Multi-Marginal Optimal Transport
Defines a Generalized Metric

José Bento
Boston College
jose.bento@bc.edu

Liang Mi
Arizona State University
liangmi@asu.edu

Abstract

Optimal transport (OT) is rapidly finding its way into machine learning. Favoring its use are its metric properties. Indeed, many problems admit solution guarantees only for objects embedded in a metric space, and the use of non-metrics can make their solving more difficult. Multi-marginal OT (MMOT) generalizes OT to simultaneously transporting multiple distributions. It captures important relations that are missed if the transport is pairwise. Research on MMOT, however, has been focused on its existence, the uniqueness and structure of transports, applications, practical algorithms, and the choice of cost functions. There is a lack of discussion on the metric properties of MMOT, which critically limits its theoretical and practical use. Here, we prove that MMOT defines a generalized metric. We first explain the difficulty of proving this via two negative results. Afterwards, we prove key intermediate steps and then prove MMOT’s metric properties. Finally, we show that the generalized triangle inequality that MMOT satisfies cannot be improved.

1 Introduction

Let \((\Omega^1, F^1, p^1)\) and \((\Omega^2, F^2, p^2)\) be two probability spaces. Given a cost function \(d^{1,2} : \Omega^1 \times \Omega^2 \to \mathbb{R} \geq 0\), and \(\ell \geq 1\), the (Kantorovich) Optimal Transport (OT) problem \([1]\) seeks to find

\[
\inf_{p^{1,2}} \left( \int_{\Omega^1 \times \Omega^2} (d^{1,2})^\ell \, dp^{1,2} \right)^{\frac{1}{\ell}} \text{ subject to } \int_{\Omega^1} dp^{1,2} = p^2 \text{ and } \int_{\Omega^2} dp^{1,2} = p^1,
\]

where the inf is over measures \(p^{1,2}\) on \(\Omega^1 \times \Omega^2\). Problem \([1]\) is typically studied under the assumptions that \(\Omega^1\) and \(\Omega^2\) are in a Polish space on which \(d\) is a metric, in which case the minimum of \([1]\) is the Wasserstein distance (WD). The WD is popular in many applications including shape interpolation \([2]\), generative modeling \([3, 4]\), domain adaptation \([5]\), and dictionary learning \([6]\).

The WD is a metric on the space of probability measures \([7]\), and this property is useful in many machine learning tasks, e.g., clustering \([8, 9]\), nearest-neighbor search \([10, 11, 12]\), and outlier detection \([13]\). Indeed, some of these tasks are tractable, or allow theoretical guarantees, when done on a metric space. E.g., finding the nearest neighbor \([10, 11, 12]\) or the diameter \([14]\) of a dataset requires a polylogarithmic computational effort under metric assumptions; approximation algorithms for clustering rely on metric assumptions, whose absence worsens known bounds \([15]\); also, \([16]\) uses the metric properties of the WD to study object matching via metric invariants.

Recently, a generalization of OT to multiple marginal measures has gained attention. Given probability spaces \((\Omega^i, F^i, p^i), i = 1, \ldots, n\), a function \(d : \Omega^1 \times \ldots \times \Omega^n \to \mathbb{R} \geq 0\), \(\ell \geq 1\), and \(\Omega^{-i} \overset{\Delta}{=} \Omega^1 \times \ldots \times \Omega^{i-1} \times \Omega^{i+1} \times \ldots \times \Omega^n\), the Multi-Marginal Optimal Transport (MMOT) seeks

\[
\inf_{p} \left( \int_{\Omega^1 \times \ldots \times \Omega^n} (d) \, dp \right)^{\frac{1}{\ell}} \text{ subject to } \int_{\Omega^{-i}} dp = p^i \ \forall i = 1, \ldots, n,
\]

Preprint. Under review.
where the infimum is taken over all measures \( p \) on \( \Omega^1 \times \ldots \times \Omega^n \).

Much of the discussion on MMOT has focused on its existence, the uniqueness and structure of both Monge and Kantorovich solutions, applications, practical algorithms, and the choice of the cost function [17][18][19][20]. The term MMOT was coined in [21], and was surveyed by the same authors in [22]. Applications of MMOT include [23] for image translation, and [24] for image registration, multi-agent matching with fairness requirement, and labeling for classification.

Unfortunately, there is a lack of discussion about the (generalized) metric properties of MMOT. Since the metric property of the WD is useful in so many applications, understanding when the (potential) minimum of (2), \( W(p^1, \ldots, p^n) \), has metric-like properties is critical, theoretically and practically.

**Example:** Given a set \( S \) with \( n \) distributions we can find its 3-diameter \( \Delta \triangleq \max_{p^1, p^2, p^3 \in S} W(p^1, p^2, p^3) \) with \( \binom{n}{3} \) evaluations of \( W \). What if \( W \) satisfies the generalized triangle inequality \( W(p^1, p^2, p^3) \leq W(p^1, p^2) + W(p^2, p^3) + W(p^1, p^3) \)? We now know that for at least \( n/3 \) distribution triplets \( W \geq \Delta/3 \). Indeed, if \( \Delta = W(p^1, p^2, p^3) \), then for all \( p^4 \in S \), we cannot simultaneously have \( W(p^1, p^2, p^4), W(p^2, p^4), W(p^3, p^4) < \Delta/3 \). Therefore, if we evaluate \( W \) on random distribution triplets, we are guaranteed to find a \((1/3)\)-approximation of \( \Delta \) with only \( O(n^2) \) evaluations of \( W \) on average, and improvement over \( \binom{n}{3} \).

In this paper, we show for the first time that MMOT defines a generalized metric. We first explain the difficulty of proving this via two negative results (Sec. 3.1 and 3.2). Afterwards, we establish key intermediate steps and prove the generalized metric properties (Sec. 3.3). Finally, we show that the triangle inequality that MMOT satisfies cannot be improved, up to a linear factor.

## 2 Definitions and setup

### 2.1 Lists

Expressions that depend on a list of symbols indexed consecutively are abbreviated using “\( \forall \)”. In particular, we write \( s_1, \ldots, s_k \) as \( s^{1:k} \), \( \Omega^1, \ldots, \Omega^k \) as \( \Omega^{1:k} \), and \( A^1, \ldots, A^n \) as \( A^{1:n} \). Note that \( A^{1:k} \) differs from \( A^{1:k-1} \). Assuming that \( s_k > s_1 \), the former means \( A^{1:k}, A^{s_1+1}, A^{s_1+2}, \ldots, A^{s_k-1} \). By it self, \( 1 : i \) has no meaning, and it does not mean \( 1, \ldots, i \). For \( i \in \mathbb{N} \), we let \( [i] \triangleq \{1, \ldots, i\} \). The symbol \( \oplus \) denotes a list join operation with no duplicate removal, e.g. \( \{x, y\} \oplus \{x, z\} = \{x, y, z\} \).

### 2.2 Bra-ket operator

Given two multidimensional arrays \( A \) and \( B \) with the same dimensions, and \( \ell \in \mathbb{N} \), we define \( (A, B)_\ell \triangleq \sum_{i_1, \ldots, i_k} (A_{i_1, \ldots, i_k})^\ell B_{i_1, \ldots, i_k} \), where \((A_{i_1, \ldots, i_k})^\ell \) is the \( \ell \)th power of \( A_{i_1, \ldots, i_k} \).

### 2.3 Probability spaces

To facilitate exposition, we state our main contributions for probability spaces with a sample space in \( \Omega \), finite, an event set \( \sigma \)-algebra which is the power set of the sample space, and a probability measure described by a probability mass function, but they extend to more general settings. We refer to probability mass functions using bold letters, e.g. \( p, q, r \), etc.

When talking about \( n \) probability spaces, the \( i \)th space has sample space \( \Omega^i = \{\Omega^i_{1:m_i}\} \subseteq \Omega \), an event space \( 2^{\Omega_i} \), and a probability mass function \( p^i \), or \( q^i \), or \( r^i \), etc. Variable \( m_i \) is the number of atoms in \( \Omega^i \). Symbol \( p^i_{s_i} \) denotes the probability of the atomic event \( \{\Omega^i_{s_i}\} \). Without loss of generality (w.l.o.g.) we assume \( p^i_{s_i} > 0 \), \( \forall i \in [n], \forall s_i \in [m_i] \). Our notation assumes that atoms can ordered, but our results extend beyond this assumption. W.l.o.g., we assume that \( \Omega^i = \Omega^i_{s_i} \) if and only if \( s_i = i \).

Symbol \( p^{i:k} \) denotes a mass function for the probability space with sample space \( \Omega^{i \times \ldots \times i^k} \) and event space \( 2^{\Omega^{i \times \ldots \times i^k}} \). In particular, \( p^{i, k}_{s_i, \ldots, s_k} \) (i.e. \( p^{i \times \ldots \times i: k}_{s_i, \ldots, s_k} \)) is the probability of the atomic event \( \{(\Omega_{s_i}, \ldots, \Omega_{s_k})\} \). We use \( p^{i:k|j:r} \) to denote a probability mass function for the probability space with sample space \( \Omega^{i \times \ldots \times i^k} \) and event space \( 2^{\Omega^{i \times \ldots \times i^k}} \), such that \( p^{i:k|j:r}_{s_i, s_i+1, \ldots, j_r} \approx p^{i:k}_{s_i, s_i+1, \ldots, s_k} / p^{j:r}_{s_i+1, \ldots, s_k} \), i.e. it represents a conditional probability.

