algorithms

MDPI

Article

Fused Gromov-Wasserstein Distance for Structured Objects

Titouan Vayer [1,*], Laetitia Chapel [1], Remi Flamary [2], Romain Tavenard [3] and Nicolas Courty [1]

[1] CNRS, IRISA, Université Bretagne-Sud, F-56000 Vannes, France; laetitia.chapel@univ-ubs.fr (L.C.); nicolas.courty@irisa.fr (N.C.)
[2] CNRS, OCA Lagrange, Université Côte d'Azur, F-06000 Nice, France; remi.flamary@unice.fr
[3] CNRS, LETG, Université Rennes, F-35000 Rennes, France; romain.tavenard@univ-rennes2.fr
* Correspondence: titouan.vayer@irisa.fr

Received: 8 July 2020; Accepted: 26 August 2020; Published: 31 August 2020

Abstract: Optimal transport theory has recently found many applications in machine learning thanks to its capacity to meaningfully compare various machine learning objects that are viewed as distributions. The Kantorovitch formulation, leading to the Wasserstein distance, focuses on the features of the elements of the objects, but treats them independently, whereas the Gromov–Wasserstein distance focuses on the relations between the elements, depicting the structure of the object, yet discarding its features. In this paper, we study the Fused Gromov-Wasserstein distance that extends the Wasserstein and Gromov–Wasserstein distances in order to encode simultaneously both the feature and structure information. We provide the mathematical framework for this distance in the continuous setting, prove its metric and interpolation properties, and provide a concentration result for the convergence of finite samples. We also illustrate and interpret its use in various applications, where structured objects are involved.

Keywords: optimal transport; GRAPHS and Structured objects; Wasserstein and Gromov-Wasserstein distances

1. Introduction

We focus on the comparison of structured objects, i.e., objects defined by both a feature and a structure information. Abstractly, the feature information cover all the attributes of an object; for example, it can model the value of a signal when objects are time series, or the node labels over a graph. In shape analysis, the spatial positions of the nodes can be regarded as features, or, when objects are images, local color histograms can describe the image's feature information. As for the structure information, it encodes the specific relationships that exist among the components of the object. In a graph context, nodes and edges are representative of this notion, so that each label of the graph may be linked to some others through the edges between the nodes. In a time series context, the values of the signal are related to each other through a temporal structure. This representation can be related with the concept of relational reasoning (see [1]), where some entities (or elements with attributes, such as the intensity of a signal) coexist with some relations or properties between them (or some structure, as described above). Including structural knowledge about objects in a machine learning context has often been valuable in order to build more generalizable models. As shown in many contexts, such as graphical models [2,3], relational reinforcement learning [4], or Bayesian nonparametrics [5], considering objects as a complex composition of entities together with their interactions is crucial in order to learn from small amounts of data.

Unlike recent deep learning end-to-end approaches [6,7] that attempt to avoid integration of prior knowledge or assumptions about the structure wherever possible, *ad hoc* methods, depending

on the kind of structured objects involved, aim to build meaningful tools that include structure information in the machine learning process. In graph classification, the structure can be taken into account through dedicated graph kernels, in which the structure drives the combination of the feature information [8–10]. In a time series context, Dynamic Time Warping and related approaches are based on the similarity between the features while allowing limited temporal (i.e., structural) distortion in the time instants that are matched [11,12]. Closely related, an entire field has focused on predicting the structure as an output and it has been deployed on tasks, such as segmenting an image into meaningful components or predicting a natural language sentence [10,13,14].

All of these approaches rely on meaningful representations of the structured objects that are involved. In this context, distributions or probability measures can provide an interesting representation for machine learning data. This allows their comparison within the Optimal Transport (OT) framework that provides a meaningful way of comparing distributions by capturing the underlying geometric properties of the space through a cost function. When the distributions dwell in a common metric space (Ω, d), the Wasserstein distance defines a metric between these distributions under mild assumptions [15]. In contrast, the Gromov–Wasserstein distance [16,17] aims at comparing distributions that support live in different metric spaces through the intrinsic pair-to-pair distances in each space. Unifying both distances, the Fused Gromov–Wasserstein distance was proposed in a previous work in [18] and used in the discrete setting to encode, in a single OT formulation, both feature and structure information of structured objects. This approach considers structured objects as joint distributions over a common feature space associated with a structure space specific to each object. An OT formulation is derived by considering a tradeoff between the feature and the structure costs, respectively, defined with respect to the Wasserstein and the Gromov–Wasserstein standpoints.

This paper presents the theoretical foundations of this distance and states the mathematical properties of the FGW metric in the general setting. We first introduce a representation of structured objects using distributions. We show that classical Wasserstein and Gromov–Wasserstein distance can be used in order to compare either the feature information or the structure information of the structured object but that they both fail at comparing the entire object. We then present the Fused Gromov–Wasserstein distance in its general formulation and derive some of its mathematical properties. Particularly, we show that it is a metric in a given case, we give a concentration result, and we study its interpolation and geodesic properties. We then provide a conditional-gradient algorithm to solve the quadratic problem resulting from FGW in the discrete case and we conclude by illustrating and interpreting the distance in several applicative contexts.

