels of the standard Laplacian, but we found some boost in quantitative performance in graph partitioning when using the normalized version. Experiments with directed graphs use Chung’s normalized Laplacian.

Table 1: Results of Energy Landscape Experiment.

Loss	Time (s)	Err. (%)	Prod. Err. (%)
Adj	.0230	24.21	4.81
Spec <sup>5</sup>	.0014	22.35	6.00
Spec <sup>10</sup>	.0013	3.86	1.34
Spec <sup>20</sup>	.0010	0.05	.02

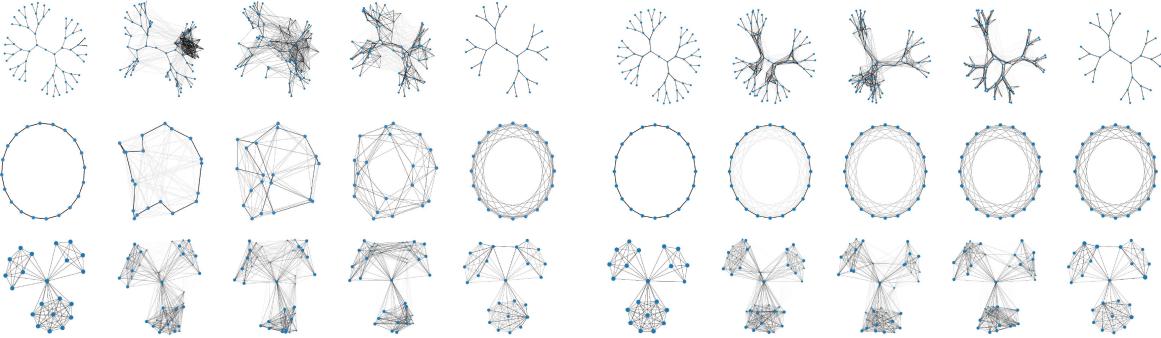


Figure 1: Visualizations of GW graph matchings via adjacency loss (3) (**left column**) and spectral loss (5) with  $t = 20$  (**right column**). For each loss, we illustrate matchings between binary trees (**Top**), circular graphs (**Middle**) and IMDB graphs (**Bottom**). In each case, the interpolation is generated from node matchings inferred from the optimal GW coupling. Uniform node distributions are used in all examples.

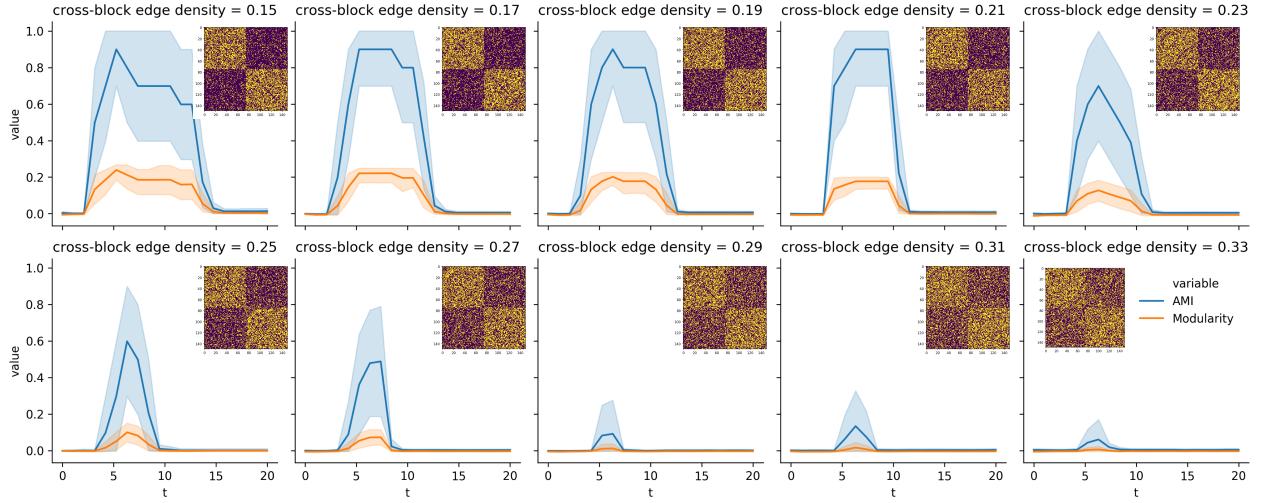
**Energy Landscapes and Convergence Rates.** Adjacency loss (3) is highly nonconvex, while Lemma 1 shows that minimizing spectral loss (5) is equivalent to maximizing a convex function over a convex polytope. While the latter optimization problem is still intractable, we now demonstrate that its approximation via gradient descent is well-behaved. In each trial, two random graphs from the **IMDB-Binary** (Yanardag and Vishwanathan, 2015) actor collaboration graph dataset (1000 graphs with 19.77 nodes and 96.53 edges on average) are selected. The nodes of the graphs are endowed with uniform distributions  $p$  and  $q$  (results obtained when using other distributions from the family (4) are similar). An ensemble of couplings between these measures is generated by running a custom Markov Chain Monte Carlo (MCMC) hit-and-run sampler (Smith, 1984) on the coupling polytope  $\tilde{\mathcal{C}}(p, q)$  (details in Supplementary Materials). We sampled 100 points in the polytope by running 100,000 MCMC steps and subsampling uniformly. Using each coupling in the ensemble as an initialization, we run projected gradient descent on adjacency loss (3) and spectral loss (5) with  $t = 5, 10$  and  $20$  and record the loss at the local minimum from each initialization. This process is repeated 100 times (100 choices of pairs of graphs).

Statistics for the experiment are reported in Table 1. For each method, we report the mean time for convergence of each gradient descent. For each trial, we obtain a distribution of losses from the 100 initializations in the ensemble, with the minimum loss treated as the putative global minimum. The “Worst Error” for each trial is  $(\max \text{ loss} - \min \text{ loss}) / \min \text{ loss}$ . We report the mean Worst Error over 100 trials. In available packages, gradient descent for GW matching is by default initialized with the product coupling  $C = pq^T$ , so we also report the “Product Error” (product loss – min loss)/min loss, averaged over all trials. We note that the absolute losses of the adjacency and heat kernel methods are not directly comparable, which is why we report these relative errors. The heat kernel representations provide an order of magnitude speed up of convergence, with decreasing error as the  $t$  parameter increases; e.g., for  $t = 20$ , all initializations converge to a coupling with less than 0.1% error. More details are provided in the Supplementary Materials.