Given the mass functions \( p^i \) and \( p^j \), for the sample spaces \( \Omega^i \) and \( \Omega^j \) respectively, symbol \( p^{i:j} \) denotes a mass function for \( \Omega^i \times \Omega^j \) where \( (p^{i:j})_{s_i,s_j} \triangleq p^{i:j}_{s_i,s_j} \) is the probability of the atomic event \( \{(\Omega^i_{s_i}, \Omega^j_{s_j})\} \). This notation extends to more than two masses. Note that \( p^{i:j} \neq p^{j:i} \), in terms of the way we index the atoms of both distributions.
Definition 1 (Gluing map). Consider a mass function $q^k$ over $\Omega^k$ and $n-1$ conditional mass functions \{ $q^{i|k}$ \}_{i \in [n]\setminus \{k \}} over \{ \Omega^i \}_{i \in [n]\setminus \{k \}}. The map $G$ defines the mass function $p$ over $\Omega^1 \times \ldots \times \Omega^n$ as

$$p = G \left( q^k, q^{i|k} \right)_{i \in [n]\setminus \{k \}} = \left( \prod_{i \in [k-1]} q^{i|k} \right) q^k \left( \prod_{i \in [n-k]} q^{i|k} \right).$$

To be more specific,

$$p_{s_1,\ldots,s_n} = d^k_{s,k} \prod_{i \in [n]\setminus \{k \}} q^{i|k}_{s_i}.$$  \hspace{1cm} (4)

2.4 Distances and metrics

We use “distance” to refer to an object that, depending on extra assumptions, might, or might not, have the properties of a (generalized) metric. We use the standard definition for a metric, and for generalized metrics we use the definition in (2).

Definition 2 (Metric). Let $d = \{d^{i,j}\}_{i,j}$ be a set of distances of the form $d^{i,j} : \Omega_i \times \Omega_j \to \mathbb{R}$ and $d^{i,j}(\Omega_s, \Omega_t) \equiv d^{s,t}. We say that $d$ is a metric when, for any $i,j,k$, and any $s \in [m^k], t \in [m^k]$, we have i) $d^{s,t} \geq 0$; ii) $d^{i,j} = d^{j,i}$; iii) $d^{i,i} = 0$ iff $\Omega_i = \Omega_i$; iv) $d^{i,j} \leq d^{s,t} + d^{s,t}.$

Definition 3 (Generalized metric). Let $d = \{d^{i,j}\}_{i,j}$ be a set of distances of the form $d^{i,j} : \Omega_i \times \ldots \times \Omega_i \to \mathbb{R}$ and $d^{i,j}(\Omega_s, \ldots, \Omega_t) \equiv d^{s_1,\ldots,s_n}. We say that $d$ is a $(n, C(n))$-metric when, for any $i_{n+1}$ and $s_{n+1}$, for any $s_r \in [m^r]$, we have i) $d^{s_{n+1}} \geq 0$; ii) $d^{s_{n+1}} = d^{s_{n+1}}$ for any permutation $\sigma$; iii) $d^{i,j} = 0$ iff $i = j \forall r \in [n];$ iv) $C(n) \leq \sum_{s_{n+1}} d^{s_{n+1}} \leq \sum_{i_{n+1}} d^{i_{n+1}}.$

Definition 4 (Generalized metric). Let $W$ be a map from $n$ probability spaces to $\mathbb{R}$ such that $W^{i_1,\ldots,i_n} \Delta W(p^{i_1,\ldots,i_n} = \text{the image of the probability spaces with indices } i_{n+1}. Map W is an $(n, C(n))$-metric when, for any $n + 1$ probability spaces with samples $\Omega^{1,\ldots,n}$ and masses $p^{i_1,\ldots,n}$, and any permutation $\sigma$ we have

1. $W^{1,\ldots,n} \geq 0$,
2. $W^{1,\ldots,n} = 0$ iff $p^i = p^j, \Omega^i = \Omega^j, \forall i, j$,
3. $W^{1,\ldots,n} = W^{n(1,\ldots,n)}$, for any map $\sigma$,
4. $C(n)W^{1,\ldots,n} \leq \sum_{r=1}^{n} W^{1,\ldots,n-1,\ldots,n+1}.$

Remark 1. Equalities $p^i = p^j$ and $\Omega^i = \Omega^j$, mean that $m^i = m^j$, and that there exists a bijection $b^{i,j}(-)$ from $[m^i]$ to $[m^j]$ such that $p^i_s = p^j_{b^{i,j}(s)}$ and $\Omega^i_s = \Omega^j_{b^{i,j}(s)} \forall s \in [m^i].$

Remark 2. When $C(n) = 1$, we abbreviate $(n, 1)$-metric by $n$-metric.

Remark 3. Our notions of metric and generalized metric are more general than usual in the sense that they support the use of different functions depending on the spaces from where we are drawing elements. This grants an extra layer of generality to our results.

In our setup, the inf in (2) is always attained and amounts to solving an LP. We refer to the minimizing distributions by $p^*, q^*, r^*$, etc. We define the following map from $n$ probability spaces to $\mathbb{R}$.

Definition 5 (MMOT distance for finite spaces). Let $d = \{d^{i,j}\}_{i,j}$ be a set of distances of the form $d^{i,j} : \Omega_i \times \ldots \times \Omega_i \to \mathbb{R}$ and $d^{i,j}(\Omega_s, \ldots, \Omega_t) \equiv d^{s_1,\ldots,s_n}. The MMOT distance associated with $d$ for $n$ probability spaces with masses $p^{i,n}$ over $\Omega^{i,n}$ is

$$W(p^{i,n}) \Delta W^{i,n} = \min_{r^i : p^i = p^i \forall s \in [n]} \left( \langle d^{i,n}, r^i \rangle \right)^{1/2},$$

where $r$ is a mass function over $\Omega^{i} \times \ldots \times \Omega^{n},$ and $r^i_s$ be the marginal probability of $r$ on $\Omega^{i,s}.$

3 Main results

To prove that MMOT leads to an $n$-metric, it is natural to extend the ideas in the classical proof that WD is a metric. The hardest property to prove is the triangle inequality. See Fig. 1 (Center) for a geometric analogy of property 3 in Def. 1. Its proof for the WD follows from (a) a gluing lemma and the assumption that (b) $d$ itself is a metric (Def. 2). Our hope is that if we can prove (a) a generalized gluing lemma and assume that (b) $d$ is a generalized metric, that we can prove what we want. Unfortunately, to our surprise, and as we explain in Sec. 3.1 and 3.2 (a) is not possible, and (b) is not enough. This requires us to develop completely new proofs.
3.1 The gluing lemma does not generalize to higher dimensions

The gluing lemma used to prove that WD is a metric is as follows. For its proof see [7], Lemma 2.1.

**Lemma 1 (Gluing lemma).** Let \( p^{1,3} \) and \( p^{2,3} \) be arbitrary mass functions for \( \Omega^1 \times \Omega^3 \) and \( \Omega^2 \times \Omega^3 \), respectively, with the same marginal, \( p^3 \), over \( \Omega^3 \). There exists a mass function \( r^{1,2,3} \) for \( \Omega^1 \times \Omega^2 \times \Omega^3 \) whose marginals over \( \Omega^1 \times \Omega^3 \) and \( \Omega^2 \times \Omega^3 \) equal \( p^{1,3} \) and \( p^{2,3} \) respectively.

The way Lemma 1 is used to prove WD’s triangle inequality is as follows. Assume \( d \) is a metric (Def. 2). Let \( \ell \) = 1 for simplicity. Let \( p^{1,2} \), \( p^{1,3} \), and \( p^{2,3} \) optimal transports such that \( \mathcal{W}^{1,2} = \langle p^{1,2}, d^{1,2} \rangle \), \( \mathcal{W}^{1,3} = \langle p^{1,3}, d^{1,3} \rangle \), and \( \mathcal{W}^{2,3} = \langle p^{2,3}, d^{2,3} \rangle \). Define \( r^{1,2,3} \) as in Lemma 1 and let \( r^{1,3} \), \( r^{2,3} \), and \( r^{1,1} \) be its bivariate marginals. Then we have that

\[
\langle p^{1,2}, d^{1,2} \rangle \leq \langle p^{1,3}, d^{1,3} \rangle + \langle p^{2,3}, d^{2,3} \rangle \] \tag{6}

Our first roadblock is that Lemma 1 does not generalize to higher dimensions. For simplicity, we now omit the sample spaces on which mass functions are defined. When a set of mass functions have all their marginals over the same sample sub-spaces equal, we will say they are compatible.

**Theorem 1 (No gluing).** There exists mass functions \( p^{1,2,4} \), \( p^{1,3,4} \), and \( p^{2,3,4} \) with compatible marginals such that there is no mass function \( r^{1,2,3,4} \) compatible with them.

**Proof.** If this was not the case, then it would be true that, given arbitrary mass functions \( p^{1,2} \), \( p^{1,3} \), and \( p^{2,3} \) with compatible univariate marginals, we should be able to find \( r^{1,2,3} \) whose bivariate marginals equal these three mass functions. But this is not the case. For example, let \( p^{1,2} = p^{1,3} = [1, 0, 1; 0, 0, 0] / \beta \) and \( p^{2,3} = [1, 1, 1; 1, 1, 1] / 9 \) (we are using matrix notation for the marginals). These marginals have compatible univariate marginals, namely, \( p^1 = [2, 1, 0] / 3 \) and \( p^2 = p^3 = [1, 1, 1] / 3 \). Yet, the following system of eqs. over \( \{ r_{i,j,k} \}_{i,j,k \in [3]} \) is easily checked to be infeasible \( \sum_i r_{i,j,k} = p^{2,3}_{j,k} \), \( \sum_j r_{i,j,k} = p^{1,3}_{i,k} \), \( \sum_k r_{i,j,k} = p^{1,2}_{i,j} \) for all \( i, j, k \).}

\( \square \)

### 3.2 Cost \( d \) being an n-metric is not a sufficient condition for MMOT to be an n-metric

**Remark 4.** The theorem can be generalized to spaces of dim. \( > 2 \), and to \( n > 3 \), and \( \ell > 1 \).

**Proof.** Let \( \Omega \) be the six points in Figure 1(left), where we assume that \( 0 < \epsilon < 1 \), and hence that there are no three co-linear points, and no two equal points. Let \( p^1 \), \( p^2 \), \( p^3 \), and \( p^4 \) be as in Fig. 1(left), each is represented by a unique color and is uniformly distributed over the points of the same color. Given any \( x, y, z \in \Omega \) let \( d(x, y, z) = \gamma \) if exactly two points are equal, and let \( d(x, y, z) = \) be the area of the corresponding triangle otherwise, where \( \gamma \) lower bounds the area of the triangle formed by any three non-co-linear points, e.g. \( \gamma = \epsilon / 4 \). A few geometric considerations (see Appendix A) show that \( d \) is an n-metric (\( n = 3 \), \( C(n) = 1 \)) and that holds as \( \frac{1}{n} + \frac{1}{n} + \frac{1}{n} \).}

\( \square \)

### 3.3 Pairwise MMOT is a generalized metric

We will prove that the properties in Def. 4 hold for the following variant of Def. 4.

