Gromov-Wasserstein Graph Coarsening

Carlos A. Taveras **Santiago Segarra** **César A. Uribe**
Department of Electrical and Computer Engineering
Ken Kennedy Institute
Rice University, Houston, TX 77005
{ctaveras, segarra, cauribe}@rice.edu

Abstract

We study the problem of graph coarsening within the Gromov-Wasserstein geometry. Specifically, we propose two algorithms that leverage a novel representation of the distortion induced by merging pairs of nodes. The first method, termed Greedy Pair Coarsening (GPC), iteratively merges pairs of nodes that locally minimize a measure of distortion until the desired size is achieved. The second method, termed k-means Greedy Pair Coarsening (KGPC), leverages clustering on pairwise distortion metrics to merge clusters of nodes directly. We provide conditions under which the algorithms are guaranteed to provide an optimal coarsening and validate their performances on six large-scale datasets and a downstream clustering task. Results show that the proposed approaches outperform existing approaches on a wide range of parameters and scenarios.

1 Introduction

With the advent of Graph Neural Networks (GNNs) [1], there has been an explosion in the development of data-driven algorithms for processing structured data including graphs and generalizations thereof (e.g., simplicial complexes, cell complexes, and hypergraphs) [2–10]. These developments are part of a broader trend in data science of leveraging topological, geometric, and algebraic structures to process non-Euclidean data [11–17], and have been applied to solve problems in drug discovery [18], social network analysis [19], finance [20], and wireless communications [21]. The performance of these models depends critically on the quantity and quality of the data used to train them. Consequently, model training can be both time- and resource-intensive, often prohibitively so.

There are three main dimensionality reduction paradigms for graphs [22]: coarsening [23–26], sparsification [26, 27], and condensation [28], differing in how they consolidate graph structures. Graph coarsening reduces graph size by partitioning and merging sets of nodes in a way that minimizes a chosen reconstruction error; graph sparsification reduces the complexity of the graph by removing a subset of nodes and edges; graph condensation learns small graphs that synthesize aspects of the original graph, including node features and topology. *We focus in this work on graph coarsening.*

Graph coarsening has a history dating back to at least the introduction of the Kron reduction along [29] and has long been applied in scientific computing to solving differential equations [23]. Beyond this, graph coarsening algorithms have been employed for learning network node embeddings [30], image segmentation [31], and neighborhood pooling in GNNs [32]. For further elaboration on the history of graph coarsening, see [22, 23, 33]. Central to the design of all coarsening algorithms, no matter the application, is the question of what notion of similarity to preserve between the original graph and its coarsened counterpart. In graph data science, there has been an emphasis on preserving some notion of spectral similarity [25, 26, 34, 35], usually with respect to graphs' Laplacian representation. While many interesting graph properties can be derived that are related to their spectra [36], there are some properties that cannot be (several graphs may have the same spectrum, i.e., *cospectral graphs* [37]).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Graph Machine Learning.

In recent years, a metric that has garnered much interest in graph data science and, more recently, in graph coarsening is the Gromov-Wasserstein (GW) distance. The GW distance [38, 39] is appealing for use in graph mining for several reasons. First, the GW distance can be computed between unaligned graphs of different sizes, unlike the commonly used Frobenius norm distance [40] and the Bures-Wasserstein distance [41, 42]. Second, the GW distance produces an alignment (transport plan) between nodes across graphs, finding utility in applications including graph matching and partitioning [43]. While solving for such an alignment is generally NP-hard, many algorithms have been proposed to efficiently approximate the alignment [44, 45]. Third, we can easily compute geodesics between a pair and barycenters between a group of networks. To this end, GW geodesic and barycenter-based methods have been used for applications including data augmentation [46], graph matching [43], graph clustering [47], time-series prediction [48], and molecule prediction [49]. Finally, the Gromov-Wasserstein distance induces an equivalence relationship between networks of different sizes, which can be used to reduce the size of a graph to its minimal representative without any loss in information (see Appendix B).

Graph coarsening in the GW setting has been previously considered in [24], where, using the signless Laplacian representation, it was shown that the GW distance can be bounded by a spectral distance. In this case, the coarsening problem is approximated using the weighted kernel k-means algorithm. Moreover, [24] experimentally demonstrates that prioritizing spectrum preservation may not be ideal for tasks including graph classification, giving merit to the utility of GW distance preservation. We focus on developing graph coarsening algorithms that minimize GW distance, though we put no restrictions on the graph representation used. Towards this goal, we propose two algorithms, Greedy Pair Coarsening (GPC) and k-means Greedy Pair Coarsening (KGPC), which exploit the similarity of a pair of nodes, as characterized by merging distortion, to coarsen graphs. We summarize our contributions as follows:

1. We propose an iterative graph coarsening method that, under appropriate assumptions, is guaranteed to recover the smallest representation of a measure network within its weak isomorphism class.
2. We propose a novel network representation, based on the distortion induced by merging node pairs, for use in a k-means clustering method, providing a more efficient alternative to the iterative method.
3. We corroborate the utility of our algorithms by comparing the distortion induced by coarsening against several established algorithms and their performance on downstream tasks like graph classification.

2 Preliminaries

Graphs: A (weighted) graph $G = (V, E, w)$ is a triplet consisting of a finite set of nodes $V = \{v_1, \cdots, v_n\}$, a set of edges $E \subseteq V \times V$, and a mapping $w : E \to \mathbb{R}$ of edges to real numbers. A graph is undirected if $w(v_i, v_j) = w(v_j, v_i)$ for all $(v_i, v_j) \in E$. We can represent a graph by its adjacency matrix $A \in \mathbb{R}^{n\times n}$ where $a_{ij} = w(v_i, v_j)$ if $(v_i, v_j) \in E$; otherwise, $a_{ij} = 0$. Note that if G is undirected, its adjacency matrix is symmetric. The adjacency matrix is but one choice of representation for graphs; undirected graphs without self-loops (i.e. $(v, v) \notin E$ for all $v \in V$) are often represented by their Laplacian matrix $L = D - A$, where $D = \text{diag}(A\mathbf{1}_n)$, and its variants, including the signless Laplacian $L' = D + A$ which is used in [23]. We denote the choice of matrix representation for a graph by $S \in \mathbb{R}^{|V|\times|V|}$ and the weight of the edge from v_i to v_j by s_{ij}.

Basics of Gromov-Wasserstein Pseudo-Metrics: The Gromov-Wasserstein (GW) "distance" [38,39] is a pseudometric on the space of measure networks [50]. A finite (measure) network (X, W_X, μ_X) is a triplet consisting of a finite set X of nodes, a weight function $W_X : X \times X \to \mathbb{R}$, and a fully-supported probability measure μ_X. Abusing notation, we also denote measure networks by (S, μ), where $S \in \mathbb{R}^{|V|\times|V|}$ is a matrix for which $s_{ij} = S_X(x_i, x_j)$ and $\mu = \mu_X$. We denote the set of all finite measure networks by $\mathcal{N}$ and of N-node measure networks by $\mathcal{N}_N$. Hereafter, we use the terms measure network and network interchangeably.

A (measure) coupling π of two networks X and Y is a probability measure on the product space $X \times Y$ satisfying the marginal constraints $\mu_Y(y) = \sum_{x_i} \pi(x_i, y)$ and $\mu_X(x) = \sum_j \pi(x, y_j)$. We denote the set of all such couplings by $\Pi(\mu_X, \mu_Y)$. The distortion of X and Y with respect to the

measure coupling $\pi \in \Pi(\mu_X, \mu_Y)$ is defined by

$$\text{dis}^2(\pi) = \langle \mathcal{L}_2^2(X,Y) \otimes \pi, \pi \rangle. \tag{1}$$

where $\mathcal{L}_2^2(X,Y) \otimes \pi$ is the tensor product defined by

$$[\mathcal{L}_2^2(X,Y) \otimes \pi]_{ik} = \sum_{jl} |S_X(x_i, x_j) - S_Y(y_k, y_l)|^2 \pi(x_j, y_l), \tag{2}$$

and $\langle \cdot, \cdot \rangle$ is the Frobenius inner product [51]. The (L^2-) GW distance between networks X and Y is then the distortion of the infimizing coupling

$$d_{GW}^2(X,Y) = \inf_{\pi \in \Pi(\mu_X, \mu_Y)} \text{dis}^2(\pi), \tag{3}$$

implying that $d_{GW}(X,Y) \leq \text{dis}(\pi)$ for any $\pi \in \Pi(\mu_X, \mu_y)$. Equipped with the GW distance, $(\mathcal{N}, d_{GW})$ is a pseudometric space [50], i.e., it is a metric space up to weak isomorphism (see Appendix B). We use measure networks to model graphs; the graph (V, E, S) is represented by the measure network $G = (V, S, \mu)$, where μ can be chosen to reflect the relative importance of nodes. We set μ to the uniform measure over V by default. We denote the vector of weights emanating to/from a node $v \in V$ by $S(v) = [S(v, v_1), \cdots, S(v, v_n)]$.