Notations. Let $\mathcal{P}(\Omega)$ be the set of all probability measures on a space Ω and $\mathcal{B}(A)$ the set of all Borel sets of a σ-algebra A. We note # the push-forward operator, such, that for a measurable function T, $B \in \mathcal{B}(A)$, $T\#\mu(B) = \mu(T^{-1}(B))$.

We note $\text{supp}(\mu)$ the support of $\mu \in \mathcal{P}(\Omega)$ is the minimal closed subset $A \subset \Omega$ such that $\mu(\Omega \backslash A) = 0$. Informally, this is the set where the measure "is not zero".

For two probability measures $\mu \in \mathcal{P}(A)$ and $\nu \in \mathcal{P}(B)$ we note $\Pi(\mu, \nu)$ the set of all couplings or matching measures of μ and ν, i.e., the set $\{\pi \in P(\Omega \times \Omega) \,|\, \forall (A_0, B_0) \in \mathcal{B}(A) \times \mathcal{B}(B), \pi(A_0 \times B) = \mu(A_0), \pi(\Omega \times \Omega_0) = \nu(B_0)\}$.

For two metric spaces (X, d_X) and (Y, d_Y), we define the distance $d_X \oplus d_Y$ on $X \times Y$ such that, for $(x, y), (x', y') \in X \times Y$, $d_X \oplus d_Y((x, y), (x', y')) = d_X(x, x') + d_Y(y, y')$.

We note the simplex of N bins as $\Sigma_N = \{a \in (\mathbb{R}_+)^N, \sum_i a_i = 1\}$. For two histograms $a \in \Sigma_n$ and $b \in \Sigma_m$ we note with some abuses $\Pi(a, b)$ the set of all couplings of a and b, i.e., the set $\Pi(a, b) = \{\pi \in \mathbb{R}_+^{n \times m} | \sum_i \pi_{i,j} = b_j; \sum_j \pi_{i,j} = a_i\}$. Finally, for $x \in \Omega$, δ_x denotes the dirac measure in x.

Assumption. In this paper, we assume that all metric spaces are non-trivial Polish metric spaces (i.e., separable and completely metrizable topological spaces) and that all measures are Borel.

2. Structured Objects as Distributions and Fused Gromov–Wasserstein Distance

The notion of structured objects used in this paper is inspired from the discrete point of view where one aims at comparing labeled graphs. More formally, we consider undirected labeled graphs as tuples of the form $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \ell_f, \ell_s)$, where $(\mathcal{V}, \mathcal{E})$ are the set of vertices and edges of the graph. $\ell_f : \mathcal{V} \to \Omega$ is a labelling function that associates each vertex $v_i \in \mathcal{V}$ with a feature $a_i \overset{\text{def}}{=} \ell_f(v_i)$ in some feature metric space (Ω, d). Similarly, $\ell_s : \mathcal{V} \to X$ maps a vertex v_i from the graph to its structure representation $x_i \overset{\text{def}}{=} \ell_s(v_i)$ in some structure space (X, d_X) specific to each graph. $d_X : X \times X \to \mathbb{R}_+$ is a symmetric application which aims at measuring the similarity between the nodes in the graph. In the graph context, d_X can either encode the neighborhood information of the nodes, the edge information of the graph or more generally it can model a distance between the nodes such as the shortest path distance or the harmonic distance [19]. When d_X is a metric, such as the shortest-path distance, we naturally endow the structure with the metric space (X, d_X).

In this paper, we propose enriching the previous definition of a structured object with a probability measure which serves the purpose of signaling the relative importance of the object's elements. For example, we can add a probability (also denoted as weight) $(h_i)_i \in \Sigma_n$ to each node in the graph. This way, we define a fully supported probability measure $\mu = \sum_i h_i \delta_{(x_i, a_i)}$, which includes all the structured object information (see Figure 1 for a graphical depiction).

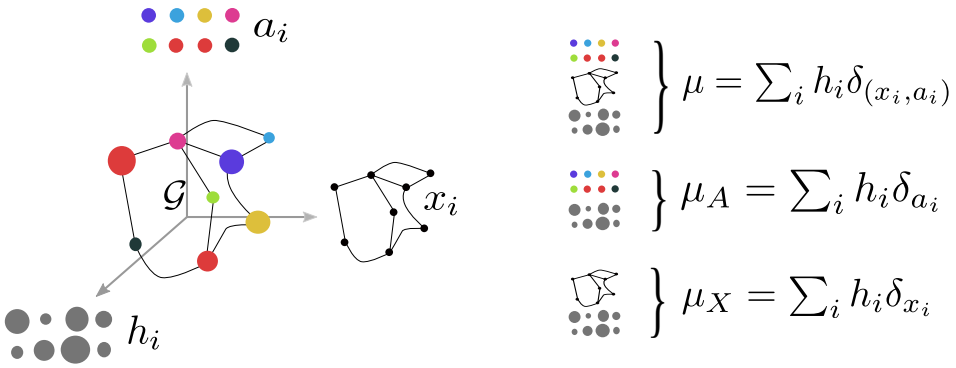

Figure 1. Discrete structured object (**left**) can be described by a labeled graph with $(a_i)_i$ the feature information of the object and $(x_i)_i$ the structure information. If we enrich this object with a histogram $(h_i)_i$ aiming at measuring the relative importance of the nodes, we can represent the structured object as a fully supported probability measure μ over the couple space of feature and structure with marginals μ_X and μ_A on the structure and the features respectively (**right**).