**Definition 6 (Pairwise MMOT distance).** Let \( d \) be a set of distances of the form \( d_{i,j} : \Omega_i \times \Omega_j \mapsto \mathbb{R} \) and \( d_{i,j} : \Omega_i \times \Omega_j \mapsto d_{i,j} \). The Pairwise MMOT distance associated with \( d \) for \( n \) probability spaces with masses \( p^{1..n} \) over \( \Omega^{1..n} \) is

\[
\mathcal{W}(p^{1..n}) \triangleq \mathcal{W}^{1..n} = \min_{r \vdash r^{1..n} \vdash \Omega^{1..n}} \sum_{1 \leq s < t \leq n} (d_{i,s}^{r_{i,s}} \times d_{i,t}^{r_{i,t}})^{1/2} \] \tag{7}

where \( r \) is a mass over \( \Omega^{1..n} \), with marginals \( r^{1..n} \) over \( \Omega^{1..n} \), resp.
We currently do not know the most general conditions under which Def. 3 is an d
is a metric (Def. 2). However, because of Theorem 2 we know that this is not sufficient to guarantee that the pairwise MMOT distance is a n-metric, which only makes the proof of the next theorem all the more interesting.

**Theorem 3.** If d is a metric (Def. 2), then the pairwise MMOT distance (Def. 6) associated with d is a (n, C(n))-metric, with C(n) ≥ 1.

We currently do not know the most general conditions under which Def. 6 is an n-metric. However, working with Def. 6 allows us sharply bound the best possible C(n), which would unlikely be possible in a general setting. As Theorem 4 shows, the best C(n) is C(n) = Θ(n).

**Theorem 4.** In Theorem 3 the constant C(n) can be made larger than (n − 1)/5 for n > 7, and there exists sample spaces Ω1:n, mass functions p1:n, and a metric d over Ω1:n such that C(n) ≤ n − 1.

## 4 Main proof ideas

Our main technical contribution is our proof that the generalized triangle inequality - property (2) in Def. 4 - holds with C(n) ≥ (n − 1)/5, n > 7, if d is a metric (Def. 2), i.e. the first part of Thrm. 4. We give this proof in this section. The other proofs are also non-trivial and novel, but for spaces reasons are in the Appendix. A full proof of Thrm. 3 is in App. C and the proof of the second part of Thrm. 4 is in App. D.

To prove that in Def. 4 holds with C(n) ≥ (n − 1)/5, n > 7, we need the following special tool.

### 4.1 Special hash function

**Definition 7.** The map $H^n$ transforms a triple $(i, j, r)$, $1 ≤ i < j ≤ n$, $r ∈ [n − 1]$ to either 2, 3, or 4 triples according to

$$(i, j, r) → H^n(i, j, r) = H'_1(i, j, r) ∪ H'_2(i, j, r),$$

where

$$H'_1(i, j, r) = \begin{cases} \{(i, r, h'(i, r))\} & \text{if } j = h'(i, r), \\ \{(i, j, h'(i, r), (j, r, h'(i, r))\} & \text{if } j ≠ h'(i, r), \end{cases}$$

and

$$h'(i, r) = \begin{cases} 1 + (i + r - 1) \mod n & \text{if } i < n \\ 1 + (r \mod (n - 1)) & \text{if } i = n. \end{cases}$$

We assume that the first two components of each output triple are ordered. For example, $(i, r, h'(j, r)) ≡ (\min\{i, r\}, \max\{i, r\}, h'(j, r))$. 

Figure 1: (Left) Sample space Ω, mass functions $\{p^i\}_{i=1}^4$, and cost function $d$ that lead to violation of the total area of any three faces in a tetrahedron is greater than that of the fourth face. (Right) Graph whose edges are pairs of scenarios that cannot both hold. Any maximum independent set has size 5, which proves Lemma 2.
The following property of $\mathcal{H}^n$ is critical to lower bound $C(n)$.

**Lemma 2.** Let $(a, b, c) \in \mathcal{H}^n(i, j, r)$, $1 \leq i < j \leq n$, $r \in [n - 1]$. Then, $1 \leq a \leq b \leq n$, $1 \leq c \leq n$, and $c \notin \{a, b\}$. Furthermore,

$$
\bigoplus_{1 \leq i < j \leq n} \mathcal{H}^n(i, j, r)
$$

has at most 5 copies of each triple, where two triples are equal if and only if they agree component-wise.

**Remark 6.** Note that we might have $a = b$ in an triple $(a, b, c)$ output by $\mathcal{H}^n$.

**Remark 7.** For example, if $n = 4$, all 5 triples $(1, 2, 3)$, $(1, 3, 2)$, $(2, 3, 2)$, $\mathcal{H}^n(1, 2, 2)$, and $(2, 3, 1)$ map to $(3, 2, 1)$ and $(1, 4, 1)$ map to $(1, 1, 2)$ whose first two components equal.

**Proof.** The fact that $1 \leq a \leq b \leq n$ and that $1 \leq c \leq n$ is immediate. The fact that $c \notin \{a, b\}$ amounts to checking that $h'(i, r) \notin \{i, r\}$ for all $i \in [n]$, and $r \in [n - 1]$. This can be checked directly from $\mathcal{H}^n$. E.g. $h'(i, r) = i$ would imply either $(i = n) \land (r \mod n - 1 = i - 1)$, or $(i < n) \land ((i + r - 1) \mod n = i - 1)$, both of which are impossible. The rest of the proof amounts to checking that if $(a, b, c)$ is in the output of $\mathcal{H}^n$, then there are at most 5 different inputs that lead to $(a, b, c)$. There are 10 possible candidate input triples that lead to output $(a, b, c)$. Namely,

1. $(a, b, c) = (i_1, r_1, h'(i_1, r_1)) = \mathcal{H}^n(i_1, j_1, r_1)$, if $j_1 = h'(i_1, r_1)$ and $i_1 < r_1$,
2. $(a, b, c) = (r_2, i_2, h'(i_2, r_2)) = \mathcal{H}^n(i_2, j_2, r_2)$, if $j_2 = h'(i_2, r_2)$ and $r_2 < i_2$,
3. $(a, b, c) = (i_3, j_3, h'(i_3, r_3)) = \mathcal{H}^n(i_3, j_3, r_3)$, if $j_3 \neq h'(i_3, r_3)$,
4. $(a, b, c) = (j_4, r_4, h'(i_4, r_4)) = \mathcal{H}^n(i_4, j_4, r_4)$, if $j_4 \neq h'(i_4, r_4)$ and $j_4 < r_4$,
5. $(a, b, c) = (r_5, i_5, h'(i_5, r_5)) = \mathcal{H}^n(i_5, j_5, r_5)$, if $j_5 \neq h'(i_5, r_5)$ and $r_5 < j_5$,
6. $(a, b, c) = (j_6, r_6, h'(i_6, r_6)) = \mathcal{H}^n(i_6, j_6, r_6)$, if $i_6 = h'(j_6, r_6)$ and $j_6 < r_6$,
7. $(a, b, c) = (r_7, r_7, h'(i_7, r_7)) = \mathcal{H}^n(i_7, j_7, r_7)$, if $i_7 = h'(j_7, r_7)$ and $r_7 < j_7$,
8. $(a, b, c) = (i_8, r_8, h'(j_8, r_8)) = \mathcal{H}^n(i_8, j_8, r_8)$, if $i_8 \neq h'(j_8, r_8)$,
9. $(a, b, c) = (i_9, r_9, h'(j_9, r_9)) = \mathcal{H}^n(i_9, j_9, r_9)$, if $i_9 \neq h'(j_9, r_9)$ and $i_9 < r_9$,
10. $(a, b, c) = (r_{10}, i_{10}, h'(j_{10}, r_{10})) = \mathcal{H}^n(i_{10}, j_{10}, r_{10})$, if $i_{10} \neq h'(j_{10}, r_{10})$ and $r_{10} < i_{10}$.

Twelve pairs of the 10 scenarios above cannot simultaneously hold. The 12 pairs of scenarios that cannot both hold are displayed as edges in a graph in Figure 4(right). E.g. edge $(4, 10)$ represents that scenarios 4 and 10 cannot both hold. The proof that the pairs represented by edges in Fig. 4(right) cannot both hold is in Appendix B. A maximum independent set of this graph is $\{3, 8, 9, 2, x\}$, where $x \in \{1, 4, 6, 10\}$. Thus at most 5 input scenarios lead to a given output triple.

### 4.2 Useful lemmas

We also need then following lemmas.

**Lemma 3.** Let $p$ be as in Def. 2 eq. 3, for some $q_i$ and $\{q_{i|k}\}_{i \in [n]}$. Let $p^i$ and $p^{i|k, i \neq k}$, be the marginals of $p$ over $\Omega^i$ and $\Omega^i \times \Omega^k$, respectively. Let $q^{i|k} = q_{i|k} q^i$, $i \neq k$, and let $q^i$ be its marginal over $\Omega^i$. We have that $p^i = q^i \forall i$, and $p^{i|k} = q^{i|k} \forall i \neq k$.

**Proof.** Think of $p$ as describing $n$ discrete random variables (r.v.’s). It follows from the factorisation in 3 that conditioned on the $k$ r.v. the other r.v.’s are independent. The result follows.