Graph Coarsening: Graph coarsening partitions the node set V into a set with M sets, where $|V| = N > M$. We denote this mapping by $p : V \to \{1, \cdots, M\}$ and the set of nodes in the j-th partition set, or supernodes, by $P_j = p^{-1}(j)$. We can encode such a partition by an assignment matrix $C_p \in \{0,1\}^{N \times M}$ where

$$C_p(i,j) = \begin{cases} 1 & \text{if } v_i \in P_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We denote the space of all assignment matrices from N-node to M-node networks by $\mathcal{C}_{N,M}$. Given an assignment matrix $C_p \in \mathcal{C}_{N,M}$, we can form the average coarsening matrix $\mathcal{C}_w$ defined as

$$C_w = \text{diag}(\mu) C_p \text{diag}(1/\mu')$$

where $\mu' = C_p^\top \mu \in \mathbb{R}^M$ and $C_w \in \mathbb{R}^{N \times M}$. Coarsened graphs are constructed using average coarsening matrices as follows:

$$S' = C_w^\top S C_w, \quad \text{and} \quad \mu' = C_p^\top \mu \tag{5}$$

The coarsened matrix representation S' consolidates edge weights by taking a convex combination of the weights being merged with respect to the relative mass of the nodes being merged – see Equation 10 for an explicit characterization. Moreover, it was shown in [24, Appendix B.3] that the transformation of S defined in Eq. (5) is a semi-relaxed Gromov-Wasserstein barycenter [52] (see Appendix C).

3 Gromov-Wasserstein Graph Coarsening

Given an N-node measure network $G = (S, \mu)$, the Gromov-Wasserstein coarsening problem seeks an assignment matrix C_p^* that solves

$$C_p^* = \underset{C_p \in \mathcal{C}_{N,M}}{\arg\min} \|(S - C_p C_w^\top S C_w C_p^\top) \odot (\mu \mu^\top)^{1/2}\|_F^2. \tag{6}$$

Recall that coarsening reduces the size of a network by finding an optimal partition $\mathcal{P} = \{P_1, \cdots, P_M\}$ merging the nodes in these subsets to produce the coarsened network G'. If C_p encodes the assignment of nodes to partition sets (supernodes), the matrix $\pi = \text{diag}(\mu) C_p$ is a transport plan from G to G', for any $G' \in \mathcal{N}_M$ that is formed by merging nodes in G (i.e. no mass splitting). Given the assignment C_p, we can construct the measure coupling $\pi = \text{diag}(\mu) C_p$ and it was shown in [23, Appendix B.3] that

$$G' = \underset{G' \in \mathcal{N}_M}{\arg\min} \langle \mathcal{L}_2^2(G, G') \otimes \text{diag}(\mu) C_p, \text{diag}(\mu) C_p \rangle = (C_w^\top S C_w, C_p^\top \mu);$$

in other words, G' minimizes $\text{dis}(\pi)$ for a fixed C_p. Therefore, minimizing the distortion over C_p gives the best coarsening. Moreover, it can be shown that $d_{GW}(G', G'') = 0$ when $G'' =$

$(C_p C_w^\top S C_w C_p^\top, \mu)$. Taken together, we get Eq. (6); this formulation is closely related to the Gromov-Wasserstein sketching problem [53] which, given $G \in \mathcal{N}_N$, seeks the network $G' \in \mathcal{N}_M$ closest to G, or $G' = \arg\min_{G' \in \mathcal{N}_M} d_{GW}(G, G')$. The connection between coarsening and sketching is expounded on in Appendix C.

We propose two approaches to tackle the problem (6), both of which are based on the distortion induced by merging a pair of nodes; we leverage the intuition that nodes within the same partition should have similar neighborhoods. The first method greedily merges node pairs by choosing the coarsening from N to $N-1$ nodes with minimal discrepancy. The second method derives a graph representation from the distortion induced by merging node pairs, which partitions nodes using the k-means algorithm.

3.1 Greedy Pair Coarsening (GPC)

The first method we propose to solve Problem (6) is Greedy Pair Coarsening (GPC). First, note that we can reformulate Problem (6) as

$$\min_{C_p^1, \cdots, C_p^{N-M}} \|(S - R^\top S R) \odot (\mu\mu^\top)^{1/2}\|_F^2, \tag{7}$$

where M is the desired coarsening level, and $R = \prod_{j=1}^{N-M} C_p^j (C_w^j)^\top$, where each $C_p^j \in \mathcal{C}_{N-j+1, N-j}$ corresponds to the merging of a pair of nodes. We can approximate a solution to Problem (7) by solving a sequence of greedy optimization problems to produce $(C_p^i)_{i=1}^{N-M}$, each of which minimizes some intermediate cost. This sequence of assignment matrices can then be consolidated to produce an assignment matrix $C_p \in \mathcal{C}_{N-M}$ within the feasible set of Problem (6). In particular, given the first $i-1$ assignment matrices $(C_p^1, \cdots, C_p^{i-1})$, we compute C_p^i by

$$C_p^i = \underset{C_p^i \in \mathcal{C}_{N-i+1, N-i}}{\arg\min} \|((S - C_p^i (C_w^i)^\top (R^{i-1})^\top S R^{i-1} C_w^i (C_p^i)^\top) \odot (\mu\mu^\top)^{1/2}\|_F^2, \tag{8}$$

where $R^i = \prod_{j=1}^{i} C_p^j (C_w^j)^\top \in \mathbb{R}^{N \times N-i}$. Since $\mathcal{C}_{N-i+1, N-i}$ is a discrete set with $\binom{N-i+1}{2}$ elements, we can determine the minimizing transport plan for Eq. (8) directly by comparing distortions. Note that when the distortion induced by a node merging is zero, the coarsened network remains in the weak isomorphism class of the original network (see Proposition 1). Once the minimal representative is achieved, we must merge nodes that incur some distortion/error. This process is repeated until a network of the desired size is recovered. As a first theoretical result, we show that the GPC algorithm produces the minimal representative of the weak isomorphism class of a measure network.

Proposition 1. *Given a measure network G, GPC recovers the smallest network weakly isomorphic to G. Moreover, when k is the size of the minimal representative, GPC solves Problem* (6)*.*

For sufficiently well-behaved graphs, GPC can recover optimal partition sets.

Proposition 2. *Let $G = (V, S, \mu)$ be a symmetric N-node network whose nodes can be partitioned into sets $\mathcal{P} = \{P_1, \cdots, P_M\}$ for which there exist $\epsilon > 0$, $\alpha > 4 + 4\sqrt{N^2/(N-1)}$, satisfying $\|s(x) - s(y)\|_\infty < \epsilon$ for all $u_1, u_2 \in P_i$ and $\inf_{u \in V} |s(u_1, u) - s(u_2, u)| \geq \alpha\epsilon$ for $u_1 \in P_i, u_2 \in P_j$ for $i \neq j$. Then, for $G' = GPC(G, N-M) = (V^{(N-M)}, S^{(N-M)}, \mu^{(N-M)})$, we have $V^{(N-M)} = \mathcal{P}$.*

3.2 k-means Greedy Pair Coarsening (KGPC)

While Algorithm 1 is guaranteed to find the minimal representative of a measure network, its time complexity is $O(N^4)$ as we have to compute pairwise distortions for each iteration. To remedy this, we propose k-means Greedy Pair Coarsening (KGPC), which runs at $O(N^2 + TNM^2)$ where T bounds the number of k-means iterations and M is the desired coarsening size. As with GPC, we characterize node similarity using the distortion induced by merging node pairs. The goal then is to group nodes with similar induced node pair merging distortion within the same partition. Towards this, we construct a matrix $H \in \mathbb{R}_+^{|V| \times |V|}$ where $H_{ij} = \mathrm{dis}(\pi^{ij})$, where π^{ij} is the transport plan merging nodes v_i and v_j. Equipped with H and assuming $\mu = \mathbf{1}_N/N$ we can solve for the assignment matrix C_p^* by

$$C_p^* = \underset{C_p \in \mathcal{C}_{N,M}}{\arg\min} \|H - C_p C_w^\top H C_w C_p^\top\|_F^2. \tag{9}$$

We have observed this derived representation to be especially useful for finding node partitions using the k-means algorithm. Moreover, while there are currently no theoretical guarantees that this method minimizes (6) in the general case, the algorithm's performance seems to indicate that the cost function, Eq. (9), can be upper-bounded by the first ordered differences comprising H.

Algorithm 1 Greedy Pair Coarsening (GPC)

function GPC($(S, \mu), M$)
  $(N, t) \leftarrow (|\mathcal{V}|, 0)$
  $(S_t, \mu_t) \leftarrow (S, \mu)$
  while $t < N - M$ **do**
    $D_{ij} \leftarrow \text{dis}^2(\pi^{ij})$
    $C_p \leftarrow \arg\min_{v_i, v_j} D_{ij}$
    $C_w \leftarrow \text{diag}(\mu) C_p \text{diag}(1/(C_p \mu))$
    $S_t, \mu_t \leftarrow (C_w^\top S_t C_w, C_p^\top \mu)$
    $t \leftarrow t + 1$
  end while
  return (S_t, μ_t)
end function

Algorithm 2 k-means Greedy Pair Coarsening (KGPC)

function KGPC($(S, \mu), M$)
  $H_{ij} \leftarrow \text{dis}(\pi^{ij})$
  $C_p \leftarrow k\text{-means}(H, M)$
  $C_w \leftarrow \text{diag}(\mu) C_p \text{diag}(1/(C_p \mu))$
  $(S', \mu') \leftarrow (C_w^\top S C_w,\ C_p^\top \mu)$
  return $G = (S', \mu')$
end function

4 Numerical Analysis

In this section, we detail the experiments conducted to validate the utility of GPC and KGPC – all experimental code can be found here: https://github.com/ctaveras1999/graph-coarsening. We contextualize the performance of our methods against the Multi-level Graph Coarsening (MGC) and Spectral Graph Coarsening (SGC) algorithms proposed in [25] and the Kernel Graph Coarsening (KGC) algorithm proposed in [24]. Note that [24] is, to our knowledge, the only other coarsening method that explicitly minimizes a Gromov-Wasserstein distance, though their method requires the use of the unsigned Laplacian representation. Our method is flexible to graph representation, but we choose the adjacency matrix. Changing the graph representation will affect the values of induced distortion; representation can therefore be seen as a hyperparameter for our coarsening method.