**Graph Matching and Averaging.** Let  $(G, p)$  and  $(H, q)$  be measure graphs. Minimizers of the loss functions (3) and (5) are couplings  $C$  giving soft node correspondences between  $G$  and  $H$ . To assess the intuitive meaning of such a coupling, it is useful to visualize this node correspondence at the graph level. We produce such visualizations using ideas from Chowdhury and Needham (2020), as shown in Figure 1. The figure shows six separate examples. For any particular example, we display an interpolation between graph  $G$  (on the left) and graph  $H$  (on the right). The optimal coupling  $C$  is used to interpolate node positions from  $G$  to  $H$ , with new edges phasing in during the interpolation—see Supplementary Materials for the details of the visualization algorithm as well as another experiment illustrating improved stability of the graph averaging algorithm of Peyré et al. (2016) when using spectral loss. We

Table 2: Node Correctness, Mean  $\pm$  St. Dev. (Time).

Dataset	Adj	Spec <sup>10</sup>
Proteins	.68 $\pm$ .22 (31.9)	.78 $\pm$ .22 (5.1)
Enzymes	.70 $\pm$ .18 (8.9)	.79 $\pm$ .17 (1.4)
Reddit	.29 $\pm$ .21 (3941.7)	.50 $\pm$ .11 (206.1)
Collab	.50 $\pm$ .27 (4.3)	.50 $\pm$ .27 (5.6)

Figure 2: Plots of AMI and modularity for stochastic block model graphs across  $t$  parameters with clearly correlated peaks. **Inset:** Sample adjacency matrices for SBMs with increasing cross-block densities.

observe that in all cases, the matching produced via spectral loss preserves large scale qualitative features of the graphs more faithfully than adjacency-based matchings.

We also assess the quality of node correspondences quantitatively. In this experiment, we consider two biological graph databases **Proteins** (Borgwardt et al., 2005) (1113 graphs with 39.06 nodes and 72.82 edges on average) and **Enzymes** (Dobson and Doig, 2003; Schomburg et al., 2004) (600 graphs with 32.63 nodes and 62.14 edges on average), and two social graph databases **Reddit** (subset of 500 graphs with 375.9 nodes and 449.3 edges on average) and **Collab** (subset of 1000 graphs with 63.5 nodes and 855.6 edges on average), both from Yanardag and Vishwanathan (2015). All datasets were downloaded from Kersting et al. (2016). For each graph  $G = (V, E)$ , we assign a node distribution  $p$  from (4) (parameters tuned for the best performance in each method). A ‘‘new’’ measure graph  $(H, q)$  is created by randomly permuting node labels of  $G$ . There is a ground truth node correspondence between  $G$  and  $H$  and the goal is to measure the ability of GWL and SpecGWL to recover it. Given a coupling  $C$  of  $(G, p)$  and  $(H, q)$ , we measure its performance by the *node correctness score*  $|S \cap S_{GT}|/|S|$ , where  $S = \{(i, j) \mid C_{ij} > \epsilon\}$ , for  $\epsilon > 0$  a small threshold parameter, is the set of node correspondences from  $C$  and  $S_{GT}$  comprises ground truth node correspondences. Xu et al. (2019a) showed that the adjacency-based GWL framework achieves state-of-the-art graph matching performance with respect to this metric.

For each graph  $G$  and permuted version  $H$ , we compute couplings minimizing adjacency loss (3) and spectral loss (5) with  $t = 10$  and compute their node correctness scores. Mean scores for each dataset are provided in Table 2, where we see that spectral loss outperforms adjacency, except on the dense **Collab** graph where results agree. Minimizers of spectral loss in SpecGWL were computed via standard gradient descent. Minimizers of adjacency loss in GWL were computed via both gradient descent and the regularized proximal gradient descent method of Xu et al. (2019a), with the best results we obtained reported in Table 2 (proximal for the biological graphs, standard for the social graphs). We have empirically found that  $t = 10$  gives good matching performance for graphs with tens or hundreds of nodes, and moreover that performance is generally quite robust to this parameter choice (cf. illustration in Supplementary Materials).

**Optimizing the Scale Parameter.** SpecGWL requires determining an appropriate scale parameter  $t$ , and here we propose two methods to this end in the context of graph partitioning. We first propose a supervised method for tuning  $t$

Table 3: GRP model results; averaged AMI (time).

$p_{out}$	Infomap	GWL	SpecGWL
.08	.676 (.34)	.611 (1.44)	<b>.845</b> (.62)
.10	—	.624 (1.56)	<b>.820</b> (.70)
.12	—	.615 (1.64)	<b>.820</b> (.75)
.15	—	.611 (1.82)	<b>.788</b> (.87)

Table 4: Comparison of adjusted mutual information scores across a variety of datasets.

Dataset		Fluid	FastGreedy	Louvain	Infomap	GWL	SpecGWL
Wikipedia	sym, raw	—	0.382	0.377	0.332	0.312	<b>0.442*</b>
	sym, noisy	—	0.341	0.329	0.329	0.285	<b>0.395</b>
	asym, raw	—	—	—	0.332	0.178	<b>0.376</b>
	asym, noisy	—	—	—	<b>0.329</b>	0.170	0.307
EU-email	sym, raw	—	0.312	0.447	0.374	0.451	<b>0.487</b>
	sym, noisy	—	0.251	0.382	0.379	0.404	<b>0.425</b>
	asym, raw	—	—	—	<b>0.443</b>	0.420	0.437
	asym, noisy	—	—	—	0.356	<b>0.422</b>	0.377
Amazon	raw	—	0.637	0.622	<b>0.940</b>	0.443*	0.692
	noisy	0.347	0.573	<b>0.584</b>	0.463	0.352	0.441
Village	raw	—	<b>0.881</b>	<b>0.881</b>	<b>0.881</b>	0.606*	0.801*
	noisy	—	0.778	<b>0.827</b>	0.190	0.560	0.758

\*Slight improvements possible with proximal gradient, but overall performance rankings are preserved.

via **cross-validation**. We generate a dataset (illustration in Supplementary Materials) of  $n = 10$  stochastic block model networks with blocks of uniformly random sizes in the range [20, 50], within-block edge density 0.5, and uniformly random across-block edge densities in the range [0, 0.35]. Following a leave-one-out cross-validation scheme, we take  $n$  subsets of the data and run  $n$  trials. In trial  $j$  we use subset  $j$  as a training set to optimize the scale parameter  $t$  via a grid search. Specifically, the optimal  $t$  was the one that maximized the sum of the *adjusted mutual information* (AMI) scores computed between the ground truth clusterings and the clusterings recovered by SpecGWL across the training set. We then evaluate the performance of SpecGWL on the test set according to the AMI score. Additionally we evaluate the following baselines: **Fluid** (Parés et al., 2017), **FastGreedy** (Clauset et al., 2004), **Louvain** (Blondel et al., 2008), **Infomap** (Rosvall and Bergstrom, 2008), and the adjacency-based GWL (Xu et al., 2019a). Average AMI scores were: (Fluid, 0.59), (FastGreedy, 0.48), (Louvain, **0.65**), (Infomap, 0.0), (GWL, 0.62), (SpecGWL, **0.65**), where the results for Fluid and Louvain were tuned via gridsearch over their parameters (number of communities and *resolution*, respectively). Additionally, SpecGWL outperformed all other baselines in 5 out of 10 trials, with Louvain winning 4 and GWL winning the remaining trial. Moreover, SpecGWL outperformed GWL in 7 out of 10 trials.