**Lemma 4.** Let $d$ be a metric (Def. 2), and $p$ a mass function over $\Omega^1 \times \ldots \times \Omega^n$. Let $p^{i|j}$ be the marginal of $p$ over $\Omega^i \times \Omega^j$. Define $w_{i,j} = (\langle d^{i,j}, p^{i,j} \rangle)^{1/2}$. For any $i, j, k \in [n]$ and $\ell \in \mathbb{N}$ we have that $w_{i,j} \leq w_{i,k} + w_{k,j}$.

**Proof.** Let $p^{i|j,k}$ be the marginal of $p$ over $\Omega^i \times \Omega^j \times \Omega^k$. Write $w_{i,j} = (\langle d^{i,j}, p^{i,j} \rangle)^{1/\ell} = \left(\sum_{s,t,r} (d^{i,j}_{s,t,r} p^{i,j}_{s,t,r})^{1/\ell}\right)^{\ell} \leq \left(\sum_{s,t,r} (d^{i,j}_{s,t,r} + d^{k}_{r,l}) (p^{i|j,k}_{s,t,r})^{1/\ell}\right)^{\ell}$. Use Minkowski’s ineq. on a $L_\ell$ space with measure $p^{i,j,k}$ to bound this by $\left(\sum_{s,t,r} (d^{i,j}_{s,t,r} p^{i,j,k}_{s,t,r})^{1/\ell}\right)^{1/\ell} \leq w_{i,k} + w_{k,j}$. □
4.3 Proof of lower bound on $G(n)$

We will show that, $(n-1)\mathcal{W}^{1,\ldots,n}$ can be upper bounded by $5\sum_{r=1}^{n} \mathcal{W}^{1,\ldots,r-1,r+1,\ldots,n+1}$, where we are using Def. [6] for $\mathcal{W}$.

For $r \in [n]$, let $p^{(sr)}$ be a minimizer for $\mathcal{W}^{1,\ldots,r-1,r+1,\ldots,n+1}$. We would normally use $r^{(sr)}$ for this minimizer, but, to avoid confusions between $r$ and $r$, we avoid doing so. For $i, j \in [n+1]\setminus\{r\}$, let $p^{(sr)ij\cdot}$ be the marginal of $p^{(sr)}$ for the sample space $\Omega^i \times \Omega^j$. Recall that since $p^{(sr)}$ satisfies the constraints in (9), its marginal for the sample space $\Omega^i$ equals $p^i$, which is given in advance.

Let $h'(\cdot, \cdot)$ be the map in (12). For each $r \in [n-1]$, define the following mass function over $\Omega^1 \times \ldots \times \Omega^n$

$$q^{(r)} = \mathcal{G} \left( p^r, \{p^{(sh'(i,r))ij\cdot}_{i \in [n]\setminus r} \right),$$

where $p^{(sh'(i,r))ij\cdot}$ is the mass function that satisfies $p^{(sh'(i,r))ij\cdot} = p^{(sh'(i,r))ij\cdot}$.

By Lemma [3], we know that $q^{(r)}$ exists. For all $i \in [n]$, and hence $q^{(r)}$ satisfies the optimization constraints in (9) for $\mathcal{W}^{1,\ldots,n}$. Therefore, we can write that

$$(n-1)\mathcal{W}^{1,\ldots,n} = \sum_{r=1}^{n-1} \sum_{1 \leq i < j \leq n} \left( \left< d^{ij}, p^{(r)} \right> \right)^{\frac{1}{2}} \leq \sum_{r=1}^{n-1} \sum_{1 \leq i < j \leq n} \left( \left< d^{ij}, q^{(r)} \right> \right)^{\frac{1}{2}},$$

where $p^{(r)}$ is the bivariate marginal over $\Omega^i \times \Omega^j$ of the minimizer $p^*$ for $\mathcal{W}^{1,\ldots,n}$.

We now bound each term in the inner most sum on the r.h.s. of (15) as

$$\left( \left< d^{ij}, q^{(r)} \right> \right)^{\frac{1}{2}} \leq \left( \left< d^{ir}, q^{(r)} \right> \right)^{\frac{1}{2}} + \left( \left< d^{jr}, q^{(r)} \right> \right)^{\frac{1}{2}},$$

where we assume $i \neq r$, $r \neq j$, and that: inequality (a) holds by Lemma [4]; (b) holds because $d$ is symmetric; and (c) holds because, by Lemma [3], $q^{(r)}$ equals $p^{(sh'(i,r))ij\cdot}$ and $p^{(sh'(i,r))ij\cdot} = p^{(sh'(i,r))ij\cdot}$.

Bounding the r.h.s. of (15) using (16) - (18), we re-write the resulting inequality using the notation

$$(n-1)\mathcal{W}^{1,\ldots,n} = \sum_{r=1}^{n-1} \sum_{1 \leq i < j \leq n} w(i,j,r) \leq \sum_{r=1}^{n-1} \sum_{1 \leq i < j \leq n} v(i,r,h'(i,r)) + v(j,r,h'(i,r)),$$

where (a) we are implicitly assuming that the first two components of each triple on the r.h.s. of (19) are ordered, i.e. if e.g. $r < i$ then $(r, i, h'(r, i))$ should be red as $(i, r, h'(i, r))$; (b) each $w(i,j,r)$ represents one $\left( \left< d^{ij}, p^{(r)} \right> \right)^{\frac{1}{2}}$ on the l.h.s. of (15); and (c) each $v(i,s,t)$ represents $\left( \left< d^{st}, p^{(s,t)} \right> \right)^{\frac{1}{2}}$ if $s \neq t$, and is zero if $s = t$. Since $h'(i, r) \notin \{i, r\}$, when $i \neq r$ the mass $p^{(sh'(i,r))ij\cdot}$ exists.

Finally, using this same compact notation, we write

$$5\sum_{r=1}^{n} \mathcal{W}^{1,\ldots,r-1,r+1,\ldots,n+1} = \sum_{r=1}^{n} \sum_{i,j \in [n+1]\setminus \{r\}, i < j} v(i,j,r),$$

and now we will show that (20) upper-bounds the r.h.s. of (19), finishing the proof.

First, by Lemma [4] and the symmetry of $d$, observe that the following inequalities are true

$$v(i,r,h'(i,r)) \leq v(i,j,h'(i,r)) + v(j,r,h'(i,r)),$$

and

$$v(j,r,h'(j,r)) \leq v(i,j,h'(j,r)) + v(i,r,h'(j,r)).$$
as long as for each triple \((a, b, c)\) in the above expressions, \(c \notin \{a, b\}\). We will use inequalities (21) and (22) to upper bound some of the terms on the r.h.s. of (19), and then we will show that the resulting sum can be upper bounded by (20). In particular, for each \((i, j, r)\) being considered by the two summations on the r.h.s. of (19), we will apply inequalities (21) and (22) such that the terms \(v_{(a,b,c)}\) that we get after their use have triples \((a, b, c)\) that match the triples in \(\mathcal{H}_n^{\mathcal{S}}((i,j,r))\), defined in Def. 7. To be concrete, for example, if \(\mathcal{H}_n^{\mathcal{S}}\) maps \((i, j, r)\) to \(\{(i, r, h'(i, r)), (r, j, h'(j, r))\}\), then we do not apply (21) and (22), and we leave \(v_{(r,i,h'(i,r))} + v_{(r,j,h'(j,r))}\) as is on the r.h.s. of (19). If, for example, \(\mathcal{H}_n^{\mathcal{S}}\) maps \((i, j, r)\) to \(\{(i, r, h'(i, r)), (i, j, h'(j, r)), (i, r, h'(j, r))\}\), then we leave the first term in \(v_{(r,i,h'(i,r))} + v_{(r,j,h'(j,r))}\) in the r.h.s. of (19) untouched, but we upper bound the second term using (22) to get \(v_{(i,r,h'(i,r))} + v_{(i,j,h'(j,r))} + v_{(i,r,h'(j,r))}\).

After proceeding in this fashion, and by Lemma 2, we know that all of the terms \(v_{(a,b,c)}\) that we obtain have triples \((a, b, c)\) with \(c \neq \{a, b\}\), \(c \in [n-1]\), and \(1 \leq a < b \leq n\). Therefore, these terms are either zero (if \(a = b\) or appear in (20)). Also because of Lemma 2, each triple \((a, b, c)\) with non-zero \(v_{(a,b,c)}\) will not appear more than 5 times. Therefore, the upper bound we build with the help of \(h'\) for the r.h.s of (19) can be upper bounded by (20).

5 Numerical experiments

We illustrate how using a MMOT that defines a \(n\)-metric, \(n > 2\), improves a task of clustering graphs of 7 different types, compared to using an OT or MMOT that defines a 2-metric or a non-\(n\)-metric.

We cluster graphs by i) computing their spectrum, ii) treating each spectrum as a distribution, iii) using WD and two different MMOT’s to compute distances among these distributions, and iv) feeding these distances to distance-based clustering algorithms to recover the true cluster memberships. We use spectral clustering based on normalized random-walk Laplacians [26] to produce one clustering solution out of the pairwise graph distances computed via WD. We also produce clustering solutions out of the triple-wise graph distances computed via Def. 6 (a \(n\)-metric), and via \(\mathcal{W}\) defined as in Thm. 2 (a non-\(n\)-metric). To do so, we use the hyper-graph-based clustering methods NH-Cut [27] and TTM [28, 29]. Details of our setup are in Appendix E.

Figures 2 (Left, Center) show that both TTM and NH-Cut work better when hyper-edges are computed using an \(n\)-metric. Figure 2 (Right) shows that clustering using only pairwise relationships among graphs leads to worse accuracy than if using triple-wise relationships as in Fig. 2 (Left, Center). This has been pointed out before in [30]. Note that a random prediction has 0.857 miss-classification rate.