We perform two experiments: 1) we compare the reconstruction error of the different methods to quantify coarsening quality, and 2) we leverage the GW dictionary learning method proposed in [47] for graph classification. Throughout these experiments, we use several well-established graph datasets, namely the IMDB-Binary [54], Mutag [55], Proteins [56], Enzymes [56, 57], MSRC [58], and Tumblr [59] datasets.

Quantifying Reconstruction Error: The goal of this experiment is to determine which of the aforementioned methods best preserves the structure of the adjacency matrix as measured by the coarsening-induced distortion. For each graph in a given dataset, we coarsen at various levels (15% and 85% of all nodes in 5% increments), then compute the distortion between the graphs and their coarsenings. At each coarsening level, we compute the distortion induced by coarsening and average over all such distortions. This average distortion is then treated as a measure of coarsening quality as a function of coarsening level. The results of this experiment for the MSRC dataset are summarized in Figure 1. In it, we can see that for MSRC and Enzymes, our algorithms achieve the lowest, or near-lowest, distortion across coarsening levels. As the coarsening level increases, the difference between the methods becomes less pronounced, leading to little meaningful difference between the methods. This suggests that the graphs cannot be well-approximated by graphs with 80%+ of nodes coarsened.

Graph Classification via Clustering: For this experiment, we leverage the Graph Dictionary Learning method (FGWF) proposed in [47] for unsupervised graph classification. Given a set of graphs, $\mathcal{G} = [G_i]_{i=1}^N$, the objective of this method is to learn a set of atoms $[B_j]_{j=1}^M$ and weights $[\lambda_i]_{i=1}^N$ such that barycenters formed by the dictionary atoms can well approximate graphs in the dataset. We initialize the dictionary with 15 graphs randomly sampled from the dataset and randomly initialize the weights λ_i associated with each graph in the dictionary. Model parameters are updated using the Adam optimizer. Individual FGWF models were trained for 15 epochs with a learning rate of 0.01. After training, we classify graphs by clustering their associated weight values using k-means.

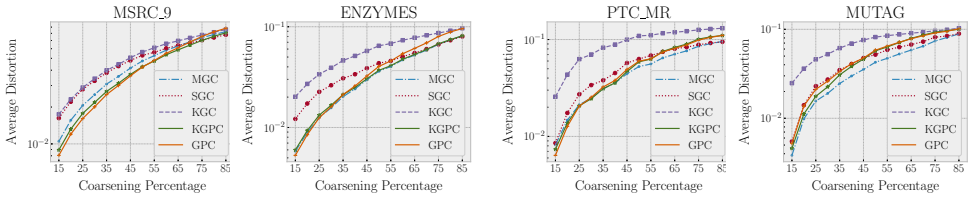

Figure 1: For each method and graph in a dataset, we coarsen between 15% to 85% of nodes, and, for each coarsening level, average the distortion over all graphs. GPC and KGPC achieve the overall lowest or near-lowest distortion on MSRC, Enzymes, and PTC-MR. MGC performs best on MUTAG.

The goal of this experiment is to determine how faithfully the different coarsening methods represent the original data. Towards this, we coarsen all data to 40% of their original size, and for each combination of coarsening method and dataset, we train separate FGWF models and compute the Rand Index of the respective classification results. Table 4 reports the average and standard deviation of the Rand Index for each combination of dataset and algorithm, and indicates that GPC and KGPC are, overall, the most compatible with the FGWF model [47] for graph classification. The Rand index for our algorithms was on par with or better than the ones produced by the original graph. This may indicate that the coarsening step helps remove extraneous structure in the graph that does not aid in classification.

Methods\Datasets	IMDB-B	MUTAG	Proteins	MSRC-9	ENZYMES	PTC-MR
MGC [25]	50.7 ± 0.2	50.1 ± 1.4	54.3 ± 3.4	$\mathbf{78.0 \pm 0.3}$	$\mathbf{72.5 \pm 0.2}$	50.7 ± 0.2
SGC [25]	50.5 ± 0.2	50.3 ± 0.6	52.8 ± 3.6	77.9 ± 0.3	72.4 ± 0.1	50.5 ± 0.2
KGC [24]	50.0 ± 0.0	51.1 ± 1.4	$\mathbf{58.1 \pm 2.0}$	77.9 ± 0.2	72.4 ± 0.2	50.0 ± 0.0
GPC (Ours)	$\mathbf{51.3 \pm 0.3}$	51.0 ± 1.6	55.7 ± 2.3	77.9 ± 0.4	72.3 ± 0.1	$\mathbf{51.3 \pm 0.2}$
KGPC (Ours)	50.8 ± 0.3	$\mathbf{53.6 \pm 2.5}$	56.7 ± 2.1	$\mathbf{78.0 \pm 0.4}$	72.3 ± 0.1	50.8 ± 0.3
Original	51.1 ± 0.8	50.6 ± 0.6	54.9 ± 2.3	77.9 ± 0.2	72.3 ± 0.1	51.1 ± 0.8

Table 1: Rand Index for coarsened graphs with 60% of nodes coarsened. For each combination of dataset and method, we coarsen the data, which we then use to train four FGWF models [47] for graph classification. We use the Rand Index to measure the quality of the classification results and find that for most datasets, either GPC or KGPC performs best.

5 Discussion

In this work, we proposed the GPC and KGPC algorithms for graph coarsening with respect to the Gromov-Wasserstein distance. We conducted two experiments, where the results show that our methods produce coarsening with less distortion and better discriminability for classification tasks. These results suggest that our methods produce coarsenings that better leverage the structure provided by the GW geometry than others and are more compatible with GW-based methods.

Future research should explore establishing an upper bound on the coarsening objective in terms of the distortion of first-order coarsening to supplement the development of the KGPC algorithm and exploring trade-offs between graph representations on the proposed algorithms. Other directions that this work opens up include incorporating graph features into the coarsening, similar to [60] for the Fused Gromov-Wasserstein distance [61], and providing guarantees for the output of GNNs that use the Gromov-Wasserstein distance, similar to the message passage guarantees derived in [62].

References

[1] T. Kipf, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[2] T. S. Cohen and M. Welling, "Group equivariant convolutional networks," 2016.

[3] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," 2019.

[4] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," 2019.

[5] T. M. Roddenberry, N. Glaze, and S. Segarra, "Principled simplicial neural networks for trajectory prediction," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 9020–9029, PMLR, 18–24 Jul 2021.

[6] C. Bodnar, F. Frasca, N. Otter, Y. G. Wang, P. Liò, G. Montúfar, and M. Bronstein, "Weisfeiler and lehman go cellular: Cw networks," 2022.

[7] M. Hajij, G. Zamzmi, T. Papamarkou, N. Miolane, A. Guzmán-Sáenz, K. N. Ramamurthy, T. Birdal, T. K. Dey, S. Mukherjee, S. N. Samaga, N. Livesay, R. Walters, P. Rosen, and M. T. Schaub, "Topological deep learning: Going beyond graph data," 2023.

[8] C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub, "What are higher-order networks?," *SIAM review*, vol. 65, no. 3, pp. 686–731, 2023.

[9] C. Battiloro, E. Karaismailoğlu, M. Tec, G. Dasoulas, M. Audirac, and F. Dominici, "E(n) equivariant topological neural networks," 2025.

[10] D. Fuchsgruber, T. Poštuvan, S. Günnemann, and S. Geisler, "Graph neural networks for edge signals: Orientation equivariance and invariance," 2025.

[11] M. Puschel and J. M. Moura, "Algebraic signal processing theory: Foundation and 1-d time," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3572–3585, 2008.

[12] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[13] A. Petersen and H.-G. Müller, "Fréchet regression for random objects with Euclidean predictors," *The Annals of Statistics*, vol. 47, no. 2, pp. pp. 691–719, 2019.

[14] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[15] N. Guigui, N. Miolane, and X. Pennec, "Introduction to riemannian geometry and geometric statistics: From basic theory to implementation with geomstats," *Foundations and Trends® in Machine Learning*, vol. 16, no. 3, pp. 329–493, 2023.

[16] G. Leus, A. G. Marques, J. M. Moura, A. Ortega, and D. I. Shuman, "Graph signal processing: History, development, impact, and outlook," *IEEE Signal Processing Magazine*, vol. 40, p. 49–60, June 2023.