We also propose a fully unsupervised method to tune the scale parameter via **modularity maximization**: here we use the *Newman modularity Q-score* as a proxy for AMI—Figure 2 illustrates how the peaks of AMI and modularity occur together across the scale parameter axis using synthetic data generated from a stochastic block model. In our unsupervised tuning method, the number of clusters  $k$  is selected by maximizing the modularity (which requires no ground truth partition) of a SpecGWL partitioning with respect to a fixed  $t$ -value (we use  $t = 10$  at this stage, which we generally found to be stable to parameter perturbation for graphs on the order of 100s to 1000s of nodes) over a range of  $k$ -values. Once a  $k$  has been selected, the scale parameter  $t$  is also selected by maximizing modularity over a range of  $t$  values. To illustrate this method with an unsupervised task, we test partitioning performance on Gaussian Random Partition model networks. This is a parametric model for random graphs, where the parameters are number of nodes (we use 1000), mean number of nodes per cluster (we use 150), a probability of within cluster edges (we use 0.5) and a probability of out-of-cluster edges  $p_{out} \in \{.08, .10, .12, .15\}$ . For each value of  $p_{out}$ , we construct a random graph on which SpecGWL is trained to determine number of clusters and optimal  $t$ -value. We then create 10 random graphs, partition them and compute AMI against ground truth clusters. Since this is a directed model, the only algorithms that apply are Infomap, GWL and SpecGWL. Average AMI scores and compute times for each  $p_{out}$  are reported in Table 3.

**Graph Partitioning on Real Data.** We compare the performance of SpecGWL on a graph partitioning task against the baseline methods described earlier for four real-world datasets. The first is a directed **Wikipedia** hyperlink network

(Leskovec and Krevl, 2014) that we preprocessed by choosing 15 webpage categories and extracting their induced subgraphs. The resulting digraph had 1998 nodes and 2700 edges. The second was obtained from an **Amazon** product network (Leskovec and Krevl, 2014) by taking the subgraph induced by the top 12 product categories. The resulting graph had 1501 nodes and 4626 edges. The third dataset was a digraph of email interactions between 42 departments of a European research institute (**EU-email**), and it comprised 1005 nodes and 25571 edges. The final dataset was a real-world network of interactions (8423 edges) among 1991 residents of 12 Indian villages (Banerjee et al., 2013), which we refer to as the **Village** dataset. We created noisy versions of each graph by adding up to 10% additional edges, and also created symmetrized versions of the Wikipedia and EU-email graphs by adding reciprocal edges.

The quality of each graph partition was measured by computing the AMI score against the ground-truth partition. The scores reported in Table 4 are obtained using parameters selected through the unsupervised procedure described in the previous section. Reported results use standard gradient descent to compute GWL and SpecGWL scores. Despite issues with numerical instability, we also computed scores via the regularized proximal gradient method of Xu et al. (2019a) where possible. Slight score improvements are possible in some cases, but overall score rankings between methods are unchanged—more experimental details are provided in the Supplementary Materials. We find that SpecGWL is the most consistent leader across all methods; in particular, it consistently produces improved results compared to GWL, the only other comprehensive method which is also applicable to graph matching and averaging. The performance of SpecGWL on directed graphs is especially relevant, considering that its closest competitor Infomap is a state-of-the-art method for digraph partitioning.

## 5 DISCUSSION

We have introduced a spectral notion of GW distance for graph comparison problems based on comparing heat kernels rather than adjacency matrices. This spectral variant is shown qualitatively and quantitatively to improve performance in graph matching and partitioning tasks. The techniques introduced here should be useful for studying further variants of GW distances. For example, work of Peyré et al. (2016) suggests that replacing the  $L^2$ -type spectral loss (5) with a loss based on KL divergence could have benefits when performing statistical analysis on graph datasets. We also remark that, while spectral GW is faster than its adjacency counterpart on smaller graphs, it may not enjoy the same scalability properties due to the lack of sparse matrix operations. A significant direction of future work will be to construct multiscale approaches to analyze large scale graphs through both adjacency and spectral methods: the divide-and-conquer techniques of Xu et al. (2019a) can break graphs into manageable chunks, and then SpecGWL can finish the computation with improved runtime and accuracy. On the theoretical front, Theorems 2 and 3 suggest that it is tractable to study spectral GW rigorously, and developing Theorem 3 into a larger theory should illuminate further connections between optimal transport and spectral graph theory. It will also be interesting to explore more fully the dependence of our algorithms on the scale parameter  $t$  and to incorporate multiple  $t$ -values into computations.

**Acknowledgments** We would like to thank the reviewers on this work for their numerous helpful comments. We wish to also thank Facundo Mémoli for suggesting references related to the original work on Spectral GW.

## References

- David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.
- David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.
- Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.
- Amélie Barbe, Marc Sebban, Paulo Gonçalves, Pierre Borgnat, and Rémi Gribonval. Graph diffusion Wasserstein distances. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2020.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl\_1):i47–i56, 2005.
- Michael M Bronstein and Klaus Glashoff. Heat kernel coupling for multiple graph analysis. *arXiv preprint arXiv:1312.3035*, 2013.

- Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *International Conference on Machine Learning*, pages 851–861, 2019.
- Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33, 2020.
- Lénaïc Chizat. *Transport optimal de mesures positives: modèles, méthodes numériques, applications*. PhD thesis, Université Paris-Dauphine, 2017.
- Samir Chowdhury and Facundo Mémoli. The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.
- Samir Chowdhury and Tom Needham. Gromov-Wasserstein averaging in a Riemannian framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 842–843, 2020.
- Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.
- Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- Nicoló De Ponti and Andrea Mondino. Entropy-transport distances between unbalanced metric measure spaces. *arXiv preprint arXiv:2009.10636*, 2020.
- Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.
- Yihe Dong and Will Sawin. COPT: Coordinated optimal transport on graphs. *Advances in Neural Information Processing Systems*, 33, 2020.
- Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.
- Rémi Flamary and Nicolas Courty. POT: Python Optimal Transport library. 2017. URL <https://github.com/rflamary/POT>.
- Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- Reigo Hendrikson. Using Gromov-Wasserstein distance to explore sets of networks. Master’s thesis, University of Tartu, 2016.
- Nan Hu, Raif M Rustamov, and Leonidas Guibas. Stable and informative spectral signatures for graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2305–2312, 2014.
- Atsushi Kasue and Hironori Kumura. Spectral convergence of Riemannian manifolds. *Tohoku Mathematical Journal, Second Series*, 46(2):147–179, 1994.
- Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. Benchmark data sets for graph kernels, 2016. URL <http://graphkernels.cs.tu-dortmund.de>.
- Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.
- Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- Facundo Mémoli. On the use of Gromov-Hausdorff distances for shape comparison. *The Eurographics Association*, 2007.
- Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011a.
- Facundo Mémoli. A spectral notion of Gromov–Wasserstein distance and related methods. *Applied and Computational Harmonic Analysis*, 30(3):363–401, 2011b.
- Facundo Mémoli and Tom Needham. Gromov-Monge quasi-metrics and distance distributions. *arXiv preprint arXiv:1810.09646*, 2018.

- Huda Nassar, Nate Veldt, Shahin Mohammadi, Ananth Grama, and David F Gleich. Low rank spectral network alignment. In *Proceedings of the 2018 World Wide Web Conference*, pages 619–628, 2018.
- Ferran Parés, Dario Garcia Gasulla, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. Fluid communities: a competitive, scalable and diverse community detection algorithm. In *International Conference on Complex Networks and their Applications*, pages 229–240. Springer, 2017.
- Rob Patro and Carl Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.
- Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.
- Martin Reuter, Franz-Erich Wolter, and Niklas Peinecke. Laplace–Beltrami spectra as ‘Shape-DNA’ of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.
- Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
- Bernhard Schmitzer and Christoph Schnörr. Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. *Journal of mathematical imaging and vision*, 46(1):143–159, 2013.
- Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl\_1):D431–D433, 2004.
- Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The Unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. *arXiv preprint arXiv:2009.04266*, 2020.
- Robert L Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.
- Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016.
- Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta mathematica*, 196(1):65–131, 2006.
- Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.
- Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.
- Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. NetLSD: hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2347–2356. ACM, 2018.
- Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alex Bronstein, Ivan Oseledets, and Emmanuel Müller. The shape of data: Intrinsic distance for data distributions. In *ICLR 2020: Proceedings of the International Conference on Learning Representations*, 2020.
- Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.
- Titouan Vayer, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284, 2019a.
- Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems*, pages 14726–14736, 2019b.
- Hongteng Xu. Gromov-Wasserstein factorization models for graph clustering. In *AAAI*, pages 6478–6485, 2020.
- Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems*, pages 3046–3056, 2019a.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-Wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning*, pages 6932–6941, 2019b.

Hongteng Xu, Dixin Luo, Ricardo Henao, Svatopluk Shah, and Lawrence Carin. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*, pages 10576–10586. PMLR, 2020.

Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374, 2015.

## A Proofs of Theorems

### A.1 Theorem 2

*Proof.* Suppose  $G = (V^G, E^G, p)$  and  $H = (V^H, E^H, q)$  have  $m, n$  nodes, respectively. Let  $C \in \mathcal{C}(p, q)$ . Then  $C \in [0, 1]^{m \times n}$  and satisfies  $(m + n - 1)$  linear equality constraints coming from the row and column sums. By Lemma 1, at least one of the minimizers of spectral loss (5) is located at an extreme point of the convex polytope  $\mathcal{C}(p, q)$ . This polytope lies in an  $(mn - (m + n - 1))$ -dimensional affine subspace of  $mn$ -dimensional space. The equality constraints automatically ensure that each  $C_{ij} < 1$ , where the strict inequality holds because the graphs are fully supported and thus each  $p_i, q_j < 1$ . Therefore, estimating the number of zero entries is equivalent to estimating the number  $k$  of active nonnegativity constraints. An extreme point corresponds to the intersection of  $k$  hyperplanes in general position with this affine subspace, and this intersection has dimension  $mn - (m + n - 1) - k$ . Because the extreme point has dimension 0, we have  $k = mn - (m + n - 1)$ . If the hyperplanes are not in general position, then the number of active nonnegativity constraints, i.e. the number of zeros, is greater than or equal to  $mn - (m + n - 1)$ . Next suppose  $m \sim n$ . Then the ratio of nonzero entries to total entries of  $C$  is roughly  $\frac{n^2 - k}{n^2} = \frac{2n - 1}{n^2}$ , and this term tends to 0 as  $n \rightarrow \infty$ .  $\square$

### A.2 Theorem 3

*Proof.* Let  $G = (V, E)$ , with  $|V| = n$ , be a graph satisfying the assumptions, endowed with uniform vertex distribution  $p$  and let  $K^t$  denote the heat kernel of  $G$ . By definition,

$$K^t = \Phi e^{-t\Lambda} \Phi^T = \sum_{j=1}^n e^{-t\lambda_j} \phi_j \phi_j^T,$$

where  $\Phi$  is a matrix whose columns are the orthonormal eigenvectors  $\phi_1, \dots, \phi_n$  of the graph Laplacian  $L$  of  $G$  and  $\Lambda$  is the diagonal matrix of sorted eigenvalues  $0 = \lambda_1 < \lambda_2 < \lambda_3 \leq \lambda_4 \leq \dots \leq \lambda_n$  of  $L$ . Let  $Q$  be the 2-way partitioning template from (7). Since  $p$  is uniform, the estimated distribution  $q$  is also uniform and  $Q = \frac{1}{2}I_2$ , where  $I_2$  is the  $2 \times 2$  identity matrix.