![Figure 2: Comparing the effect that different distances and metrics have on clustering graphs.](image)

6 Future work

We have solved the problem of showing that a generalization of the optimal transport to multiple distributions, the pairwise multi-marginal optimal transport (pairwise MMOT), leads to a multi-distance that satisfies generalized metric properties. In particular, we have proved that the generalized triangle inequality that it satisfies cannot be improved, up to a linear factor. This now opens the door to us using pairwise MMOT in combination with several algorithms whose good performance depends on metric properties. At the same time, for a general MMOT, we have proved that the cost function being a generalized metric is not enough to guarantee that MMOT defines a generalized metric. In future work, we seek to find new sufficient conditions under which other variants of MMOT lead to generalized metrics, and, for certain families of MMOTs, find necessary conditions for these same properties to hold.
7 Broader impact

Our main contributions are theoretical in nature. Hence, this work does not present any direct foreseeable societal consequences. However, we do recognize that, indirectly, it empowers the movement towards using machine learning algorithms that deal with complex relationships among different data distributions, as opposed to those that only relate two data distributions at a time. This movement is leading to unprecedented inference powers, and, with it, a wide range of negative and positive potential social benefits. For example, [23] uses the types of methods discussed in this paper for image translation, which can lead to both better security, and a more accurate justice system, but also to major privacy violations [31,32]. The discussion of these thrice removed broader impacts (our contribution → algorithms → applications → impact) is beyond the scope of our paper.
References


A Details for proof of Theorem 2

Proof. Note that Definition 3 supports using a different function $d_{i,j,k}^{i,j,k}$ for different product sample spaces $\Omega^i \times \Omega^j \times \Omega^k$. In the case of Theorem 2 however, we only use $\Omega \times \Omega \times \Omega$, so, when checking the $n$-metric properties, we can drop the upper indices in $d$ in Definition 3.

For simplicity, we will abuse the notation and use $d(x,y,z)$ and $d_{i,j,k}$ interchangeably, where $i$, $j$, and $k$ are the index of $x$, $y$, and $z$, in $\Omega$.

Given $x, y, z, w \in \Omega$, it is immediate to see that (i) $d(x,y,z) \geq 0$, (ii) $d(x,y,z)$ is permutation invariant, and that (iii) $d(x,y,z) = 0$ if and only $x = y = z$ (remember that there are no three co-linear points in $\Omega$). It is also not hard to see that, $d(x,y,z) \leq d(x,y,w) + d(x,w,z) + d(w,y,z)$.

To be specific, if $d(x,y,z) = 0$, then the inequality is obvious. If $d(x,y,z) = \gamma$, then without loss of generality we can assume that $x = y \neq z$. In this case, if furthermore $w = x$, then $d(x,w,z) = \gamma$, and the inequality follows. If $w = z$, then $d(x,y,w) = \gamma$, and the inequality follows. If $w$ is different from $x, y, z$ then $\gamma \leq d(x,w,z)$, and the inequality follows. If $d(x,y,z) > \gamma$, it must be that $x, y$ and $z$ are different. In which case we need do consider two special cases. If $w$ is equal to one among $x, y, z$, say $w = x$ without loss of generality, then $d(x,y,z) = d(y,z,w)$, and the inequality follows. If $w$ is different from $x, y, z$, then we have $d(x,y,z) = d(x,y,w) + d(x,w,z) + d(w,y,z)$ if $w$ is contained by the triangle formed by $x, y$, and $z$, and otherwise, we have $d(x,y,z) < d(x,y,w) + d(x,w,z) + d(w,y,z)$. In other words, $d$ is an $n$-metric ($n = 3$).

Given a mass function $p^{i,j,k}$, the value $(d_{i,j,k}^{i,j,k}, p_{i,j,k}^{i,j,k})$ represents the average area of the triangle whose three vertices are sampled from $p_{i,j,k}^{i,j,k}$. Computing the MMOT distance $\mathcal{W}_{i,j,k}^{i,j,k}$ for the mass functions $p^i, p^j, p^k$, amounts to finding the mass function $p^{*i,j,k}$ with univariate marginals $p^i, p^j, p^k$ that minimizes this average area.

Now consider $p^1, p^2, p^3$, and $p^4$ as depicted in Figure 1. The mass functions $p^1$, $p^2$ assign probability one to each one of the blue and red points, and zero probability to every other point in $\Omega$. The mass functions $p^3$ and $p^4$ assign equal probability to each one of the green points, and orange points, respectively, and zero probability to every other point in $\Omega$.

Now we compute the distances $\mathcal{W}^{1,2,3}, \mathcal{W}^{1,2,4}, \mathcal{W}^{1,3,4}$, and $\mathcal{W}^{2,3,4}$. The MMOT distance $\mathcal{W}^{1,2,3}$ is equal to the average of the area of the two shaded triangles in Figure 3 (left), which is $\mathcal{W}^{1,2,3} = 0.5 \times (0.5) + 0.5 \times (0.5) = 0.5$. The distance $\mathcal{W}^{1,2,3}$ is equal to the average of the area of the two shaded triangles in Figure 3 (right), which is $\mathcal{W}^{1,2,4} = 0.5 \times (0.5) + 0.5 \times (0.25 - 0.5\varepsilon) = 0.125$.

![Figure 3](image_url)

Figure 3: (Left) Triangles associated with the optimal distribution of triples $p^{*1,2,3}$ associated with $\mathcal{W}^{1,2,3}$. (Right) Triangles associated with the optimal distribution of triples $p^{*1,2,4}$ associated with $\mathcal{W}^{1,2,4}$.

The MMOT distances for $\mathcal{W}^{1,3,4}$ and $\mathcal{W}^{2,3,4}$ are the same by symmetry. We focus on the computation of $\mathcal{W}^{2,3,4}$. Since both $p^3$ and $p^4$ are uniform over their respective supports, it must be the case that $p^{*2,3,4}$ - the optimal joint distribution in the computation of $\mathcal{W}^{2,3,4}$ - has a bi-variate marginal $p^{*3,4}$ of the form

$$
\{ (p^{*3,4}_{1,1} p^{*3,4}_{2,1}), (p^{*3,4}_{1,2} p^{*3,4}_{2,2}) \} = \left\{ \left\{ \alpha, \frac{1}{2} - \alpha \right\}, \left\{ \frac{1}{2} - \alpha, \alpha \right\} \right\},
$$

where $\alpha \in \left[ \frac{1}{2} \right]$. Therefore, the distance $\mathcal{W}^{2,3,4}$ is equal to the weighted average of the area of the four shaded triangles in Figure 4, where we split the four triangles into two different drawings for clarity sake. In other words,
where we are using the fact that \( \min_{\alpha \in [0, \frac{1}{2}]} (\text{linear function of } \alpha) \) must be minimized at either extreme \( \alpha = 0 \) or \( \alpha = \frac{1}{2} \).

It is finally straightforward to observe that \( \mathcal{W}^{1,2,3} = \frac{1}{2} > \frac{1}{8} + \left( \frac{1}{8} + \frac{\epsilon}{2} \right) + \left( \frac{1}{8} + \frac{\epsilon}{4} \right) = \mathcal{W}^{1,2,4} + \mathcal{W}^{3,4} + \mathcal{W}^{2,3,4}. \) \( \square \)

**B Details for proof of Lemma 2**

Recall the definitions:

\[
\mathcal{H}'_1(i,j,r) = \begin{cases} 
\{(i,r,h'(i,r))\} & \text{if } j = h'(i,r), \\
\{(i,j,h'(i,r),(j,r,h'(i,r))\} & \text{if } j \neq h'(i,r),
\end{cases}
\]

\[
\mathcal{H}'_2(i,j,r) = \begin{cases} 
\{(j,r,h'(j,r))\} & \text{if } i = h'(j,r), \\
\{(i,j,h'(j,r)),(i,r,h'(j,r))\} & \text{if } i \neq h'(j,r).
\end{cases}
\]

\[
h'(i,r) = \begin{cases} 
1 + ((i+r-1) \mod n) & \text{if } i < n, \\
1 + (r \mod (n-1)) & \text{if } i = n.
\end{cases}
\]

What remains to be proved is that several pairs of the 10 scenarios described in the proof of Lemma 2 cannot both hold.

**Proof.** Recall that for any input triple \((i,j,r)\) we always have \(1 \leq i < j \leq n\), and \(r \in [n-1]\).

Scenario 1) and 5) cannot both hold, because that would imply \(r_5 = i_1, j_5 = r_1\), which would imply \(h'(r_5,j_5) = h'(i_1,r_1) = h'(i_5,r_5)\), which since \(r_5, i_5 < n\) would imply \(i_5 = j_5\), contradicting \(i_5 < j_5\).

Scenarios 1) and 6) cannot both hold, because that would imply \(a = i_1 < j_1 = h'(i_1,r_1) = c = h'(j_6,r_6) = i_6 < j_6 = a\).

Scenarios 2) and 7) cannot both hold, because that would imply \(j_2 = h'(i_2,r_2) = h'(j_7,r_7) = i_7 < j_7 = i_2\), contradicting \(i_2 < j_2\).

Scenarios 1) and 4) cannot both hold, because that would imply \(j_4 = i_4 < n\) and \(r_4 = r_4\), which would imply \(h'(j_4,r_4) = h'(i_1,r_1) = h'(i_4,r_4)\), which since \(i_4, j_4 < n\) would imply \(j_4 = i_4\), contradicting \(i_4 < j_4\).

Scenarios 2) and 5) cannot both hold, because that would imply \(j_5 = i_2 < n, r_5 = r_2\), which would imply \(h'(j_5,r_5) = h'(i_2,r_2) = h'(i_5,r_5)\), which since \(j_5 < n\) would imply \(i_5 = j_5\), contradicting \(i_5 < j_5\).