[17] M. Papillon, S. Sanborn, J. Mathe, L. Cornelis, A. Bertics, D. Buracas, H. J Lillemark, C. Shewmake, F. Dinc, X. Pennec, and N. Miolane, "Beyond euclid: an illustrated guide to modern machine learning with geometric, topological, and algebraic structures," *Machine Learning: Science and Technology*, vol. 6, p. 031002, Aug. 2025.

[18] C. Wang, G. A. Kumar, and J. C. Rajapakse, "Drug discovery and mechanism prediction with explainable graph neural networks," *Scientific Reports*, vol. 15, no. 1, p. 179, 2025.

[19] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," 2019.

[20] J. Wang, S. Zhang, Y. Xiao, and R. Song, "A review on graph neural network methods in financial applications," *Journal of Data Science*, vol. 20, no. 2, pp. 111–134, 2022.

[21] R. Olshevskyi, Z. Zhao, K. Chan, G. Verma, A. Swami, and S. Segarra, "Fully distributed online training of graph neural networks in networked systems," *arXiv preprint arXiv:2412.06105*, 2024.

[22] M. Hashemi, S. Gong, J. Ni, W. Fan, B. A. Prakash, and W. Jin, "A comprehensive survey on graph reduction: Sparsification, coarsening, and condensation," in *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, (Jeju, South Korea), p. 8058–8066, International Joint Conferences on Artificial Intelligence Organization, Aug. 2024.

[23] J. Chen, Y. Saad, and Z. Zhang, "Graph coarsening: from scientific computing to machine learning," *SeMA Journal*, vol. 79, no. 1, pp. 187–223, 2022.

[24] Y. Chen, R. Yao, Y. Yang, and J. Chen, "A gromov-wasserstein geometric view of spectrum-preserving graph coarsening," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org, 2023.

[25] Y. Jin, A. Loukas, and J. JaJa, "Graph coarsening with preserved spectral properties," in *International Conference on Artificial Intelligence and Statistics*, pp. 4452–4462, PMLR, 2020.

[26] G. Bravo Hermsdorff and L. Gunderson, "A unifying framework for spectrum-preserving graph sparsification and coarsening," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] Y. Chen, H. Ye, S. Vedula, A. Bronstein, R. Dreslinski, T. Mudge, and N. Talati, "Demystifying graph sparsification algorithms in graph properties preservation," 2023.

[28] X. Gao, J. Yu, T. Chen, G. Ye, W. Zhang, and H. Yin, "Graph condensation: A survey," 2025.

[29] G. Kron, *Tensor analysis of networks*. J. Wiley & Sons New York, 1939.

[30] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "Harp: Hierarchical representation learning for networks," 2017.

[31] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[32] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *Advances in neural information processing systems*, vol. 31, 2018.

[33] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, "Graph summarization methods and applications: A survey," *ACM Comput. Surv.*, vol. 51, June 2018.

[34] A. Loukas and P. Vandergheynst, "Spectrally approximating large graphs with smaller graphs," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 3237–3246, PMLR, 10–15 Jul 2018.

[35] A. Loukas, "Graph reduction with spectral and cut guarantees," *Journal of Machine Learning Research*, vol. 20, no. 116, pp. 1–42, 2019.

[36] F. R. Chung, *Spectral graph theory*, vol. 92. American Mathematical Soc., 1997.

[37] E. R. van Dam and W. H. Haemers, "Which graphs are determined by their spectrum?," *Linear Algebra and its Applications*, vol. 373, pp. 241–272, 2003. Combinatorial Matrix Theory Conference (Pohang, 2002).

[38] K.-T. Sturm, "On the geometry of metric measure spaces," *Acta Mathematica*, vol. 196, no. 1, p. 65–131, 2006.

[39] F. Mémoli, "Gromov-wasserstein distances and the metric approach to object matching," *Foundations of Computational Mathematics*, vol. 11, pp. 417–487, Aug. 2011.

[40] T. Gervens and M. Grohe, "Graph similarity based on matrix norms," 2022.

[41] R. Bhatia, T. Jain, and Y. Lim, "On the bures-wasserstein distance between positive definite matrices," 2017.

[42] I. Haasler and P. Frossard, "Bures-wasserstein means of graphs," 2024.

[43] H. Xu, D. Luo, H. Zha, and L. C. Duke, "Gromov-wasserstein learning for graph matching and node embedding," in *International conference on machine learning*, pp. 6932–6941, PMLR, 2019.

[44] H. Xu, D. Luo, and L. Carin, "Scalable gromov-wasserstein learning for graph partitioning and matching," *Advances in neural information processing systems*, vol. 32, 2019.

[45] L. Zheng, Y. Xiao, and L. Niu, "A brief survey on computational gromov-wasserstein distance," *Procedia Computer Science*, vol. 199, pp. 697–702, 2022.

[46] Z. Zeng, R. Qiu, Z. Xu, Z. Liu, Y. Yan, T. Wei, L. Ying, J. He, and H. Tong, "Graph mixup on approximate gromov–wasserstein geodesics," in *Forty-first International Conference on Machine Learning*, 2024.

[47] H. Xu, "Gromov-wasserstein factorization models for graph clustering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, p. 6478–6485, Apr. 2020.

[48] Y. Xiang, D. Luo, and H. Xu, "Privacy-preserved evolutionary graph modeling via gromov-wasserstein autoregression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37(12), pp. 14566–14574, 2023.

[49] L. Brogat-Motte, R. Flamary, C. Brouard, J. Rousu, and F. d'Alché Buc, "Learning to predict graphs with fused gromov-wasserstein barycenters," in *International Conference on Machine Learning*, pp. 2321–2335, PMLR, 2022.

[50] S. Chowdhury and F. Mémoli, "The gromov-wasserstein distance between networks and stable network invariants," *Information and Inference: A Journal of the IMA*, vol. 8, p. 757–787, Dec. 2019.

[51] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-wasserstein averaging of kernel and distance matrices," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 2664–2672, PMLR, 20–22 Jun 2016.

[52] C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer, and N. Courty, "Semi-relaxed Gromov Wasserstein divergence with applications on graphs," in *ICLR 2022 - 10th International Conference on Learning Representations*, (Virtual, France), pp. 1–28, Apr. 2022. preprint under review.

[53] F. Mémoli, A. Sidiropoulos, and K. Singhal, "Sketching and clustering metric measure spaces," *CoRR*, vol. abs/1801.00551, 2018.

[54] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (New York, NY, USA), p. 1365–1374, Association for Computing Machinery, 2015.

[55] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch, "Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity," *Journal of Medicinal Chemistry*, vol. 34, no. 2, pp. 786–797, 1991.

[56] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, pp. i47–i56, 06 2005.

[57] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, "Brenda, the enzyme database: updates and major new developments," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D431–D433, 2004.

[58] M. Neumann, R. Garnett, C. Bauckhage, and K. Kersting, "Propagation kernels," 2014.

[59] L. Oettershagen, N. M. Kriege, C. Morris, and P. Mutzel, "Temporal graph kernels for classifying dissemination processes," 2021.

[60] M. Kumar, A. Sharma, S. Saxena, and S. Kumar, "Featured graph coarsening with similarity guarantees," in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 17953–17975, PMLR, 23–29 Jul 2023.

[61] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty, "Fused gromov-wasserstein distance for structured objects," *Algorithms*, vol. 13, no. 9, 2020.

[62] A. Joly and N. Keriven, "Graph coarsening with message-passing guarantees," *Advances in Neural Information Processing Systems*, vol. 37, pp. 114902–114927, 2024.

[63] S. Chowdhury and F. Mémoli, "Distances and isomorphism between networks: Stability and convergence of network invariants," *Journal of Applied and Computational Topology*, vol. 7, p. 243–361, june 2023. arXiv:1708.04727 [cs].

[64] S. Chowdhury and T. Needham, "Gromov-wasserstein averaging in a riemannian framework," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3676–3684, 2020.

A Proofs Omitted from the Main Text

A.1 Proof of Proposition 1

If there exists a pair of nodes for which the induced distortion is equal to zero, then G', the graph resulting from merging said nodes, is weakly isomorphic to G. This follows from the fact that the distortion is an upper-bound for the GW distance (GW evaluates distortion of infimizing coupling), and the GW distance is unique up to weak isomorphism [50]. We repeat this process until there are no node pairs with zero distortion, in which case we have reached a minimal representative. Since the distortion between the original network and a minimal representative is zero, and the distortion is an upper-bound on the GW distance, it follows that the minimal representative solves Problem (6). Note that minimal representatives may not be unique, but all minimal representatives are strongly isomorphic (i.e. unique up to node re-labeling) [63].

A.2 Proof of Proposition 2

Let $V^{(t)} = \{v_1^{(t)}, \cdots, v_{N-t}^{(t)}\}$ denote the partition, or supernode set, constructed after t iterations of GPC. The i-th supernode $v_i^{(t)} \in V^{(t)}$ contains a set of nodes and we denote its size by $N_i^{(t)} = |v_i^{(t)}|$. We call a pair of supernodes $v_i^{(t)}, v_j^{(t)} \in V^{(t)}$ consistent if $v_i^{(t)} \cup v_j^{(t)} \subseteq P$ for some $P \in \mathcal{P}$. We proceed by induction towards showing that GPC only merges consistent pairs of supernodes for $1 \leq t < N - M$. This will then imply that after $N - M$ iterations we achieve $V^{(N-M)} = \mathcal{P}$.