The 2-way spectral GW partitioning of  $G$  is obtained from a coupling minimizing the spectral partitioning loss (8). Using Lemma 1, we see that this optimization task is equivalent to maximizing

$$C \mapsto \langle K^t C, C Q \rangle = \frac{1}{2} \langle K^t C, C \rangle$$

over the coupling polytope  $\mathcal{C}(p, q)$ . Since the factor of  $\frac{1}{2}$  does not effect the optimization, we suppress it and further simplify the objective function as

$$\langle K^t C, C \rangle = \text{tr}((K^t C)^T C) = \text{tr}\left(CC^T \sum_{j=1}^n e^{-t\lambda_j} \phi_j \phi_j^T\right) = \sum_{j=1}^n e^{-t\lambda_j} \text{tr}(CC^T \phi_j \phi_j^T).$$

Since the leading eigenvector is  $\phi_1 = \frac{1}{\sqrt{n}}1^{n \times 1}$  (the normalized vector of all ones), it is easy to check that the term  $\text{tr}(CC^T \phi_1 \phi_1^T)$  is constant for all  $C \in \mathcal{C}(p, q)$ . The objective therefore becomes to maximize over  $C \in \mathcal{C}(p, q)$  the quantity

$$\sum_{j=2}^n e^{-t\lambda_j} \text{tr}(CC^T \phi_j \phi_j^T) = e^{-t\lambda_2} \left( \text{tr}(CC^T \phi_2 \phi_2^T) + \sum_{j=3}^n e^{-t(\lambda_j - \lambda_2)} \text{tr}(CC^T \phi_j \phi_j^T) \right),$$

which is in turn equivalent to maximizing

$$C \mapsto \text{tr}(CC^T \phi_2 \phi_2^T) + \sum_{j=3}^n e^{-t(\lambda_j - \lambda_2)} \text{tr}(CC^T \phi_j \phi_j^T). \quad (9)$$

Observe that the summation term goes to zero as  $t \rightarrow \infty$  (using  $\lambda_j > \lambda_2$  for all  $j \geq 3$ ), while the first term is independent of  $t$ . It follows that, for sufficiently large  $t$ , maximization of (9) is equivalent to maximizing

$$C \mapsto \text{tr}(CC^T \phi_2 \phi_2^T) \quad (10)$$

over  $\mathcal{C}(p, q)$ . It then remains to study the structure of maximizers of (10).

We further simplify the objective function (10) as

$$\text{tr}(CC^T\phi_2\phi_2^T) = \text{tr}((C^T\phi_2)^T(C^T\phi_2)) = \|C^T\phi_2\|^2,$$

where the norm in the last line is the Frobenius norm. We denote the column vectors of  $C$  by  $C_1, C_2 \in \mathbb{R}^{1 \times n}$ , so that

$$\|C^T\phi_2\|^2 = \left\| \begin{pmatrix} C_1 \cdot \phi_2 \\ C_2 \cdot \phi_2 \end{pmatrix} \right\|^2,$$

where the norm on the right is the Euclidean norm. Since  $C \in \mathcal{C}(p, q)$  and  $p$  is uniform, we have  $C_2 = \frac{1}{n}\mathbf{1}^{n \times 1} - C_1$ , whence

$$C_2 \cdot \phi_2 = \left( \frac{1}{n}\mathbf{1}^{n \times 1} - C_1 \right) \cdot \phi_2 = -C_1 \cdot \phi_2,$$

since  $\phi_2$  is orthogonal to  $\mathbf{1}^{n \times 1} = \sqrt{n}\phi_1$ . The objective (10) is finally reduced to

$$C \mapsto 2(C_1 \cdot \phi_2)^2. \quad (11)$$

Let  $\phi_2^+$  be the vector of positive entries of  $\phi_2$  with all negative entries thresholded to zero and likewise define  $\phi_2^-$  to be the vector of negative entries of  $\phi_2$ . Assume without loss of generality that  $\|\phi_2^+\| \geq \|\phi_2^-\|$  (the other case follows entirely similarly). Then in order to maximize (11), one should set each entry of  $C_1$  to be nonzero if and only if the corresponding entry of  $\phi_2$  is positive. The spectral GW partitioning therefore agrees with the Fiedler partitioning, and the proof is complete.  $\square$

## B An MCMC Sampler for Couplings.

Both the adjacency (3) and spectral (5) loss functions are nonconvex, and solving such problems effectively often relies on a clever choice of initialization. A limitation of the current practice is that this initialization is often chosen to be the *product coupling*  $pq^T$ , which we empirically find to be sub-optimal in even simple cases. This is accomplished by running gradient descent from each point in an ensemble of initializations generated by a Markov Chain Monte Carlo Hit-And-Run sampler (Smith, 1984). This algorithm is well-known, but we describe it below for the convenience of the reader. Our code includes a lean Python implementation written specifically for sampling the coupling polytope; we hope such an implementation will be useful to the broader optimal transport community.

---

### Algorithm 1 Markov chain sampler

---

```

1: function MARKOVSTEP( $A, p, q, C$ )
2:   //  $A$ : matrix of linear constraints
3:   //  $p, q : m \times 1, n \times 1$  probability vectors
4:   //  $C : m \times n$  initial coupling matrix
5:
6:    $V \leftarrow$  random  $m \times n$  matrix as direction
7:    $Q \leftarrow$  o.n. basis for row space of  $A$ 
8:    $V \leftarrow V - QQ^TV$  ▷ project  $V$  to correct subspace
9:   pos  $\leftarrow$  indices where  $V > 0$ 
10:  neg  $\leftarrow$  indices where  $V < 0$ 
11:   $\alpha \leftarrow \max(-C[\text{pos}]/V[\text{pos}])$ 
12:   $\beta \leftarrow \min(-C[\text{neg}]/V[\text{neg}])$  ▷  $[\alpha, \beta]$  is maximal range of step sizes
13:   $\gamma \leftarrow$  random element of  $[\alpha, \beta]$ 
14:  return  $C + \gamma V$  ▷ new coupling matrix
15: end function

```

---

## C Additional experiments and implementation details

### C.1 Additional Landscape Results

Figure 3 gives a more detailed view of the results reported in Table 1. For each plot, the  $x$ -axis is (Worst or Product) error percentage. The  $y$ -axis shows the percentage of samples whose error was above the relative error rate. We see that a significant number of samples have high error rates for adjacency loss (3) and spectral loss (5) with  $t = 5$ . For spectral loss with  $t = 10$  or 20, these error rates are greatly decreased. In particular, spectral loss with  $t = 20$  has essentially zero samples with error rate above 2%.