This graph representation for objects with a finite number of points/vertices can be generalized to the continuous case and leads to a more general definition of structured objects:

Definition 1 (Structured objects)**.** *A structured object over a metric space (Ω, d) is a triplet $(X \times \Omega, d_X, \mu)$, where (X, d_X) is a metric space and μ is a probability measure over $X \times \Omega$. (Ω, d) is denoted as the* feature space*, such that $d : \Omega \times \Omega \to \mathbb{R}_+$ is the distance in the feature space and (X, d_X) the* structure space*, such that $d_X : X \times X \to \mathbb{R}_+$ is the distance in the structure space. We will note μ_X and μ_A the structure and feature marginals of μ.*

Definition 2 (Space of structured objects)**.** *We note $\mathbb{X}$ the set of all metric spaces. The space of all structured objects over (Ω, d) will be written as $\mathbb{S}(\Omega)$ and is defined by all the triplets $(X \times \Omega, d_X, \mu)$ where $(X, d_X) \in \mathbb{X}$ and $\mu \in \mathcal{P}(X \times \Omega)$. To avoid finiteness issues in the rest of the paper we define for $p \in \mathbb{N}^*$ the space $\mathbb{S}_p(\Omega) \subset \mathbb{S}(\Omega)$ such that if $(X \times \Omega, d_X, \mu) \in \mathbb{S}_p(\Omega)$ we have:*

$$\int_\Omega d(a, a_0)^p d\mu_A(a) < +\infty \tag{1}$$

(the finiteness of this integral does not depend on the choice of a_0)

$$\int_{X\times X} d_X(x,x')^p d\mu_X(x) d\mu_X(x') < +\infty. \tag{2}$$

For the sake of simplicity, and when it is clear from the context, we will sometimes denote only by μ the whole structured object. The marginals μ_X, μ_A encode very partial information since they focus only on independent feature distributions or only on the structure. This definition encompasses the discrete setting discussed in above. More precisely, let us consider a labeled graph of n nodes with features $A = (a_i)_{i=1}^n$ with $a_i \in \Omega$ and $X = (x_i)_{i=1}^n$ the structure representation of the nodes. Let $(h_i)_{i=1}^n$ be an histogram, then the probability measure $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ defines structured object in the sense of Definition 1, since it lies in $\mathcal{P}(X \times \Omega)$. In this case, an example of μ, μ_X, and μ_A is provided in Figure 1.

Note that the set of structured objects is quite general and also allows considering discrete probability measures of the form $\mu = \sum_{i,j=1}^{p,q} h_{i,j} \delta_{(x_i,a_j)}$ with p, q possibly different than n. We propose focusing on a particular type of structured objects, namely the generalized labeled graphs, as described in the following definition:

Definition 3 (Generalized labeled graph)**.** *We call generalized labeled graph a structured object $(X \times \Omega, d_X, \mu) \in \mathbb{S}_p(\Omega)$ such that μ can be expressed as $\mu = (I \times \ell_f)\#\mu_X$ where $\ell_f : X \to \Omega$ is surjective and pushes μ_X forward to μ_A, i.e., $\ell_f \# \mu_X = \mu_A$.*

This definition implies that there exists a function ℓ_f, which associates a feature $a = l_f(x)$ to a structure point $x \in X$ and, as such, one structure point can not have two different features. The labeled graph described by $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ is a particular instance of a generalized labeled graph in which l_f is defined by $l_f(x_i) = a_i$.

2.1. Comparing Structured Objects

We now aim to define a notion of equivalence between two structured objects $(X \times \Omega, d_X, \mu)$ and $(Y \times \Omega, d_Y, \nu)$. We note in the following ν_Y, ν_B the marginals of ν. Intuitively, two structured objects are the same if they share the same feature information, if their structure information are lookalike, and if the probability measures are corresponding in some sense. In this section, we present mathematical tools for individual comparison of the elements of structured objects. First, our formalism implies comparing metric spaces, which can be done *via* the notion of isometry.