Before the first iteration, the partition $V^{(0)}$ consists of singleton supernodes $v_i^{(0)} = \{v_i\}$. The first iteration of GPC thus merges the pair of nodes v_i, v_j, inducing the least distortion. By hypothesis and Lemma 1, the distortion induced by the merging of nodes in the same partition set is upper-bounded by ϵ, and lower-bounded by $\alpha\epsilon$ for nodes in different partition sets. Therefore, GPC must merge consistent nodes at the first iteration.

Suppose now that GPC has run for $1 \leq t < N-M-1$ iterations during which only consistent supernodes were merged, to produce the supernode set $V^{(t)} = \{v_1^{(t)}, \cdots, v_{N-t}^{(t)}\}$. Let $v_i^{(t)}, v_j^{(t)}, v_k^{(t)} \in V^{(t)}$ be such that $v_i^{(t)}$ and $v_j^{(t)}$ are consistent, and $v_i^{(t)}$ and $v_k^{(t)}$ are not. We want to show that the distortion induced by merging the former pair is strictly less than that of the latter. Lemma 2 allows us to compute upper and lower bounds on $D_{ij} = \text{dis}(\pi^{ij})$ and $D_{ik} = \text{dis}(\pi^{ik})$, respectively, where π^{rs} is the transport map merging supernodes $v_r^{(t)}$ and $v_s^{(t)}$. By Lemma 3, we have that $\max D_{ij} < \min D_{ik}$, as long as $\alpha > (4 + 4N/\sqrt{N-1})$. Therefore, at iteration t, GPC will merge the pair of consistent nodes with the least distortion.

After $N - M$ iterations, no pair of supernodes in $V^{(N-M)}$ is consistent. This implies that the supernodes in $V^{(N-M)}$ correspond exactly with the sets in the partition $\mathcal{P}$, thus $V^{(N-M)} = \mathcal{P}$, as desired.

Lemma 1. *Let $G = (V, w, \mu)$ be a symmetric measure network and π^{12} the transport map induced by merging nodes $v_1, v_2 \in V$. Then, dis$^2(\pi^{12}) = \mu_1\mu_2(A_1 + 2A_2)/(\mu_1 + \mu_2)^4$ where $\Delta_{klmn} = (s_{kl} - s_{mn})$ and*

$$\begin{aligned} A_1 &= \mu_1^3\mu_2(4\Delta_{1112}^2 + \Delta_{2211}^2) + \mu_1\mu_2^3(\Delta_{1122}^2 + 4\Delta_{2212}^2) + 2(\Delta_{1211}^2\mu_1^4 + \Delta_{1222}^2\mu_2^4) \\ &\quad + 4\mu_1^2\mu_2^2[(|\Delta_{1112}| - |\Delta_{2212}|)^2 + 2|\Delta_{1112}||\Delta_{2212}| + \Delta_{1112}\Delta_{2212}] \\ A_2 &= (\mu_1 + \mu_2)^3 \sum_{n=3}^{N} \mu_n \Delta_{1n2n}^2. \end{aligned}$$

Proof. We can represent the assignment matrix merging nodes v_1 and v_2 by

$$C_p^{12} = \begin{bmatrix} 1_{2\times 1} & 0_{2\times(N-2)} \\ 0_{(N-2)\times 1} & I_{N-2} \end{bmatrix},$$

where $1_{n_1\times n_2}$ (resp. $1_{n_1\times n_2}$) is the ones (resp. zeros) matrix n_1 rows and n_2 columns and I_n is the identity matrix with n rows and columns. Then, letting $\pi = \pi^{12}$ and $C_p = C_p^{12}$, we get

$$\pi = \text{diag}(\mu)C_p,\ C_w = \pi\text{diag}\left(1/C_p\mu\right),\ \text{and } S' = (V', C_w^\top S C_w, C_p\mu).$$

Carrying out the multiplications, we get

$$w' = \begin{bmatrix} \sum_{i,j=1}^{2} \theta_i \theta_j s_{ij} & \sum_{i=1}^{2} \theta_i s_{i3} & \cdots & \sum_{i=1}^{2} \theta_i s_{iN} \\ \sum_{j=1}^{2} \theta_i s_{3j} & s_{33} & \cdots & s_{3N} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{2} \theta_j s_{Nj} & s_{N3} & \cdots & s_{NN} \end{bmatrix} \tag{10}$$

where $\theta_1 = \mu_1/(\mu_1+\mu_2)$ and $\theta_2 = \mu_2/(\mu_1+\mu_2)$. We now start computing $\text{dis}^2(\pi)$ of the coupling. Let $d_{ijkl} = |s_{ij} - s'_{kl}|^2 \pi_{ik}\pi_{jl}$. Then,

$$\begin{aligned} \text{dis}^2(\pi) &= \sum_{i,j=1}^{N} \sum_{k,l=1}^{N-1} d_{ijkl} \\ &= \sum_{i,j=1}^{N} \left[d_{ij11} + \sum_{k=2}^{N-1} d_{ijk1} + \sum_{l=2}^{N-1} d_{ij1l} + \sum_{k,l=2}^{N-1} d_{ijkl} \right] \end{aligned}$$

Let

$$D_1 = \sum_{i,j=1}^{N} d_{ij11} \qquad D_2 = \sum_{i,j=1}^{N} \sum_{k=2}^{N-1} d_{ijk1} \qquad D_3 = \sum_{i,j=1}^{N} \sum_{l=2}^{N-1} d_{ij1l} \qquad D_4 = \sum_{i,j=1}^{N} \sum_{k,l=2}^{N-1} d_{ijkl}$$

and $\Delta_{ijkl} = (s_{ij} - s_{kl})$. We proceed with a term-by-term expansion

$$\begin{aligned} D_1 &= d_{1111} + d_{1211} + d_{2111} + d_{2211} \\ &= |s_{11} - s'_{11}|^2 \pi_{11}\pi_{11} + |s_{12} - s'_{11}|^2 \pi_{11}\pi_{21} + |s_{21} - s'_{11}|^2 \pi_{21}\pi_{11} + |s_{22} - s'_{11}|^2 \pi_{21}\pi_{21} \\ &= |s_{11} - s'_{11}|^2 \mu_1^2 + |s_{12} - s'_{11}|^2 \mu_1\mu_2 + |s_{21} - s'_{11}|^2 \mu_2\mu_1 + |s_{22} - s'_{11}|^2 \mu_2^2 \end{aligned}$$

Let $D_{11} = |s_{11} - s'_{11}|^2 \mu_1^2$, $D_{12} = |s_{12} - s'_{11}|^2 \mu_1\mu_2$, $D_{13} = |s_{21} - s'_{11}|^2 \mu_2\mu_1$, and $D_{14} = |s_{22} - s'_{11}|^2 \mu_2^2$; note that by the symmetry of w we have $D_{12} = D_{13}$. Then,

$$\begin{aligned} D_{11} &= \mu_1^2 |s_{11} - s'_{11}|^2 \\ &= \mu_1^2 \left| \sum_{i,j=1}^{2} (s_{11} - s_{ij}) \theta_i \theta_j \right|^2 = \mu_1^2 |2\Delta_{1112}\theta_1\theta_2 + \Delta_{1122}\theta_2^2|^2 \\ &= \mu_1^2 \theta_2^2 (4\Delta_{1112}^2 \theta_1^2 + \Delta_{1122}^2 \theta_2^2 + 4\Delta_{1112}\Delta_{1122}\theta_1\theta_2) \\ &= \frac{\mu_1^2\mu_2^2}{(\mu_1+\mu_2)^4} \left(4\Delta_{1112}^2 \mu_1^2 + \Delta_{1122}^2 \mu_2^2 + 4\Delta_{1112}\Delta_{1122}\mu_1\mu_2 \right). \end{aligned}$$

Similar computations yield

$$\begin{aligned} D_{12} = D_{13} &= \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^4} (\Delta_{1211}^2 \mu_1^4 + \Delta_{1222}^2 \mu_2^4 + 2\Delta_{1211}\Delta_{1222}\mu_1^2\mu_2^2) \\ D_{14} &= \frac{\mu_2^2\mu_1^2}{(\mu_1+\mu_2)^4} (\Delta_{1122}^2 \mu_1^2 + 4\Delta_{1222}\mu_2^2 + 4\Delta_{1122}\Delta_{1222}\mu_1\mu_2) \end{aligned}$$

Combining these terms, we get

$$\begin{aligned} D_1 \frac{(\mu_1+\mu_2)^4}{\mu_1\mu_2} &= \mu_1^3\mu_2(4\Delta_{1112}^2 + \Delta_{1122}^2) + \mu_1\mu_2^3(\Delta_{1112}^2 + \Delta_{1222}^2) + 2(\Delta_{1211}^2\mu_1^4 + \Delta_{1222}^2\mu_2^4) \\ &\quad + 4\mu_1^2\mu_2^2(\Delta_{1112}\Delta_{1122} + \Delta_{1211}\Delta_{1222} + \Delta_{1122}\Delta_{1222}) \\ &= \mu_1^3\mu_2(4\Delta_{1112}^2 + \Delta_{1122}^2) + \mu_1\mu_2^3(\Delta_{1112}^2 + \Delta_{1222}^2) + 2(\Delta_{1211}^2\mu_1^4 + \Delta_{1222}^2\mu_2^4) \\ &\quad + 4\mu_1^2\mu_2^2[(|\Delta_{1112}| - |\Delta_{1222}|)^2 + 2|\Delta_{1112}||\Delta_{1222}| + \Delta_{1112}\Delta_{1222}] \end{aligned}$$