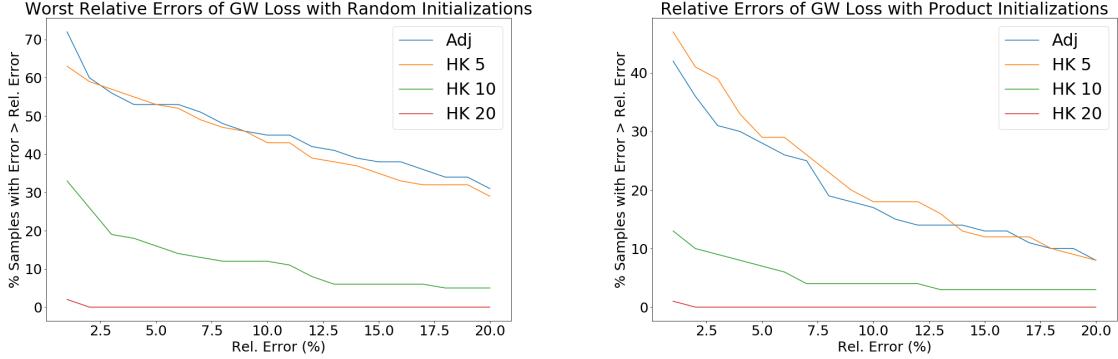
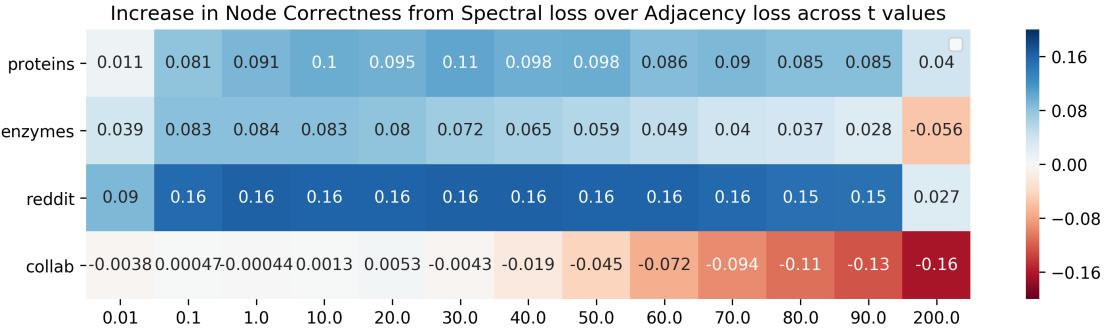


Figure 3: Results of energy landscape experiments.

Figure 4: Improvement (blue cells) obtained by using spectral loss (5) instead of adjacency loss (3) across a range of  $t$  values ( $x$  axis).

## C.2 Additional figures

Here we add some figures that help better understand some of the quantitative results presented in the main text. Figure 4 shows that the improvement in the graph matching experiment obtained via SpecGWL remains stable across a wide range of scale parameters. Specifically, we computed SpecGWL loss for  $t \in \{10, 20, \dots, 90\}$  for each of the four datasets. The **Collab** dataset was the only one where there was no appreciable improvement from using SpecGWL, but there was no significant decrease in performance either for scales in the range  $t \in \{10, 20, 30\}$ .

Figure 5 shows the 10 stochastic block model networks used in the synthetic graph partitioning experiment. Each network has 5 blocks, and the block sizes were chosen uniformly at random from the range [20, 50] at the beginning of the experiment. Within-block edge densities were fixed at 0.5, and across-block edge densities were chosen uniformly at random in the range [0, 0.3].

## C.3 Visualizing Graph Matchings

Here we describe how the interpolations used to visualize coupling quality in Figure 1 were produced. Let  $(G, p)$  and  $(H, q)$  be measure graphs and  $C \in \mathcal{C}(p, q)$  a coupling. To produce an interpolation, we first “blow up”  $C$  so that it has the form of a weighted permutation matrix. This is done by first scanning across rows; any row with more than a single nonzero entry is split into “dummy” copies, each of which contains a single nonzero entry from the original row. The splits allow us to split nodes of  $G$  into dummy copies, with weights given by entries in the corresponding row of  $C$ . The same procedure is applied to split columns of  $C$  and to split nodes of  $H$ . The result is a pair of expanded measure graphs  $(G', p')$  and  $(H', q')$  together with an expanded coupling  $C'$  which provides a bijective correspondence between the nodes of  $G'$  and  $H'$ . Once such a bijective correspondence is obtained, we position each graph  $G'$  and  $H'$  in the plane using a common embedding modality and then performing Procrustes alignment of the resulting embeddings. To interpolate the graphs, we simply interpolate positions of the bijectively matched nodes, while phasing in new edges that are formed. This visualization method has strong theoretical justification: building on work of Sturm (2012), it

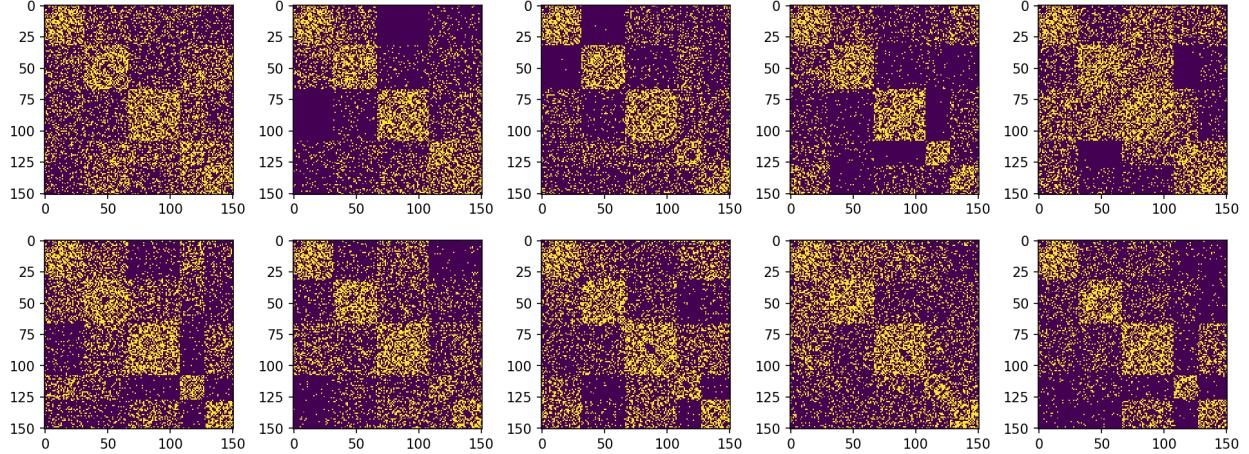


Figure 5: Stochastic block models used in the supervised partitioning task with cross-validation

is shown by [Chowdhury and Needham \(2020\)](#) that this process represents a geodesic (in the metric geometry sense) in the space of edge-weighted measure graphs. We observe that the conclusion of Lemma 1 is useful here, since the theoretical guarantee on the sparsity of  $C$  implies that  $C$  will not get too large in the “blow up” phase of the algorithm.

To produce each example in Figure 1, we sampled 100 couplings from the coupling polytope via the MCMC algorithm (1000 MCMC steps between each coupling) as initializations. We then computed an optimal coupling between the graphs by optimizing the relevant loss function from each initialization and keeping the coupling with the lowest loss from the resulting ensemble.