Definition 4 (Isometry)**.** *Let (X, d_X) and (Y, d_Y) be two metric spaces. An isometry is a surjective map $f : X \to Y$ that preserves the distances:*

$$\forall x, x' \in X, d_Y(f(x), f(x')) = d_X(x, x'). \tag{3}$$

An isometry is bijective, since for $f(x) = f(x')$ we have $d_Y(f(x), f(x')) = 0 = d_X(x, x')$ and hence $x = x'$ (in the same way f^{-1} is also a isometry). When it exists, X and Y share the same "size" and any statement about X, which can be expressed through its distance is transported to Y by the isometry f.

Example 1. *Let us consider the two following graphs whose discrete metric spaces are obtained as shortest path between the vertices (see corresponding graphs in Figure 2)*

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{pmatrix}}_{d_X(x_i,x_j)} \quad \textit{and} \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}}_{d_Y(y_i,y_j)}.$$

These spaces are isometric since the map f, such that $f(x_1) = y_1$, $f(x_2) = y_3$, $f(x_3) = y_4$, $f(x_4) = y_2$ verifies Equation (3).

Figure 2. Two isometric metric spaces. Distances between the nodes are given by the shortest path, and the weight of each edge is equal to 1.

The previous definition can be used in order to compare the structure information of two structured objects. Regarding the feature information, because they all lie in the same ambient space Ω, a natural way for comparing them is by the standard set equality $A = B$. Finally, in order to compare measures on different spaces, the notion of preserving map can be used.

Definition 5 (Measure preserving map)**.** *Let $\Omega_1, \mu_1 \in \mathcal{P}(\Omega_1)$ and $\Omega_2, \mu_2 \in \mathcal{P}(\Omega_2)$ be two measurable spaces. A function (usually called a map) $f : \Omega_1 \to \Omega_2$ is said to be measure preserving if it transports the measure μ_1 on μ_2 such that*

$$f\#\mu_1 = \mu_2.$$

If there exists such a measure preserving map, the properties about measures of Ω_1 are transported via f to Ω_2.

Combining these two ideas together leads to the notion of measurable metric spaces (often called *mm-spaces* [17]), i.e., a metric space (X, d_X) enriched with a probability measure and described by a triplet $(X, d_X, \mu_X \in \mathcal{P}(X))$. An interesting notion for comparing mm-spaces is the notion of isomorphism.

Definition 6 (Isomorphism)**.** *Two mm-spaces $(X, d_X, \mu_X), (Y, d_Y, \mu_Y)$ are isomorphic if there exists a surjective measure preserving isometry $f : supp(\mu_X) \to supp(\nu_Y)$ between the support of the measures μ_X, ν_Y.*

Example 2. *Let us consider two mm-spaces $(X = \{x_1, x_2\}, d_X = \{1\}, \mu_X = \{\frac{1}{2}, \frac{1}{2}\})$ and $(Y = \{y_1, y_2\}, d_Y = \{1\}, \mu_Y = \{\frac{1}{4}, \frac{3}{4}\})$, as depicted in Figure 3. These spaces are isometric, but not isomorphic, as there exists no measure preserving map between them.*

Figure 3. Two isometric but not isomorphic spaces.

All of this considered, we can now define a notion of equivalence between structured objects.

Definition 7 (Strong isomorphism of structured objects)**.** *Two structured objects are said to be strongly isomorphic if there exists an isomorphism I between the structures such that $f = (I, id)$ is bijective between $supp(\mu)$ and $supp(\nu)$ and measure preserving. More precisely, f satisfies the following properties:*

P.1 $f\#\mu = \nu$.

P.2 *The function f statisfies:*

$$\forall (x,a) \in supp(\mu), f(x,a) = (I(x), a). \tag{4}$$

P.3 *The function $I : supp(\mu_X) \to supp(\nu_Y)$ is surjective, satisfies $I\#\mu_X = \nu_Y$ and:*

$$\forall x, x' \in supp(\mu_X)^2, d_X(x,x') = d_Y(I(x), I(x')). \tag{5}$$

It is easy to check that the strong isomorphism defines an equivalence relation over $\mathbb{S}_p(\Omega)$.

Remark 1. *The function f described in this definition can be seen as a feature, structure, and measure preserving function. Indeed, from **P.1** f is measure preserving. Moreover, (X, d_X, μ_X) and (Y, d_Y, ν_Y) are isomorphic through I. Finally using **P.1** and **P.2** we have that $\mu_A = \nu_B$, so that the feature information is also preserved.*

Example 3. *To illustrate this definition, we consider a simple example of two discrete structured objects:*

$$\underbrace{\begin{pmatrix}(x_1,a_1)\\(x_2,a_2)\\(x_3,a_3)\\(x_4,a_4)\end{pmatrix}}_{x_i,a_i}, \underbrace{\begin{pmatrix}0&1&1&1\\1&0&1&2\\1&1&0&2\\1&2&2&0\end{pmatrix}}_{d_X(x_i,x_j)}, \underbrace{\begin{pmatrix}1/4\\1/4\\1/4\\1/4\end{pmatrix}}_{h_i} \quad \textit{and} \quad \underbrace{\begin{pmatrix}(y_1,b_1)\\(y_2,b_2)\\(y_3,b_3)\\(y_4,b_4)\end{pmatrix}}_{y_i,b_i}, \underbrace{\begin{pmatrix}0&1&1&1\\1&0&2&2\\1&2&0&1\\1&2&1&0\end{pmatrix}}_{d_Y(y_i,y_j)}, \underbrace{\begin{pmatrix}1/4\\1/4\\1/4\\1/4\end{pmatrix}}_{h'_i}$$