We proceed with D_2, noting that $D_3 = D_2$ by the symmetry of w,

$$\begin{aligned}
D_2 &= \sum_{i,j=1}^{N}\sum_{k=2}^{N-1} d_{ijk1} = \sum_{i,j=1}^{N}\sum_{k=2}^{N-1} |s_{ij} - s'_{k1}|^2 \pi_{ik}\pi_{j1} \\
&= \sum_{i=3}^{N} \mu_i \left(\left| s_{i1} - \sum_{l=1}^{2} \theta_l s_{il} \right|^2 \mu_1 + \left| s_{i2} - \sum_{l=1}^{2} \theta_l s_{l2} \right|^2 \mu_2 \right) \\
&= \sum_{i=3}^{N} \mu_i \left(|\theta_2(s_{i1} - s_{i2})|^2 \mu_1 + |\theta_1(s_{i2} - s_{i1})|^2 \mu_2 \right) \\
&= \sum_{i=3}^{N} \mu_i \left(|s_{i1} - s_{i2}|^2 \mu_1 \theta_2^2 + |s_{i2} - s_{i1}|^2 \theta_1^2 \mu_2 \right) \\
&= \frac{1}{(\mu_1+\mu_2)^2} \sum_{i=3}^{N} \mu_i \left(|s_{i1} - s_{i2}|^2 \mu_1 \mu_2^2 + |s_{i2} - s_{i1}|^2 \mu_1^2 \mu_2 \right) \\
&= \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2} \sum_{i=3}^{N} \mu_i \left(|s_{i1} - s_{i2}|^2 \mu_2 + |s_{i2} - s_{i1}|^2 \mu_1 \right) \\
&= \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2} \sum_{i=3}^{N} \mu_i |s_{i1} - s_{i2}|^2 (\mu_1+\mu_2) \\
&= \frac{\mu_1\mu_2}{\mu_1+\mu_2} \sum_{i=3}^{N} \mu_i |s_{i1} - s_{i2}|^2 \\
&= \frac{\mu_1\mu_2}{\mu_1+\mu_2} \sum_{i=3}^{N} \mu_i |\Delta_{i1i2}|^2.
\end{aligned}$$

Finally, we compute D_4,

$$\begin{aligned}
D_4 &= \sum_{i,j=1}^{N}\sum_{k,l=2}^{N-1} d_{ijkl} \\
&= \sum_{i,j=3}^{N}\sum_{k,l=2}^{N-1} |s_{ij} - s'_{kl}|^2 \pi_{ik}\pi_{jl} \\
&= \sum_{i,j=3}^{N} |s_{ij} - s_{ij}| \mu_i \mu_j = 0
\end{aligned}$$

Combining these terms, we get

$$\begin{aligned}
\frac{(\mu_1+\mu_2)^4}{\mu_1\mu_2} D &= 2(\mu_1+\mu_2)^3 \sum_{i=3}^{N} \mu_i |\Delta_{i1i2}|^2 \\
&\quad + \mu_1^3\mu_2(4\Delta_{1112}^2 + \Delta_{1122}^2) + \mu_1\mu_2^3(\Delta_{1112}^2 + \Delta_{1222}^2) + 2(\Delta_{1211}^2\mu_1^4 + \Delta_{1222}^2\mu_2^4) \\
&\quad + 4\mu_1^2\mu_2^2[(|\Delta_{1112}| - |\Delta_{1222}|)^2 + 2|\Delta_{1112}||\Delta_{1222}| + \Delta_{1112}\Delta_{1222}]
\end{aligned}$$

□

Lemma 2. *Let $G = (V, w, \mu)$ satisfy the hypotheses in Proposition 1, with partition $\mathcal{P} = \{P_1, \cdots, P_M\}$, $\epsilon > 0$, and $\alpha > 4$. Let G' be the coarsening of G induced by C_p, and $\phi : V \to V'$ the mapping corresponding to C_p. Suppose there exist nodes $v'_i, v'_j, v'_k \in V'$ satisfying $\|w(u_1) - w(u_2)\|_\infty < \epsilon$ for all $u_1, u_2 \in \phi^{-1}(v'_i) \cup \phi^{-1}(v'_j)$ and $\inf_{z\in V} |w(u_1, z) - w(u_3, z)| > \alpha\epsilon$ for $u_1 \in \phi^{-1}(v'_i)$ and $u_3 \in \phi^{-1}(v'_k)$. Then,*

$$\textit{dis}^2(\pi^{ij}) < \frac{32\epsilon^2\mu'_1\mu'_2}{\mu'_1+\mu'_2} \textit{ and } \textit{dis}^2(\pi^{ik}) \geq \frac{\epsilon^2(\alpha-4)^2\mu'_1\mu'_3}{2(\mu'_1+\mu'_3)}.$$

Proof. Without loss of generality, we let $v'_i = v'_1$ and $v'_j = v'_2$. To prove this lemma, we must compute upper and lower bounds on several terms of the form $|\Delta'_{ijkl})$ where $\Delta'_{ijkl} = s'_{ij} - s'_{kl}$. Note that by the symmetry of w' we have $|\Delta'_{ijkl}| = |\Delta'_{jikl}| = |\Delta'_{ijlk}| = |\Delta'_{jilk}|$.

We first compute upper bounds on weight differences of G and G', which we use in later computations. Without loss of generality we order the nodes in V such that $\phi^{-1}(v'_l) = \left[v_{\hat{N}_{l-1}+1}, \cdots, v_{\hat{N}_{l-1}+N_l}\right]$, where $N_0 = 0$, $N_l = \left|\phi^{-1}(v'_l)\right|$, and $\hat{N}_l = \sum_{r=0}^{l} N_r$. Moreover, given a supernode v'_l, we define $[\theta_m]_{m=1}^{N_l}$ where $\theta_m = \mu\left(v_{\hat{N}_l+m}\right)/\mu(v'_l)$. Finally, we abuse notation below, using $m \in \phi^{-1}(v'_l)$ in place of $v_m \in \phi^{-1}(v'_l)$.

$$
\begin{aligned}
|s'_{11} - s_{11}| &= \left| \sum_{l,m \in \phi^{-1}(v'_1)} \theta_l \theta_m (s_{lm} - s_{11}) \right| \\
&\leq \left| \sum_{l,m \in \phi^{-1}(v'_1)} \theta_l \theta_m |s_{lm} - s_{11}| \right| \\
&< 2\epsilon \left| \sum_{l,m \in \phi^{-1}(v'_1)} \theta_l \theta_m \right| = 2\epsilon \\
|s'_{12} - s_{11}| &= \left| \sum_{l \in \phi^{-1}(v'_1)} \sum_{m \in \phi^{-1}(v'_2)} \theta_l \theta_m (s_{kl} - s_{11}) \right| \\
&\leq \left| \sum_{l \in \phi^{-1}(v'_1)} \sum_{m \in \phi^{-1}(v'_2)} \theta_l \theta_m |s_{lm} - s_{11}| \right| \\
&< 2\epsilon \left| \sum_{l \in \phi^{-1}(v'_1)} \sum_{m \in \phi^{-1}(v'_2)} \theta_l \theta_m \right| = 2\epsilon \\
|s'_{22} - s_{11}| &= |s'_{22} - s_{1,N_1+1} + s_{1,N_1+1} - s_{11}| \\
&\leq |s'_{22} - s_{1,N_1+1}| + |s_{1,N_1+1} - s_{11}| \\
&< 2\epsilon + 2\epsilon = 4\epsilon \\
|s'_{1n} - s'_{2n}| &= |s'_{1n} - s_{1,\hat{N}_{n-1}+1} + s_{1,\hat{N}_{n-1}+1} - s'_{2n}| \\
&\leq |s'_{1n} - s_{1,\hat{N}_{n-1}+1}| + |s_{\hat{N}_{n-1}+1,\hat{N}_{n-1}+1} - s'_{2n}| \\
&\leq 2\epsilon + 2\epsilon \\
&< 4\epsilon
\end{aligned}
$$