#### C.4 Dependence of Matchings on $t$ -Value

To understand the landscape of optimal couplings that occur for different scale parameters, we took two graphs from the **Enzymes** dataset and computed optimal couplings for spectral loss using 100 linearly spaced  $t$  values in the range  $[0, 50]$ . Figure 6 shows a t-SNE embedding (with perplexity = 15) obtained after flattening these couplings into vectors in Euclidean space. The inset coupling matrices are representatives of the points in each significant cluster. Interpolation visualizations for some of the coupling matrices from different clusters are provided in Figure 7.

#### C.5 Averaging

We use the observations regarding the energy landscape and the quality of matchings to show that in the *GW averaging* problem, using the heat kernel leads to 10x faster convergence than the adjacency matrix, and moreover, the heat kernel yields a more “unique” barycenter. Specifically, given measure network representations  $X_1, X_2, \dots, X_n$ , a Fréchet mean is an element of  $\arg \min_X \sum_i d_{\text{GW}}(X, X_i)^2$ . The objective of the GW averaging problem is to compute this barycenter, i.e. an average representation. In the Python OT package ([Flamary and Courty, 2017](#)), this barycenter is computed iteratively from a random initialization (cf. the `gromov_barycenter` function). As a proxy for the “uniqueness” of the barycenter, we compute the barycenter for multiple random initializations, and then take the variance of the distribution of Fréchet losses achieved by the barycenters.

We demonstrate this claim on the **Village** dataset. We ran a bootstrapping procedure to sample 10 sets of 30 nodes, and took the induced subgraphs to obtain 10 subgraphs. To keep the samples from being too sparse, we first sorted the nodes in order of decreasing betweenness centrality, and then selected 30 nodes (for each iteration) from the top 40 nodes with the highest centrality. Next we computed both adjacency and heat kernel representations (for  $t = 3, 7, 11$ ) of these subgraphs. Then we used the `gromov_barycenter` function to compute averages of the adjacency and heat kernel representations. Each call to `gromov_barycenter` uses a random initialization. Using this randomness as a source of stochasticity, we repeated the set of barycenter computations 10 times to obtain four distribution of Fréchet losses. After mean-centering the distributions, the variance of the adjacency distribution was found to be two orders of magnitude higher than any of the heat kernel distributions, and each of the three comparisons was found to be statistically significant by computing Bartlett tests for unequal variance ( $p < 10^{-6}$  for all, adjusted for multiple comparison via Bonferroni correction). Boxplots of the results are shown in Figure 8.

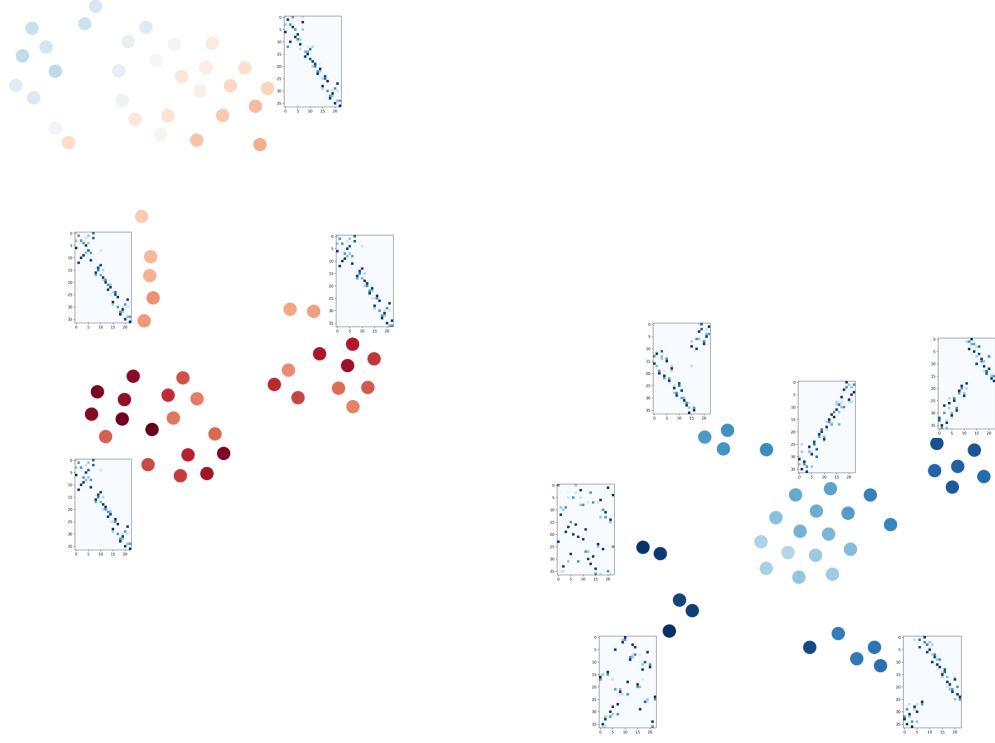


Figure 6: t-SNE embedding of optimal couplings between two graphs from the **Enzymes** dataset obtained via spectral loss (5) using 100 equally spaced  $t$  values in the range  $[0, 50]$ . The ground metric is the  $l^2$  norm between vectorized representations of the couplings. The range  $[0, 50]$  is mapped linearly from the blue to red color scheme.

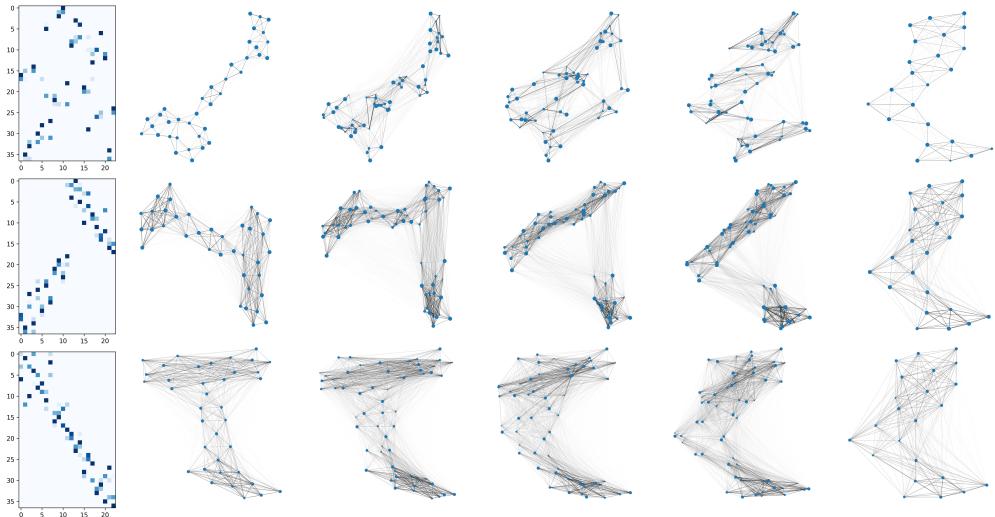


Figure 7: Interpolation visualizations obtained via some of the representative couplings in Figure 6 arranged top to bottom with increasing  $t$ . Note that the interpolation in the third row, corresponding to the largest  $t$  value, represents global structure more faithfully.