Scenarios 6) and 10) cannot both hold, because that would imply \(r_{10} = j_6 < r_6 = i_{10} < n\), which would imply \(h'(r_{10},i_{10}) = h'(j_6,r_6) = h'(j_{10},r_{10})\). This in turn would imply one of two things. If \(j_{10} < n\), then \(h'(r_{10},i_{10}) = h'(j_{10},r_{10})\) would imply \(j_{10} = i_{10}\), contradicting \(i_{10} < j_{10}\). If on the other hand \(j_{10} = n\), then \(h'(r_{10},i_{10}) = h'(j_{10},r_{10})\) would imply \(1 + (r_{10} + i_{10} - 1 \mod n) = 1 + (r_{10} \mod n - 1)\). Recalling that \(r_{10} < r_6 \leq n - 1\), we would get \(i_{10} - 1 = 0 \mod n\). This would imply \(i_{10} = 1\), contradicting \(i_{10} > r_6 \geq 1\).
Scenarios 7) and 10) cannot both hold, because that would imply $r_7 = r_{10} < j_7 = i_{10} \leq n - 1$, which would imply $h'(i_{10}, r_{10}) = h'(j_7, r_7) = h'(j_{10}, r_{10})$. This in turn would imply one of two things. If $j_{10} < n$, then $h'(i_{10}, r_{10}) = h'(j_{10}, r_{10})$ would imply $r_{10} = j_{10}$, contradicting $i_{10} < j_{10}$. If, on the other hand, $j_{10} = n$, then $h'(i_{10}, r_{10}) = h'(j_{10}, r_{10})$ would imply $1 + (i_{10} + r_{10} - 1 \text{ mod } n) = 1 + (r_{10} \text{ mod } n - 1)$. Recalling that $r_{10} < j_7 = i_{10} \leq n - 1$, we would get $i_{10} - 1 = 0 \text{ mod } n$. This would imply $i_{10} = 1$, contradicting $i_{10} > r_{10} \geq 1$.

Scenarios 5) and 6) cannot both hold, because that would imply $j_6 = r_5 < j_5 = r_6 \leq n - 1$, and $h'(i_5, r_5) = h'(j_6, r_6)$, which would imply $h'(i_5, j_6) = h'(j_5, j_6)$, which since $j_6 \leq n - 2$ would imply $i_5 = j_5$, contradicting $i_5 < j_5$.

Scenarios 1) and 10) cannot both hold, because that would imply $r_1 = i_{10} > i_4 = r_{10} \geq 1$, which would imply $h'(r_{10}, i_{10}) = h'(i_1, r_1) = h'(j_{10}, r_{10}) = h'(j_1, r_4)$. This would imply one of two things. If $j_{10} < n$, and, recalling that $r_{10} \leq n$, this would imply $i_{10} = j_{10}$, contradicting $i_{10} < j_{10}$. If on the other hand, $j_{10} = n$, this would imply $1 + (r_{10} \text{ mod } n - 1) = 1 + (r_{10} + i_{10} - 1 \text{ mod } n)$, which would imply $i_{10} = 1$, contradicting $i_{10} > 1$.

Scenarios 4) and 6) cannot both hold, because that would imply $j_4 = j_6 < r_4 = r_6 < n$, which would imply $h'(i_4, r_4) = h'(j_6, r_6) = h'(j_4, r_4)$, which recalling that $j_4 < n$ would imply $i_4 = j_4$, contradicting $i_4 < j_4$.

Scenarios 4) and 7) cannot both hold, because that would imply $j_4 = r_7 < j_2 = r_4 < n$, which would imply $h'(i_4, r_4) = h'(r_7, r_7) = h'(j_4, r_4)$, which recalling that $j_4 < n$ would imply $i_4 = j_4$, contradicting $i_4 < j_4$.

Scenarios 4) and 10) cannot both hold, because that would imply $i_4 < j_4 = r_{10} < r_4 = i_{10} < j_{10}$, which would imply $h'(i_4, i_{10}) = h'(i_4, r_4) = h'(j_{10}, r_{10}) = h'(j_4, j_{10})$. This would imply one of two things. If $j_{10} < n$, this would imply $1 + (i_4 + i_{10} + 1 \text{ mod } n) = 1 + (j_4 + j_{10} + 1 \text{ mod } n)$, which would imply $(j_4 - i_4 + (j_{10} - i_{10}) = 0 \text{ mod } n$, which since $j_4 > i_4, j_{10} > i_{10}$ would imply $(j_4 - i_4 + (j_{10} - i_{10}) = n$. This in turn would imply $j_{10} = n + (i_{10} - j_4) + i_4 > n + i_4 > n$, since $i_{10} - j_4 > 0$, and $i_4 > 0$, contradicting $j_{10} \leq n$. If on the other hand $j_{10} = n$, this would imply $1 + (i_4 + i_{10} + 1 \text{ mod } n) = 1 + (j_4 \text{ mod } n - 1)$, which since $j_4 = r_{10} < r_4 \leq n - 1$ would imply $(j_4 - i_4 + (j_{10} - i_{10}) + 1 \text{ mod } n = 0$, which imply either $j_4 - i_4 - i_{10} + 1 = 0$ or $j_4 - i_4 - i_{10} + 1 = -n$. The first option would imply $j_4 = i_{10} + i_4 - 1 \geq i_{10}$, contradicting $j_4 < i_{10}$. The second option would imply $i_{10} = n + 1 + j_4 - i_4 > n$, contradicting $i_{10} < n$.

\[\Box\]

C Proof of Theorem 3

We will need the following hash function in this proof.

C.1 Special hash function

**Definition 8.** The map $H^a$ transforms a tuple $(i, j), 1 \leq i < j \leq n$, into either 2, 3 or 4 triples according to

\[(i, j) \mapsto H^a(i, j) = H^a_1(i, j) \oplus H^a_2(i, j), \tag{23}\]

where two tuples (resp. triples) are assumed duplicates iff all of their components agree and

\[H^a_1(i, j) = \begin{cases} \{ (i, n + 1, h(i)) \} & \text{if } j = n \land i = 1, \\ \{ (i, j, h(i)), (j, n + 1, h(i)) \} & \text{if otherwise}, \end{cases}\]

and

\[H^a_2(i, j) = \begin{cases} \{ (j, n + 1, h(j)) \} & \text{if } i = j = 1, \\ \{ (i, j, h(j)), (i, n + 1, h(j)) \} & \text{if } i < j = 1. \end{cases}\]

$h(\cdot)$ is also a function of $n$ but for simplicity we omit it in the notation. $h(\cdot)$ is defined as

\[h(i) = 1 + ((i - 2) \text{ mod } n).\]

\[\tag{24}\]

**Lemma 5.** Let $(a, b, c) \in H^a(i, j)$ for $1 \leq i < j \leq n$. Then, $1 \leq a < b \leq n + 1, 1 \leq c \leq n$, and $c \notin \{ a, b \}$. Furthermore, the set

\[
\bigoplus_{1 \leq i < j \leq n} H^a(i, j)
\]

\[\tag{25}\]
has no duplicates.

**Proof.** The fact that $1 \leq a < b \leq n + 1$ and that $1 \leq c \leq n$ is immediate. To see that $c \notin \{a, b\}$, we just need to notice that $h(i) \notin \{i, n + 1\}$ for $i \in [n]$. The fact that $h(i) \neq n + 1$ follows the range of $h$ being $[n]$. If we had $h(i) = i$, then we would have $(i - 2) \text{ mod } n = i - 1$, which is not possible.

To see that the set $\{1, 2, 3\}$ does not have duplicates, we just need to see that, starting from two different tuples, the different expressions that define the triples that go into $(25)$ can never be equal.

Given $1 \leq i < j \leq n$, $1 \leq i' < j' \leq n, (i, j) \neq (i', j')$ we will show that

1. $H_n(i, j)$ does not have duplicates;
2. $H_n(i, j)$ and $H_n(i', j')$ do not have overlaps, that is, $H_n^1(i, j), H_n^2(i, j), H_n^1(i', j'),$ and $H_n^2(i', j')$ do not have overlaps with each other.

It is obvious that $H_n^2(i, j)$ does not have duplicates and nor does $H_n^2(i, j)$ according to their definitions.

For 2., we show that the four sets have no overlaps with each other. We show this two sets at a time, there are in total 6 pairs to consider. As an immediate result of the discussion in 1., the following four combinations do not have overlaps: $H_n^2(i, j)$ vs. $H_n^2(i, j)$, $H_n^2(i', j')$ vs. $H_n^2(i', j')$, $H_n^2(i, j)$ vs. $H_n^2(i', j')$, $H_n^2(i', j')$ vs. $H_n^2(i', j')$. The two combinations left are $H_n^2(i, j)$ vs. $H_n^2(i', j')$ and $H_n^2(i', j')$ vs. $H_n^2(i, j)$. We notice that they are symmetric and, because the choice of the tuples $(i, j), (i', j')$ is arbitrary, we only need to show that $H_n^2(i, j)$ and $H_n^2(i', j')$ do not have overlaps, given $(i, j) \neq (i', j')$.

$H_n^2(i, j)$ and $H_n^2(i', j')$ each have two possibilities for the form of their output. Thus, together, there are four possibilities to consider. None of them have an overlap, which we show by contradiction.

1. $H_n^2(i, j) = \{(i, n + 1, h(i))\}$ and $H_n^2(i', j') = \{(i', n + 1, h(j'))\}$. If these single-element sets have an overlap, that implies that $i = i'$, but, according to the definition, $i = 1$ and $i' = i' - 1$ which implies $j' > 1$.
2. $H_n^2(i, j) = \{(i, n + 1, h(i))\}$ and $H_n^2(i', j') = \{(i', j', h(j'))\}$. For them to have an overlap, $h(i) = h(j')$. That requires $i = j'$ which contradictory to $i = 1$ and $i' < j' - 1$ at the same time.
3. $H_n^2(i, j) = \{(i, j, h(i))\}$ and $H_n^2(i', j') = \{(i', j', h(j'))\}$. For the first two components to equal, $i = i'$, $j = j'$, and $i = j'$, which is contradictory to $i' < j' - 1$. For the second two components to equal, $j = i'$ and $i = j'$, which is contradictory to $i < j$ or $i' < j'$. Because of the existence of “$n + 1$”, the components at different positions cannot collide.
4. $H_n^2(i, j) = \{(i, j, h(i))\}$ and $H_n^2(i', j') = \{(i', n + 1, h(j'))\}$. This implies $j' = j$ and $j' = i$, which is contradictory to $i < j$.