We can now produce bounds for the relevant terms in the distortion

$$
\begin{aligned}
|\Delta'_{1112}| &= |s'_{11} - s'_{12}| \\
&= |s'_{11} - s_{11} + s_{11} - s'_{12}| \\
&\leq |s'_{11} - s_{11}| + |s'_{12} - s_{11}| \\
&< 2\epsilon + 2\epsilon = 4\epsilon
\end{aligned}
$$

$$
\begin{aligned}
|\Delta'_{1222}| &= |s'_{12} - s'_{22}| \\
&= |s'_{12} - s_{\hat{N}_1+1,\hat{N}_1+1} + s_{\hat{N}_1+1,\hat{N}_1+1} - s'_{22}| \\
&\leq |s'_{12} - s_{\hat{N}_1+1,\hat{N}_1+1}| + |s'_{22} - s_{\hat{N}_1+1,\hat{N}_1+1}| \\
&< 2\epsilon + 2\epsilon = 4\epsilon
\end{aligned}
$$

$$
\begin{aligned}
|\Delta'_{1122}| &= |s'_{11} - s'_{22}| \\
&\leq |s'_{11} - s'_{12}| + |s'_{12} - s'_{22}| \\
&= |\Delta'_{1112}| + |\Delta'_{1222}| \\
&< 4\epsilon + 4\epsilon = 8\epsilon
\end{aligned}
$$

$$
\begin{aligned}
|\Delta'_{1n2n}| &= |s'_{1n} - s'_{2n}| \\
&< 4\epsilon
\end{aligned}
$$

$$
\begin{aligned}
||\Delta'_{1112}| - |\Delta'_{1222}|| &\leq |\Delta'_{1112} - \Delta'_{1222}| \\
&= |s'_{11} - s'_{12} + s'_{12} - s'_{22}| \\
&= |\Delta'_{1122}| \\
&< 8\epsilon
\end{aligned}
$$

$$
\begin{aligned}
2|\Delta'_{1112}||\Delta'_{1222}| + \Delta'_{1112}\Delta_{1222} &\leq 3|\Delta'_{1112}||\Delta'_{1222}| \\
&< 48\epsilon^2
\end{aligned}
$$

In summary

$$
\begin{aligned}
|\Delta'_{1112}|, |\Delta'_{1222}|, |\Delta'_{2212}|, |\Delta'_{1n2n}| &< 4\epsilon \\
|\Delta'_{1122}|, |\Delta'_{2211}|, ||\Delta'_{1112}| - |\Delta'_{2212}|| &< 8\epsilon \\
2|\Delta'_{1112}||\Delta'_{2212}| + \Delta'_{1112}\Delta'_{2212} &< 48\epsilon^2
\end{aligned}
$$

We now proceed with computing lower bounds. Without loss of generality, we now let $v'_i = v'_1$, $v'_l = v'_2$.

$$
\begin{aligned}
\alpha\epsilon &\leq |s_{11} - s_{1,N_1+1}| \\
&= |s_{11} - s'_{11} + s'_{11} - s'_{12} + s'_{12} - s_{1,N_1+1}| \\
&\leq |s_{11} - s'_{11}| + |s'_{11} - s'_{12}| + |s'_{12} - s_{1,N_1+1}| \\
&< 2\epsilon + |s'_{11} - s'_{12}| + 2\epsilon \\
\implies |\Delta'_{1112}| &\geq (\alpha - 4)\epsilon
\end{aligned}
$$

$$
\begin{aligned}
\alpha\epsilon &< |s_{1,\hat{N}_n+1} - s_{N_1+1,\hat{N}_n+1}| \\
&= |s_{1,\hat{N}_n+1} - s'_{1n} + s'_{1n} - s'_{2n} + s'_{2n} - s_{N_1+1,\hat{N}_n+1}| \\
&\leq |s_{1,\hat{N}_n+1} - s'_{1n}| + |s'_{1n} - s'_{2n}| + |s'_{2n} - s_{N_1+1,\hat{N}_n+1}| \\
&< 2\epsilon + |s'_{1n} - s'_{2n}| + 2\epsilon \\
\implies |\Delta'_{1n2n}| &> (\alpha - 4)\epsilon
\end{aligned}
$$

$$
\begin{aligned}
2|\Delta'_{1112}||\Delta'_{2212}| + \Delta'_{1112}\Delta'_{2212} &\geq |\Delta'_{1112}||\Delta'_{2212}| \\
&> [(\alpha - 4)\epsilon][(\alpha - 4)\epsilon] \\
&> (\alpha - 4)^2\epsilon^2
\end{aligned}
$$

$$
|\Delta'_{1122}| \geq 0.
$$

Plugging in these bounds to the result from Lemma 1 we get

$$
D_{12} < \frac{\epsilon^2 {\mu'}_1 \mu'_2}{({\mu'}_1 + \mu'_2)^4} \left[\left(2({\mu'}_1 + {\mu'}_2)^3 \sum_{i=3}^{N} {\mu'}_i (4\epsilon)^2 \right) + {\mu'}_1{}^3 {\mu'}_2 [4(4\epsilon)^2 + (8\epsilon)^2)] + {\mu'}_1 {\mu'}_2^3 [(8\epsilon)^2 + 4(4\epsilon)^2] \right.
$$

$$
\begin{aligned}
&+ 2((4\epsilon)^2{\mu'}_1^4 + (4\epsilon)^2{\mu'}_2^4) + 4{\mu'}_1^2{\mu'}_2^2[(8\epsilon)^2 + 48\epsilon^2]\Bigg] \\
&= \frac{\epsilon^2{\mu'}_1{\mu'}_2}{({\mu'}_1+{\mu'}_2)^4}\left[32({\mu'}_1+{\mu'}_2)^3(1-{\mu'}_1-{\mu'}_2) + 64({\mu'}_1^3{\mu'}_2+{\mu'}_1{\mu'}_2^3) + 32({\mu'}_1^4+{\mu'}_2^4) + 112{\mu'}_1^2{\mu'}_2^2\right] \\
&\leq \frac{\epsilon^2{\mu'}_1{\mu'}_2}{({\mu'}_1+{\mu'}_2)^4}[32({\mu'}_1+{\mu'}_2)^3 - 32({\mu'}_1+{\mu'}_2)^4 + 32({\mu'}_1+{\mu'}_2)^4] \\
&= \frac{32\epsilon^2{\mu'}_1{\mu'}_2}{({\mu'}_1+{\mu'}_2)}
\end{aligned}
$$

$$
\begin{aligned}
D_{13} >& \frac{\epsilon^2{\mu'}_1{\mu'}_3}{({\mu'}_1+{\mu'}_3)^4}\Bigg[\left(2({\mu'}_1+{\mu'}_3)^3\sum_{i\neq 1,3}{\mu'}_n(\alpha-4)^2\right) + 4(\alpha-4)^2({\mu'}_1^3{\mu'}_3+{\mu'}_1{\mu'}_3^3) \\
&+8(\alpha-4)^2({\mu'}_1^4+{\mu'}_3^4) + 4(\alpha-4)^2{\mu'}_1^2{\mu'}_3^2]\Bigg] \\
&= \frac{\epsilon^2(\alpha-4)^2{\mu'}_1{\mu'}_3}{({\mu'}_1+{\mu'}_3)^4}\left[2({\mu'}_1+{\mu'}_3)^3(1-{\mu'}_1-{\mu'}_3) + 4({\mu'}_1^3{\mu'}_3+{\mu'}_1{\mu'}_3^3) + 8({\mu'}_1^4+{\mu'}_3^4) + 4{\mu'}_1^2{\mu'}_3^2\right] \\
&> \frac{\epsilon^2(\alpha-4)^2{\mu'}_1{\mu'}_3}{({\mu'}_1+{\mu'}_3)^4}\left[\frac{1}{2}({\mu'}_1+{\mu'}_3)^3(1-{\mu'}_1-{\mu'}_3) + \frac{1}{2}({\mu'}_1+{\mu'}_3)^4\right] \\
&= \frac{\epsilon^2(\alpha-4)^2{\mu'}_1{\mu'}_3}{({\mu'}_1+{\mu'}_3)^4}\left[\frac{1}{2}({\mu'}_1+{\mu'}_3)^3 - \frac{1}{2}({\mu'}_1+{\mu'}_3)^4 + \frac{1}{2}({\mu'}_1+{\mu'}_3)^4\right] \\
&= \frac{\epsilon^2(\alpha-4)^2{\mu'}_1{\mu'}_3}{2({\mu'}_1+{\mu'}_3)},
\end{aligned}
$$

thus concluding the proof. □

Lemma 3. *Let $G = (V, w, \mu)$ be a measure network and $G' = (V', w', \mu')$ be a coarsening of G satisfying the hypotheses of Lemma 2 with $\epsilon > 0$ and $\alpha > 4 + 4N/\sqrt{N-1}$. Then, dis$(\pi^{ij}) <$ dis(π^{ik}).*

Proof. Let

$$
G_{12} = \frac{32\epsilon^2\mu_1\mu_2}{(\mu_1+\mu_2)} \qquad \text{and} \qquad G_{13} = \frac{\frac{1}{2}\epsilon^2(\alpha-4)^2\mu_1\mu_3}{(\mu_1+\mu_3)}.
$$

Then, $\text{dis}^2(\pi^{13}) > \text{dis}^2(\pi^{12})$ if $G_{13} > G_{12}$, or, equivalently,

$$
(\alpha-4)^2 > \frac{32\epsilon^2\mu_1\mu_2}{(\mu_1+\mu_2)} \times \frac{(\mu_1+\mu_3)}{\frac{1}{2}\epsilon^2\mu_1\mu_3} = 64\frac{\mu_2}{\mu_3}\frac{\mu_1+\mu_3}{\mu_1+\mu_2}.
$$