*with for i, $a_i = b_i$ and for $i \neq j$, $a_i \neq a_j$ (see Figure 4). The two structured objects have isometric structures and same features individually, but they are not strongly isomorphic. One possible map $f = (f_1, f_2)$, such that f_1 leads to an isometry is $f(x_1,a_1) = (y_1,b_1)$, $f(x_2,a_2) = (y_3,b_3)$, $f(x_3,a_3) = (y_4,b_4)$, $f(x_4,a_4) = (y_2,b_2)$. Yet, this map does not satisfy $f_2(x,.) = I_d$ for any x, since $f(x_2,a_2) = (y_3,b_3)$ and $a_2 \neq b_3$. The other possible functions, such that f_1 leads to an isometry are simply permutations of this example, yet it is easy to check that none of them verifies **P.2** (for example, with $f(x_2,a_2) = (y_4,b_4)$).*

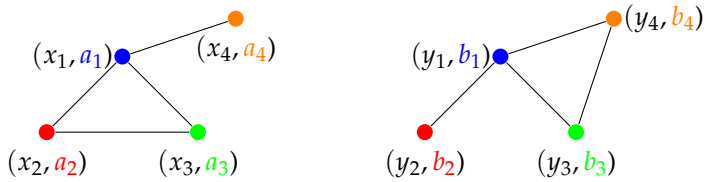

Figure 4. Two structured objects with isometric structures and identical features that are not strongly isomorphic. The color of the nodes represent the node feature and each edge represents a distance of 1 between the connected nodes.

2.2. Background on OT Distances

The Optimal Transport (OT) framework defines distances between probability measures that describe either the feature or the structure information of structured objects.

Wasserstein distance. The classical OT theory aims at comparing probability measures $\mu_A \in \mathcal{P}(\Omega), \nu_B \in \mathcal{P}(\Omega)$. In this context the quantity:

$$d_{W,p}(\mu_A, \nu_B) = \left(\inf_{\pi \in \Pi(\mu_A, \nu_B)} \int_{\Omega \times \Omega} d(a,b)^p d\pi(a,b) \right)^{\frac{1}{p}} \tag{6}$$

is usually called the p-Wasserstein distance (also known, for $p = 1$, as Earth Mover's distance [20] in the computer vision community) between distributions μ_A and ν_B. It defines a distance on probability measures, especially $d_{W,p}(\mu_A, \nu_B) = 0$ *iff* $\mu_A = \nu_B$. This distance also has a nice geometrical interpretation as it represents an optimal cost (*w.r.t.* d) to move the measure μ_A onto ν_B with $\pi(a, b)$ the amount of probability mass shifted from a to b (see Figure 5). To this extent, the Wasserstein distance quantifies how "far" μ_A is from ν_B by measuring how "difficult" it is to move all the mass from μ_A onto ν_B. Optimal transport can deal with smooth and discrete measures and it has proved to be very useful for comparing distributions in a shared space, but with different (and even non-overlapping) supports.

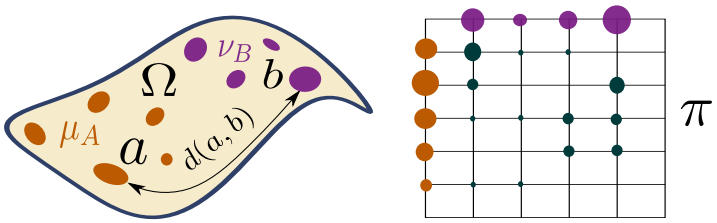

Figure 5. Example of a coupling between two discrete measures on the same ground space equipped with a distance d that will define the Wasserstein distance. (**Left**): the discrete measures on Ω. (**Right**): one possible coupling between these measures that conserves the mass. Image adapted from [21] [Figure 2.6].

Gromov–Wasserstein distance. In order to compare measures whose support are not necessarily in the same ambient space [16,17] define a new OT distance. By relaxing the classical Hausdorff distance [15,17], authors build a distance over the space of all mm-spaces. For two compact mm-spaces $(X, d_X, \mu_X \in \mathcal{P}(X))$ and $(Y, d_Y, \nu_Y \in \mathcal{P}(Y))$, the Gromov–Wasserstein distance is defined as:

$$d_{GW,p}(\mu_X, \nu_Y) = \left(\inf_{\pi \in \Pi(\mu_X, \nu_Y)} \int_{X \times Y \times X \times Y} L(x, y, x', y')^p d\pi(x, y) d\pi(x', y') \right)^{\frac{1}{p}} \tag{7}$$

where:

$$L(x, y, x', y') = |d_X(x, x') - d_Y(y, y')|$$

The Gromov–Wasserstein distance depends on the choice of the metrics d_X and d_Y and with some abuse of notation we denote the entire mm-space by its probability measure. When it is not clear from the context, we will specify using $d_{GW,p}(d_X, d_Y, \mu_X, \nu_Y)$. The resulting coupling tends to associate pairs of points with similar distances within each pair (see Figure 6). The Gromov–Wasserstein distance allows for the comparison of measures over different ground spaces and defines a metric over the space of all mm-spaces quotiented by the isomoprhisms (see Definitions 4 and 5). More precisely, it vanishes if the two mm-spaces are isomorphic. This distance has been used in the context of relational data e.g., in shape comparison [17,22], deep metric alignment [23], generative modelling [24] or to align single-cell multi-omics datasets [25].

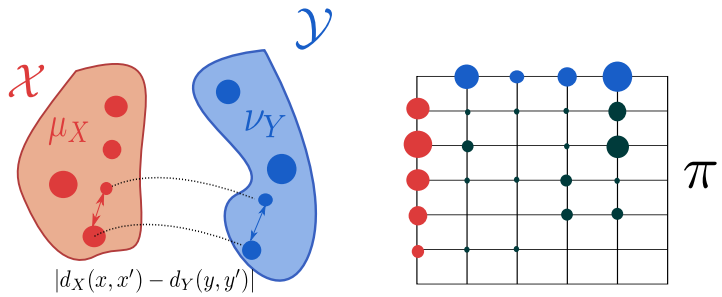

Figure 6. Gromov–Wasserstein coupling of two mm-spaces $\mathcal{X} = (X, d_X, \mu_X)$ and $\mathcal{Y} = (Y, d_Y, \nu_Y)$. **Left**: the mm-spaces have nothing in common. Similarity between pairwise distances is measured by $|d_X(x,x') - d_Y(y,y')|$. **Right**: an admissible coupling of μ_X and ν_Y. Image adapted from [21] [Figure 2.6].

2.3. Fused Gromov–Wasserstein Distance

Building on both Gromov–Wasserstein and Wasserstein distances, we define the Fused Gromov–Wasserstein (*FGW*) distance on the space of structured objects:

Definition 8 (Fused Gromov-Wasserstein distance)**.** *Let* $\alpha \in [0,1]$ *and* $p, q \geq 1$. *We consider* $(X \times \Omega, d_X, \mu) \in \mathbb{S}_{pq}(\Omega)$ *and* $(Y \times \Omega, d_Y, \nu) \in \mathbb{S}_{pq}(\Omega)$. *The Fused-Gromov–Wasserstein distance is defined as:*

$$d_{FGW,\alpha,p,q}(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} E_{p,q,\alpha}(\pi) \right)^{\frac{1}{p}} \tag{8}$$

where

$$E_{p,q,\alpha}(\pi) = \int_{(X\times\Omega\times Y\times\Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\right)^p d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b'))$$

Figure 7 illustrates this definition. When it is clear from the context we will simply note d_{FGW} instead for $d_{FGW,\alpha,p,q}$ for brevity. α acts as a trade-off parameter between the cost of the structures represented by $L(x,y,x',y')$ and the feature cost $d(a,b)$. In this way, the convex combination of both terms leads to the use of both information in one formalism resulting on a single map π that "moves" the mass from one joint probability measure to the other.

Many desirable properties arise from this definition. Among them, one can define a topology over the space of structured objects using the *FGW* distance to compare structured objects, in the same philosophy as for Wasserstein and Gromov–Wasserstein distances. The definition also implies that *FGW* acts as a generalization of both Wasserstein and Gromov-Wasserstein distances, with *FGW* achieving an interpolation between these two distances. More remarkably, *FGW* distance also realizes geodesic properties over the space of structured objects, allowing the definition of gradient flows. All of these properties are detailed in the next section. Before reviewing them, we first compare *FGW* with *GW* and *W* (by assuming for now that *FGW* exists, which will be shown later in Theorem 1).

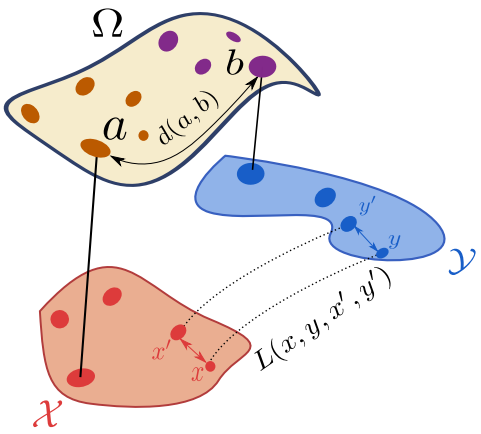

Figure 7. Illustration of Definition 8. The figure shows two structured objects $(X \times \Omega, d_X, \mu)$ and $(Y \times \Omega, d_Y, \mu)$. The feature space Ω is the common space for all features. The two metric spaces (X, d_X) and (Y, d_Y) represent the structures of the two structured objects, the similarity between all pair-to-pair distances of the structure points is measured by $L(x, y, x', y')$. μ and ν are the joint measures on the structure space and the feature space.