For example, if $n = 3$, then the possible tuples $(1, 2)$, and $(1, 3)$, and $(2, 3)$, get mapped respectively to $(1, 2, 3), (2, 4, 3), (2, 4, 1), (1, 4, 3), (1, 3, 2), (1, 4, 2), (2, 3, 1), (3, 4, 1), (3, 4, 2)$, all of which are different and satisfy the claims in Lemma 5.

We now prove the four metric properties in order. It is trivial to prove the first three properties given the definition of our distance function for the transport problem. Then, we provide a detailed proof for the triangle inequality.

**C.2 Non-Negativity**

**Proof.** The non-negativity of $d_{i,j}$ and $r_{i,j}$, implies that $\langle d_{i,j}, r_{i,j} \rangle \geq 0$, and hence that $W \geq 0$. □
C.3 Symmetry

Proof. Recall that the computation of $\mathcal{W}(p_{s:i:n})$ involves a set of distances $\{d_{a,b}\}_{a,b}$. Consider a generic permutation map $\sigma$, and let $\sigma^{-1}$ be its inverse. Let $\sigma$ and $\sigma^{-1}$ apply component-wise to its arguments. The computation of $\mathcal{W}(p_{s:i:n})$ involves a set of distances $\{d_{a,b}\}_{a,b}$ that satisfy $d_{i,j} = d_{\sigma^{-1}(i),\sigma^{-1}(j)}$. Therefore, each term $\left(\tilde{d}_{i,j}, r_{i,j}\right)$ involved in the computation of $\mathcal{W}(p_{s:i:n})$, can be rewritten as $\left(\tilde{d}_{\sigma^{-1}(i),\sigma^{-1}(j)}, r_{i,j}\right)$, which a simple reindexing of the summation $\sum_{i<j}$ allow us to write as $\left(d_{\sigma^{-1}(i),\sigma^{-1}(j)}, r_{i,j}\right)$. Since the mass function $r$ has as supporting sample space $\Omega_{\sigma(i)} \times \ldots \times \Omega_{\sigma(n)}$, the marginal $r_{\sigma(i)}$ can be seen as the marginal $q_{i,j}$ of a mass function $q$ with support $\Omega_{\sigma(i)} \times \ldots \times \Omega_{\sigma(n)}$. Therefore, minimizing $\sum_{i<j}\left(\tilde{d}_{i,j}, r_{i,j}\right)^{1/\ell}$ for $r$ over $\Omega_{\sigma(i)} \times \ldots \times \Omega_{\sigma(n)}$ is the same as minimizing $\sum_{i<j}\left(\langle d_{\sigma^{-1}(i),\sigma^{-1}(j)}, q_{i,j}\rangle\right)^{1/\ell}$ for $q$ over $\Omega_{\sigma(i)} \times \ldots \times \Omega_{\sigma(n)}$. $\Box$

C.4 Identity

Proof. We prove each direction of the equivalence separately. Recall that $\{p_i\}$ are given, they are the masses for which we want to compute the pairwise MMOT.

"$\Leftarrow$": If for each $i, j \in [n]$ we have $\Omega^i = \Omega^j$, then $m^i = m^j$, and there exists a bijection $b_{i,j}(-)$ from $[m^i]$ to $[m^j]$ such that $\Omega^j_s = \Omega_{b_{i,j}(s)}^i$ for all $s$. If furthermore $p^i = p^j$, we can define a $r$ for $\Omega^i \times \Omega^j$ such that its univariate marginal over $\Omega^i$, $r^i$, satisfies $r^i = p^i$, and such that its bivariate marginal over $\Omega^i \times \Omega^j$, $r_{i,j}$, satisfies $r_{i,j} = p_{i,j}$, if $t = b_{i,j}(s)$, and zero otherwise. Such a $r$ achieves an objective value of 0 in $\mathcal{F}$. Therefore, $\mathcal{W}^{1,\ldots,n} = 0$.

"$\Rightarrow$": Now let $r^*$ be a minimizer of $\mathcal{F}$ for $\mathcal{W}^{1,\ldots,n}$. Let $\{r^{s,i}\}$ and $\{r^{s,j}\}$ be its univariate and bivariate marginals respectively. If $\mathcal{W}^{1,\ldots,n} = 0$ then $\langle d_{i,j}, r^{s,i,j}\rangle_\ell = 0$ for all $i, j$. Let us consider a specific pair $i, j$, and, without loss of generality, let us assume that $m^i \leq m^j$. Since, by assumption, we have that $r^{s,i}_s = p^i_s > 0$ for all $s \in [m^i]$, and $r^{s,j}_s = p^j_s > 0$ for all $s \in [m^j]$, there exists an injection $b_{i,j}(-)$ from $[m^i]$ to $[m^j]$ such that $r^{s,i,j}_s > 0$ for all $s \in [m^i]$. Therefore, $\langle d_{i,j}, r^{s,i,j}_s\rangle_\ell = 0$ implies that $d^{s,i,j}_{b_{i,j}(s)} = 0$ for all $s \in [m^i]$. Therefore, since $d$ is a metric, it must be that $\Omega^i_s = \Omega^j_{b_{i,j}(s)}$ for all $s \in [m^i]$. Now lets us suppose that there exists an $r \in [m^j]$ that is not in the range of $b_{i,j}$. Since, by assumption, all of the elements of the sample spaces are different, it must be that $d^{s,i,j}_{l} > 0$ for all $s \in [m^i]$. Therefore, since $\langle d_{i,j}, r^{s,i,j}_s\rangle_\ell = 0$, it must be that $r^{s,i,j}_s = 0$ for all $s \in [m^i]$. This contradicts the fact that $\sum_{s \in [m^i]} r^{s,i,j}_s = r^{s,j}_r > 0$ (the last inequality being true by assumption). Therefore, $m^i = m^j$, and the existence of $b_{i,j}$ proves that $\Omega^i = \Omega^j$. The same time, since $d^{s,i,j}_{l} > 0$ for all $t \neq b_{i,j}(s)$, it must be that $r^{s,i,j}_{s,t} = 0$ for all $t \neq b_{i,j}(s)$. Therefore, $p^i_s = p^j_{b_{i,j}(s)}$ for all $s$, i.e. $p^i = p^j$. $\Box$

C.5 Generalized Triangle Inequality

Proof. Let $p^*$ be a minimizer for (the optimization problem associated with) $\mathcal{W}^{1,\ldots,n}$, and let $p^{s,i,j}$ be the marginal induced by $p^*$ for the sample space $\Omega^i \times \Omega^j$. We would normally use $r^*$ for this minimizer, but, to avoid confusions between $r$ and $r$, we avoid doing so. We can write that

$$\mathcal{W}^{1,\ldots,n} = \sum_{1 \leq i < j \leq n-1} \left(\langle d_{i,j}, p^{s,i,j}\rangle_\ell\right)^{\frac{1}{\ell}}.$$  (26)

For $r \in [n]$, let $p^{(s,r)}$ be a minimizer for $\mathcal{W}^{1,\ldots,n-1,r+1,\ldots,n+1}$. We would normally use $r^{(s,r)}$ for this minimizer, but, to avoid confusions between $r$ and $r$, we avoid doing so. For $i, j \in [m+1]\{r\}$, let $p^{(s,r)}_{i,j}$ be the marginal of $p^{(s,r)}$ for the sample space $\Omega^i \times \Omega^j$. Recall that since $p^{(s,r)}$ satisfies the constraints in $\mathcal{F}$, its marginal for the sample space $\Omega^i$ is $p^{s,i}$, which is given in advance.
Let \( h(\cdot) \) be the map defined as \( \text{(24)} \).

Define the following mass function for \( \Omega^1 \times \ldots \times \Omega^{n+1} \),
\[
q = \mathcal{G} \left( p^{n+1}, \{ (p^{(s+h(i))}\} \right),
\]
where \( p^{(s+h(i))} \) is defined as the mass function that satisfies \( p^{(s+h(i))} p^{n+1} = p^{(s+h(i))} \). Notice that since \( h(i) \not\in \{ i, n+1 \} \), the probability \( p^{(s+h(i))} \) exists for all \( i \in [n] \).

Let \( q_1 \times \ldots \times q_n \) be the marginal of \( q \) for sample space \( \Omega^1 \times \ldots \times \Omega^n \), and \( q_{i,j} \) be the marginal of \( q \) for \( \Omega^i \times \Omega^j \).

By Lemma 3, we know that the \( i \)-th univariate marginal of \( q \) is \( p^i \) (given) and hence \( q_1 \times \ldots \times q_n \) satisfies the constraints associated with \( \mathcal{W}^1, \ldots, n \). Therefore, we can write that
\[
\sum_{1 \leq i < j \leq n} \left( \left\langle d_{i,j}, p^{i,j} \right\rangle \right)^{\frac{1}{2}} \leq \sum_{1 \leq i < j \leq n} \left( \left\langle d_{i,j}, q^{i,j} \right\rangle \right)^{\frac{1}{2}}.
\]

By Lemma 4, inequality (a) below holds; because \( d \) is symmetric, (b) below holds; by the definition of \( q \), (c) below follows. Therefore,
\[
\left( \left\langle d_{i,j}, i,j \right\rangle \right)^{\frac{1}{2}} \leq \left( \left\langle d_{i,j}, n+1 \right\rangle, q^{n+1,j} \right)^{\frac{1}{2}} + \left( \left\langle d_{i,j}, q^{n+1}\right\rangle \right)^{\frac{1}{2}}
\]
\[
= \left( \left\langle d_{i,j}, n+1 \right\rangle, q^{n+1,j} \right)^{\frac{1}{2}} + \left( \left\langle d_{i,j}, q^{n+1}\right\rangle \right)^{\frac{1}{2}}
\]
\[
= \left( \left\langle d_{i,j}, p^{(s+h(i))} \right\rangle \right)^{\frac{1}{2}} + \left( \left\langle d_{i,j}, p^{(s+h(i))} \right\rangle \right)^{\frac{1}{2}}.
\]

Let \( w_{i,j} \) denote each term on the r.h.s. of (26), and \( w_{i,j,r} \) denote \( \left( \left\langle d_{i,j}, p^{(s+r)i,j} \right\rangle \right)^{\frac{1}{2}} \). Combining (28) - (30), we have
\[
\sum_{1 \leq i < j \leq n-1} w_{i,j} \leq \sum_{1 \leq i < j \leq n-1} w_{i,n,h(i)} + w_{j,n,h(j)}.
\]

Finally, we write
\[
\sum_{r=1}^{n} \mathcal{W}^{1,\ldots,r,1-r+1,\ldots,n+1} = \sum_{i,j \in [n] \setminus \{r\}, i < j} w_{i,j,r},
\]
and show that (32) upper-bounds the r.h.s of (31).