We want to find a lower bound that is valid for any choice of μ_1, μ_2, μ_3, where $\frac{1}{N} \leq \mu_1, \mu_2, \mu_3 \leq 1 - \frac{2}{N}$ and $\mu_1 + \mu_2 + \mu_3 \leq 1$. This bound is largest when $\mu_1 + \mu_2 + \mu_3 = 1$, thus, we let $\mu_2 = 1 - \mu_1 - \mu_3$. We thus want to solve

$$
\max_{\mu_1,\mu_3} G(\mu_1,\mu_3) := 64\frac{1-\mu_1-\mu_3}{\mu_3}\frac{\mu_1+\mu_3}{1-\mu_3} = 64\frac{(\mu_1+\mu_3)-(\mu_1+\mu_3)^2}{\mu_3(1-\mu_3)}.
$$

We first maximize for μ_1:

$$
\begin{aligned}
\frac{\partial}{\partial\mu_1}G(\mu_1,\mu_3) = 64\frac{1-2(\mu_1+\mu_3)}{\mu_3(1-\mu_3)} &= 0 \\
\implies \mu_1^* &= 1/2 - \mu_3.
\end{aligned}
$$

Note that μ_1^* is guaranteed to be a maximizer as $G(\mu_1, \mu_3)$ is a negative quadratic in μ_1. After plugging in $\mu_1^* = 1/2 - \mu_3$ we optimize for μ_3:

$$
\frac{\partial}{\partial\mu_3}G(\mu_1^*,\mu_3) = \frac{\partial}{\partial\mu_3}\left(64\frac{(1/2)-(1/2)^2}{\mu_3(1-\mu_3)}\right) = \frac{\partial}{\partial\mu_3}\left(64\frac{1/4}{\mu_3(1-\mu_3)}\right) \tag{11}
$$

$$= -16\frac{1-2\mu_3}{\mu_3^2(1-\mu_3)^2} = 0 \tag{12}$$

$$\implies \mu_3' = 1/2 \tag{13}$$

It can be shown that μ_3' is a minimizer, not a maximizer. Therefore, the maximizer μ_3^* must occur at the boundary of its domain. By inspection, we get that $\mu_3^* = 1/N$. Taken together, we have $\mu_1^* = 1/2 - 1/N, \mu_2^* = 1/2, \mu_3^* = 1/N$, and

$$G(\mu_1^*, \mu_2^*, \mu_3^*) = 64\frac{\mu_2^*}{\mu_3^*}\frac{\mu_1^* + \mu_3^*}{\mu_1^* + \mu_2^*} = 64\frac{1/2}{1/N}\frac{1/2}{1-1/N} = 16\frac{N^2}{N-1}.$$

Therefore, $\text{dis}^2(\pi^{13}) > \text{dis}^2(\pi^{12})$ if $\alpha > 4 + 4N/\sqrt{N-1}$. Since $f(x) = x^2$ is a strictly monotonic bijection on $[0, \infty)$, it follows that $\text{dis}(\pi^{12}) < \text{dis}(\pi^{13})$. □

B Weak Isomorphism and Coarsening

As previously mentioned, the space of (compact) measure networks equipped with the Gromov-Wasserstein distance is a pseudometric space. A pseudometric space consists of a set X and a pseudo-metric $\hat{d}$, differing from a metric in that we can have $\hat{d}(x, y) = 0$ for $x \neq y$. Moreover, the Gromov-Wasserstein distance is unique up to weak-isomorphism [50], that is $\hat{d}(x, y) = 0$ if and only if x and y are weakly isomorphic. There are two notions of weak-isomorphism discussed in [50], the latter of which is only necessary for infinite measure networks and is therefore beyond the scope of this work.

A pair of measure networks $X = (X, S_X, \mu_X)$ and $Y = (Y, S_Y, \mu_Y)$ is called weakly isomorphic if there exists a third measure network $Z = (Z, S_Z, \mu_Z)$ and injective maps $\phi : Z \to X$ and $\psi : Z \to Y$ such that the following conditions hold:

1. $\phi_*(\mu_Z) = \mu_X$ and $\psi_*(\mu_Z) = \mu_Y$
2. $\sup_{z_1, z_2 \in Z} |\phi^* S_X(z_1, z_2) - \psi^* S_Y(z_1, z_2)| = 0$

where $\phi_* = \mu_Z \circ \phi^{-1}$ and $\phi^* w_X(z_1, z_2) = w_Z(\phi(z_1), \phi(z_2))$; ψ_* and ψ^* are defined *mutatis mutandis*. The concept of a terminal network is discussed in [50, 63] and is the most concise representation of a measure network in the GW geometry. Any measure network in a weak isomorphism class can be represented using its minimal representative via blow-ups [64]. We can determine if a measure network is the minimal representative by checking if there exists a pair of nodes with identical neighborhoods, or equivalently, a pair of nodes that, once merged, induce zero GW distortion (as in Prop. 1.)

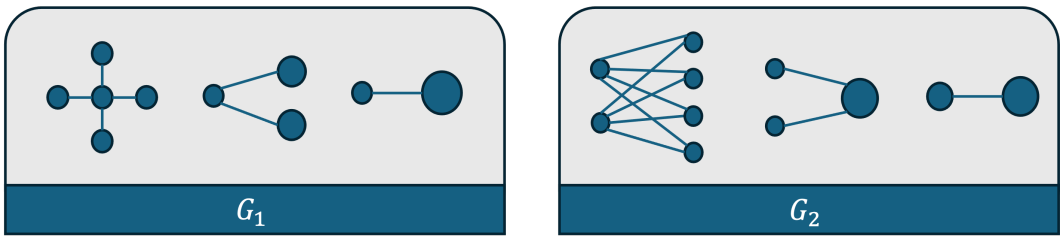

Figure 2: We show here two weak isomorphism classes of graphs. The leftmost networks in each class have uniform mass on nodes and the weights of all edges in each class are equal. The rightmost graphs are minimal representatives or terminal networks in their respective class. Visually, the minimal representatives of the graphs G_1 and G_2 appear quite similar, but comparing their leftmost representation reveals how different the graphs are. Classes G_1 and G_2 are both examples of complete bi-partite graphs; these classes of graphs benefit most from coarsening to the minimal representative, as a complete k-partite network can be reduced to a k-node minimal representative.

C Gromov-Wasserstein Coarsening and Sketching

The purpose of this section is to clarify the connection between the graph coarsening problem in the GW setting and the GW measure network sketching problem. The GW sketching problem has been previously considered in [53] in the case of metric measure spaces (a subset of measure networks [50]), where notions of duality were established between sketching and clustering with respect to the GW distance (albeit for the GW distance proposed by [38] which is not computationally feasible). The sketching problem in GW space of an N-point network G to a M-point network G' can be formulated as

$$\operatorname*{arg\,min}_{G' \in \mathcal{N}_M} d_{GW}^2(G, G') = \operatorname*{arg\,min}_{G' \in \mathcal{N}_M} \min_{\pi \in \Pi(\mu, \mu')} \text{dis}_2^2(\pi) \tag{14}$$

We can upper-bound (14) by restricting the feasibility set of the GW distance to those transport plans induced by coarsening matrices, i.e. $\pi = \text{diag}(\mu) C_p$,

$$\operatorname*{arg\,min}_{G' \in \mathcal{N}_M} d_{GW}^2(G, G') \leq \operatorname*{arg\,min}_{G' \in \mathcal{N}_M} \min_{C_p \in \mathcal{C}_{N,M}} \text{dis}_2^2(\text{diag}(\mu) C_p). \tag{15}$$

As is pointed out in [24], given C_p the G' that minimizes the upper bound in Eq. (15) is the semi-relaxed GW barycenter [52]

$$\begin{aligned}
\min_{G' \in \mathcal{N}_M} d_{GW}^2(G, G') &\leq \min_{G' \in \mathcal{N}_M} \min_{C_p \in \mathcal{C}_{N,M}} \text{dis}^2(\text{diag}(\mu) C_p) && (16)\\
&= \min_{G' \in \mathcal{N}_M} \min_{C_p \in \mathcal{C}_{N,M}} \langle \mathcal{L}_2^2(G, G') \otimes \text{diag}(\mu) C_p, \text{diag}(\mu) C_p \rangle && (17)\\
&= \min_{C_p \in \mathcal{C}_{N,M}} \min_{G' \in \mathcal{N}_M} \langle \mathcal{L}_2^2(G, G') \otimes \text{diag}(\mu) C_p, \text{diag}(\mu) C_p \rangle && (18)\\
&= \min_{C_p \in \mathcal{C}_{N,M}} \langle \mathcal{L}_2^2(G, C_w^\top G C_w) \otimes \text{diag}(\mu) C_p, \text{diag}(\mu) C_p \rangle && (19)
\end{aligned}$$

where Eq. (19) follows from the fact that $G' = C_w^\top G C_w$ is the closed-form solution of the inner minimization problem in Eq. (18), as shown in [24, Appendix B], [51, Equation 14]. Taken together, we have

$$\operatorname*{arg\,min}_{G' \in \mathcal{N}_M} d_{GW}^2(G, G') \leq \operatorname*{arg\,min}_{C_p \in \mathcal{C}_{N,M}} \| (G - C_p C_w^\top G C_w C_p^\top) \odot (\mu \mu^\top)^{1/2} \|_F^2. \tag{20}$$

Notice that the restriction to transport maps of the form $\pi = \text{diag}(\mu) C_p$ is equivalent to requiring that there exists a Gromov-Monge map between G and G'.