Figure 8: **Left:** Differences in Fréchet loss of the GW average across representations. **Center:** Mean-centered Fréchet loss, indicating the greater variance and sensitivity to initialization for the adjacency representation. **Right:** Distribution of runtimes shows 10x speedup for the heat kernel.

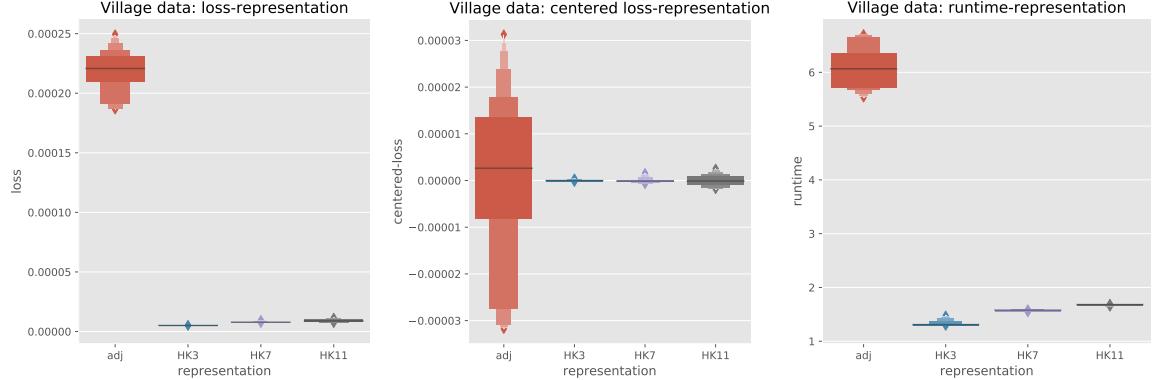


Table 5: Comparison between runtime of GWL and average runtime of SpecGWL across  $t$  parameters. “-Prox” rows use regularized proximal gradient with Sinkhorn iterations as used by Xu et al. (2019a). Other rows use vanilla gradient descent.

Method	Wikipedia				EU-email				Amazon		Village	
	sym		asym		sym		asym		raw	noisy	raw	noisy
	raw	noisy										
GWL	14.1	16.1	16.0	14.2	1.5	6.7	<b>0.9</b>	1.6	8.6	13.1	3.3	5.4
SpecGWL	<b>1.8</b>	<b>2.3</b>	<b>2.4</b>	<b>2.4</b>	<b>1.0</b>	<b>0.9</b>	<b>0.9</b>	<b>1.0</b>	<b>1.4</b>	<b>0.9</b>	<b>1.8</b>	<b>2.7</b>
GWL-Prox	—	—	—	—	<b>0.9</b>	<b>0.8</b>	—	—	<b>1.2</b>	<b>1.0</b>	2.2	2.2
SpecGWL-Prox	2.9	2.6	2.9	2.9	<b>0.9</b>	1.0	1.0	1.0	1.3	1.8	<b>1.8</b>	<b>2.0</b>

## C.6 Graph Partitioning

Runtimes for GWL and SpecGWL on the graph partitioning experiment are reported in Table 5. For SpecGWL, the times are averaged over several values of  $t$ , with the idea that finding the correct  $t$ -value is a preprocessing hyperparameter tuning step. For both GWL and SpecGWL, partitionings were obtained using standard projected gradient descent. Speedups are obtained for GWL via the regularized proximal gradient method, but we were not able to obtain results on all datasets with this method due to numerical issues (see below). Runtimes for this method are also reported as GWL-Prox. We observe that spectral loss provides up to 10x acceleration in convergence rate for standard gradient descent and even outperforms the proximal gradient in compute time.

When employing the regularized proximal gradient method, we found that the results were sensitive to the choice of regularization parameter  $\beta$  (as is also observed by Xu et al. (2019a)), leading to numerical blowups if not chosen carefully. In reporting each of the results below, we hand-tuned  $\beta$  after testing in the  $10^{-1}, 10^{-2}, 10^{-3}, \dots, 10^{-9}$  regimes. For **Wikipedia**, we used  $\beta = 2 \cdot 10^{-5}$  for SpecGWL, but were unable to find a  $\beta$  that provided stable results for GWL. For **EU-email**, we used  $2 \cdot 10^{-7}$  for GWL and  $3 \cdot 10^{-8}$  for SpecGWL. For **Amazon**, we used  $\beta = 4 \cdot 10^{-3}$  for GWL and  $1.5 \cdot 10^{-6}$  SpecGWL. Finally, for **Village** we used  $\beta = 5 \cdot 10^{-6}$  for SpecGWL. This  $\beta$  led to numerical instability for GWL, but  $\beta = 5 \cdot 10^{-5}$  worked and yielded the results we report below. In summary, it appears that the structure of the graph has a significant effect on the optimal choice of regularization parameter (e.g. the Wikipedia graph is relatively very sparse). Because the numerical instability issues are very sensitive to the regularization, one avenue for future work could be to incorporate the strategies described in the PhD thesis of Chizat (2017) (e.g. “absorption into the log domain”) to stabilize the regularized proximal gradient method.

Table 6: Performance of GWL and SpecGWL using regularized proximal gradient descent and Sinkhorn iterations as in Xu et al. (2019a). ‘—’ denotes that an AMI score could not be calculated due to numerical instability.

Method	Wikipedia				EU-email				Amazon		Village	
	asym		sym		asym							
	raw	noisy	raw	noisy	raw	noisy	raw	noisy	raw	noisy	raw	noisy
GWL-Prox	—	—	—	—	<b>0.45</b>	<b>0.40</b>	—	—	0.49	0.39	0.72*	0.58
SpecGWL-Prox	0.51	0.39	0.39	0.29	0.01	0.01	0.03	0.03	<b>0.66</b>	<b>0.43</b>	<b>0.84</b>	<b>0.72</b>

\*The code provided by Xu et al. (2019a) included a representation matrix as `database['cost']`, and this yielded the score of 0.72. However, this matrix was asymmetric and not equal to the symmetrized adjacency matrix that was used in experiments with other benchmarks. When using a symmetrized adjacency matrix, the score drops from 0.72 to 0.66.