First, by Lemma 4 and the symmetry of \( d \), we have
\[
w_{i,n,h(i)} \leq w_{i,j,h(i)} + w_{j,n,h(j)}, \quad w_{j,n,h(j)} \leq w_{i,j,h(j)} + w_{i,n,h(j)},
\]
as long as for each triple \((a, b, c)\) in the above expressions, \( c \not\in \{a, b\} \). We will use these inequalities to upper bound some of the terms on the r.h.s. of (31), which can be further upper bounded by (32). In particular, we will apply inequalities (33) and (34) such that the terms \( w_{a,b,c} \) that we get after their use have triples \((a, b, c)\) that match the triples obtained via the map \( H^n \) defined in Section 4.2.

To be concrete, for example, if \( H^n \) maps \((i, j)\) to \{(i, n+1, h(i)), (j, n+1, h(j))\}, then we do not apply (33) and (34), and we leave \( w_{i,n+1,h(i)} + w_{j,n+1,h(j)} \) as is on the r.h.s. of (31). If, for example, \( H^n \) maps \((i, j)\) to \{(i, n+1, h(i)), (i, h(j)), (i, n+1, h(j))\}, then we leave the first term in \( w_{i,n+1,h(i)} + w_{j,n+1,h(j)} \) in the r.h.s. of (32) untouched, but we upper bound the second term using (32) to get \( w_{i,n+1,h(i)} + w_{i,j,h(j)} + w_{j,n+1,h(j)} \).

After proceeding in this fashion, and by Lemma 5, we know that all of the terms \( w_{a,b,c} \) that we obtain have triples \((a, b, c)\) with \( c \not\in \{a, b\} \), with \( c \in [n] \), and \( 1 \leq a < b \leq n+1 \). Therefore, these terms appear in (32). Also by Lemma 5, we know that we do not get each triple more than once. Therefore, the upper bound that we just constructed with the help of \( H^n \) for the r.h.s of (31) can be upper bounded by (32).
D Proof of upper bound in Theorem 4

Proof. Consider the following setup. Let \( m^t = m \) for all \( i \in [n] \), and \( \Omega^t_s \in \mathbb{R} \) for all \( i \in [n] \), \( s \in [m] \). Define \( d \) such that \( d^t_{s,i} = |\Omega^s_i - \Omega^j_i| \), if \( s = t \), and infinity otherwise. Let \( p^s_i = \frac{1}{m} \) for all \( i \in [n] \), \( s \in [m] \).

Any optimal solution \( r^* \) to the pairwise MMOT problem must have bivariate marginals that satisfy \( r^*_{s,i} = \frac{1}{m} \delta_{s,i} \), and thus \( \left( \left( d^t_{s,i}, r^*_{s,i} \right) \right)^{1/\ell} = \frac{1}{m} \cdot \| \Omega^s_i - \Omega^j_i \|_{\ell} \), where we interpret \( \Omega^s_i \) has a vector in \( \mathbb{R}^{m^n} \), and \( \| \cdot \|_\ell \) is the vector \( \ell \)-norm. Therefore, ignoring the factor \( \frac{1}{m} \), we only need to prove that \( C \) in Def. 4 holds with \( C(n) = n - 1 \) when \( W^1:1 \) is defined as \( \sum_{1 \leq i < j \leq n} \| \Omega^s_i - \Omega^j_i \|_{\ell} \). This in turn is a standard result, whose proof (in a more general form) can be found e.g. in Example 2.4 in [25]. \( \square \)

E Details of numerical experiments

Graphs are everywhere, and classifying and clustering graphs are important in diverse areas. For example, [36] clusters graphs of app’s code execution to find malware; [38] clusters graphs that represent chemical compounds to understand their anti-cancer and cancer-inducing characteristics; [39] represents text as word-based dependency trees and clusters them to classify cellphone reviews; [40] clusters graphs that represent the secondary structure of proteins; [37] reviews other graph clustering applications.

A powerful and general approach to clustering is distance-based clustering, a type of connectivity-based clustering: objects that are similar, according to a given distance measure, are put into the same cluster. Its use for graph clustering requires a measure of the distance between graphs. The purpose of our experiments is to illustrate via distance-based graph clustering i) the advantages of using MMOT over OT, and ii) the advantages of using an \( n \)-metric MMOT over a non-\( n \)-metric MMOT.

Graphs: We create 7 clusters, each with 10 graphs. Each graph is a random perturbation (edge addition/removal with \( p = 0.01 \)) of either 1) a complete graph, 2) a complete bipartite graph, 3) a cyclic chain, 4) a \( k \)-dimensional cube, 5) a \( K \)-hop lattice, 6) a periodic 2D grid, or 7) an Erdős–Rényi graph.

Vector data: We transform the graphs \( \{G_i^t\}_{i=1}^{70} \) into vectors \( \{v_i^t\}_{i=1}^{70} \) to be clustered. Each \( v_i^t \) is the (complex-valued) spectrum of a matrix \( M^t_i \) representing non-backtracking walks on \( G_i^t \), which approximates the length spectrum \( \mu^t \) of \( G_i^t \) [38]. Object \( \mu^t \) uniquely identifies the (2-core [34] of) \( G_i^t \) (up to an isomorphism) [35], but is too abstract to be used directly. Hence, we use its approximation \( v_i^t \). The length of \( v_i^t \) and \( v_j^t \) for equal-sized \( G_i^t \) and \( G_j^t \) can be different, depending on how we approximate \( \mu^t \). We use distance-based clustering and OT (multi) distances, since OT allows comparing objects of different lengths. Note that unlike the length spectrum, the classical spectrum of a graph (the eigenvalues of e.g. an adjacency matrix, Laplacian matrix, or random-walk matrix) has the advantage of having the same length for graphs with the same number of nodes. However, it does not uniquely identify a graph. For example, a star graph with 5 nodes and the graph that is the union of a square with an isolated node are co-spectral but are not isomorphic.

Distances: Each \( v_i^t \) is interpreted as a uniform distribution \( p_i^t \) over \( \Omega^t_i = \{v_{i,k}^t, k = 1, \ldots, 70\} \), the eigenvalues of \( M^t_i \). We compute a sampled version \( \hat{T}^A \) of the matrix \( T_A = \{W_{i,j}^t\}_{i,j \in [70]} \), where \( W_{i,j}^t \) is the WD between \( p_i^t \) and \( p_j^t \) using a \( d_{i,j}^t \) defined by \( d_{i,j}^t = |v_i^t - v_j^t| \). We compute a sampled version \( \hat{T}^B \) of the tensor \( T_B = \{W_{i,j,k}^t\}_{i,j,k \in [70]} \), where \( W_{i,j,k}^t \) is defined as in Def. 6 with \( d_{i,j}^t \) as for \( T_A \). We compute a sampled version \( \hat{T}^C \) of the tensor \( T_C \) with \( T_{i,j,k}^C = W_{i,j,k}^j \), where \( W_{i,j,k}^j \) is defined as in Thrm. 2 but now considering points in the complex plane. The sampled tensors \( \hat{T}^B \) and \( \hat{T}^C \) are built by randomly selecting 100 triples \( (i, j, k) \) and setting \( \hat{T}_{i,j,k}^B = T_{i,j,k}^B \) and \( \hat{T}_{i,j,k}^C = T_{i,j,k}^C \).

The non-sampled triples are given a very large value. The sampled matrix \( \hat{T}^A \) is built by sampling \( (3/2) \times 100 = 150 \) pairs \( (i, j) \) and setting \( \hat{T}_{i,j}^A = T_{i,j}^A \), and setting a large value for non-sampled pairs.

Clustering: We feed the distances \( \hat{T}^A \) to a spectral clustering algorithm [26] based on normalized random-walk Laplacians to produce one clustering solution \( C^A \), we feed the distances \( \hat{T}^B \) to the two hyper-graph-based clustering methods NH-Cut [27] and TTM [28] to produce clustering solutions \( C^{B1} \) and \( C^{B2} \) respectively, and we use \( \hat{T}^C \) and NH-Cut and TTM to produce clustering solutions \( C^{C1} \)
and $C^{C2}$. Both NH-Cut and TTM determine clusters by finding optimal cuts of a hyper-graph where each hyper-edge’s weight is the MMOT distance among three graphs. Both NH-Cut and TTM require a threshold that is used to prune the hyper-graph. Edges whose weight (multi-distance) is larger than a given threshold are removed. This threshold is tuned to minimize each clustering solution error. All clustering solutions output 7 clusters.

**Errors:** For each clustering solution, we compute the fraction of miss-classified graphs. In particular, if we use $C^x(i) = k$, $x = A, B1, B2, C1, C2$, to represent that clustering solution $C^x$ assigns graph $G^i$ to cluster $k$, then the error of this solution is

$$\min_{\sigma} \frac{1}{70} \sum_{i=1}^{70} I(C^{\text{ground truth}}(i) = C^x(\sigma(i)),$$

where the min is over all permutations $\sigma$, since the specific cluster IDs output by different algorithms have no real meaning. This experiment is repeated multiple times (random numbers being drawn independently among experiments) and the frequency of the errors are plotted in histograms in Fig. 2.

**References**