Proposition 1 (Comparaison between *FGW*, *GW* and *W*)**.** *We have the following results for two structured objects μ and ν:*

- *The following inequalities hold:*

$$d_{FGW,\alpha,p,q}(\mu,\nu) \geq (1-\alpha) d_{W,pq}(\mu_A, \nu_B)^q \tag{9}$$

$$d_{FGW,\alpha,p,q}(\mu,\nu) \geq \alpha d_{GW,pq}(\mu_X, \nu_Y)^q \tag{10}$$

- *Let us suppose that the structure spaces (X, d_X),(Y, d_Y) are part of a single ground space (Z, d_Z) (i.e., $X, Y \subset Z$ and $d_X = d_Y = d_Z$). We consider the Wasserstein distance between μ and ν for the distance on $Z \times \Omega$: $\tilde{d}((x,a),(y,b)) = (1-\alpha) d(a,b) + \alpha d_Z(x,y)$. Then:*

$$d_{FGW,\alpha,p,1}(\mu,\nu) \leq 2 d_{W,p}(\mu,\nu). \tag{11}$$

Proof of this proposition can be found in Section 7.1. In particular, following this proposition, when the *FGW* distance vanishes then both *GW* and *W* distances vanish so that the structure and the feature of the structure object are individually "the same" (with respect to their corresponding equivalence relation). However, the converse is not necessarily true, as shown in the following example.

Example 4 (Toy trees)**.** *We construct two trees as illustrated in Figure 8 where the 1D node features are shown with colors. The shortest path between the nodes is used to capture the structures of the two structured objects and the Euclidean distance is used for the features. We consider uniform weights on all nodes. Figure 8 illustrates the differences between FGW, GW, and W distances. The left part is the Wasserstein coupling between the features: red nodes are transported on red ones and the blue nodes on the blue ones but tree structures are completely discarded. In this case, the Wasserstein distance vanishes. In the right part, we compute the Gromov–Wasserstein distance between the structures: all couples of points are transported to another couple of points, which enforces the matching of tree structures without taking into account the features. Because structures are isometric, the Gromov–Wasserstein distance is null. Finally, we compute the FGW using intermediate α (center), the bottom and first level structure is preserved as well as the feature matching (red on red and blue on blue) and FGW discriminates the two structured objects.*

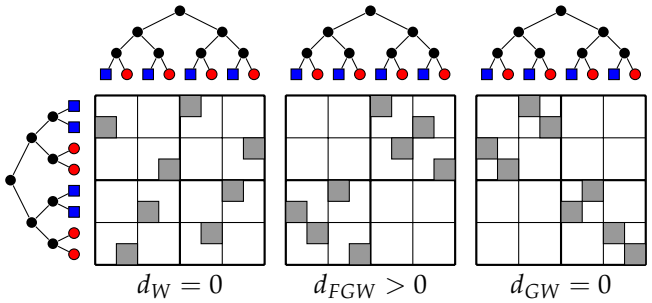

Figure 8. Difference on transportation maps between *FGW*, *GW*, and *W* distances on synthetic trees. (**Left**) the *W* distance between the features is null since feature information are the same, (**Middle**) *FGW* with $\alpha \in]0,1[$ is different from zero and discriminate the two structured objects, and (**Right**) *GW* between the two isometric structures is null.

3. Mathematical Properties of *FGW*

In this section, we establish some mathematical properties of the *FGW* distance. The first result relates to the existence of the *FGW* distance and the topology of the space of structured objects. We then prove that the *FGW* distance is indeed a distance regarding the equivalence relation between structured objects, as defined in Defintion 7, allowing us to derive a topology on $\mathbb{S}(\Omega)$.

3.1. Topology of the Structured Object Space

The *FGW* distance has the following properties:

Theorem 1 (Metric properties). *Let $p, q \geq 1$, $\alpha \in]0,1[$ and $\mu, \nu \in \mathbb{S}_{pq}(\Omega)$. The functional $\pi \to E_{p,q,\alpha}(\pi)$ always achieves an infimum π^* in $\Pi(\mu,\nu)$* s.t. *$d_{FGW,\alpha,p,q}(\mu,\nu) = E_{p,q,\alpha}(\pi^*) < +\infty$. Moreover:*

- *$d_{FGW,\alpha,p,q}$ is symmetric and, for $q = 1$, satisfies the triangle inequality. For $q \geq 2$, the triangular inequality is relaxed by a factor 2^{q-1}.*
- *For $\alpha \in]0,1[$, $d_{FGW,\alpha,p,q}(\mu,\nu) = 0$ if an only if there exists a bijective function $f = (f_1, f_2) : supp(\mu) \to supp(\nu)$ such that:*

$$f\#\mu = \nu \tag{12}$$

$$\forall (x,a) \in supp(\mu)\ , f_2(x,a) = a \tag{13}$$

$$\forall (x,a),(x',a') \in supp(\mu)^2,\ d_X(x,x') = d_Y(f_1(x,a), f_1(x',a')) \tag{14}$$

- *If (μ,ν) are generalized labeled graphs then $d_{FGW,\alpha,p,q}(\mu,\nu) = 0$ if and only if $(X \times \Omega, d_X, \mu)$ and $(Y \times \Omega, d_Y, \nu)$ are strongly isomorphic.*

Proof of this theorem can be found in Section 7.2. The identity of indiscernibles is the most delicate part to prove and it is based on using the Gromov–Wasserstein distance between the spaces $X \times \Omega$ and $Y \times \Omega$. The previous theorem states that *FGW* is a distance over the space of generalized labeled graphs endowed with the strong isomorphism as equivalence relation defined in Definition 7. More generally, for any structured objects the equivalence relation is given by (12)–(14). Informally, invariants of the *FGW* are structured objects that have both the same structure and the same features in the same place. Despite the fact that $q = 1$ leads to a proper metric, we will further see in Section 4.1 that the case $q = 2$ can be computed more efficiently using a separability trick from [26].

Theorem 1 allows a wide set of applications for *FGW*, such as *k*-nearest-neighbors, distance-substitution kernels, pseudo-Euclidean embeddings, or representative-set methods [27–29]. Arguably, such a distance allows for a better interpretation than to end-to-end learning machines, such as neural networks, because the π matrix exhibits the relationships between the elements of the objects in a pairwise comparison.

3.2. Can We Adapt W and GW for Structured Objects?

Despite the appealing properties of both Wasserstein and Gromov–Wasserstein distances, they fail at comparing structured objects by focusing only on the feature and structure marginals, respectively. However, with some hypotheses, one could adapt these distances for structured objects.

Adapting Wasserstein. If the structure spaces (X, d_X) and (Y, d_Y) are part of a same ground space (Z, d_Z), i.e., ($X, Y \subset Z$ and $d_X = d_Y = d_Z$), one can build a distance $\hat{d} = d_Z \oplus d$ between couples (x, a) and (y, b) and apply the Wasserstein distance, so as to compare the two structured objects. In this case, when the Wasserstein distance vanishes it implies that the structured objects are equal in the sense $\mu = \nu$. This approach is very related with the one discussed in [30], where the authors define the Transportation L^p distance for signal analysis purposes. Their approach can be viewed as a transport between two joint measures $\mu(X \times \Omega) = \lambda(\{(x, f(x)) \mid x \in X \subset Z = \mathbb{R}^d;\ f(x) \in A \subset \mathbb{R}^m\})$, $\nu(Y \times \Omega) = \lambda(\{(y, g(y)) \mid y \in Y \subset Z = \mathbb{R}^d;\ g(y) \in B \subset \mathbb{R}^m\})$ for function $f, g : Z \to \mathbb{R}^m$ representative of the signal values and λ the Lebesgue measure. The distance for the transport is defined as $\hat{d}((x, f(x)), (y, g(y))) = \frac{1}{\alpha}\|x - y\|_p^p + \|f(x) - g(y)\|_p^p$ for $\alpha > 0$ and $\|\cdot\|_p$ the l_p norm. In this case, $f(x)$ and $g(y)$ can be interpreted as encoding the feature information of the signal, while x, y encode its structure information. This approach is very interesting, but cannot be used on structured objects, such as graphs that will not share a common structure embedding space.

Adapting Gromov-Wasserstein. The Gromov–Wasserstein distance can also be adapted to structured objects by considering the distances $(1 - \beta)d_X \oplus \beta d$ and $(1 - \beta)d_Y \oplus \beta d$ within each space $X \times \Omega$ and $Y \times \Omega$, respectively, and $\beta \in]0, 1[$. When the resulting GW distance vanishes, structured objects are isomorphic with respect to $(1 - \beta)d_X \oplus \beta d$ and $(1 - \beta)d_Y \oplus \beta d$. However, the strong isomorphism is stronger than this notion, since the isomorphism allows for "permuting the labels", but not the strong isomorphism. More precisely, we have the following lemma:

Lemma 1. *Let $(X \times \Omega, d_X, \mu), (Y \times \Omega, d_Y, \nu)$ be two structured objects and $\beta \in]0, 1[$.*

If $(X \times \Omega, d_X, \mu)$ and $(Y \times \Omega, d_Y, \nu)$ are strongly isomorphic then $(X \times \Omega, (1 - \beta)d_X \oplus \beta d, \mu)$ and $(Y \times \Omega, (1 - \beta)d_Y \oplus \beta d, \nu)$ are isomorphic. However the converse is not true in general.

Proof. To see this, if we consider f as defined in Theorem 1, then, for $(x, a), (x', b) \in (\mathrm{supp}(\mu))^2$, we have $d_X(x, x') = d_Y(I(x), I(x'))$. In this way:

$$(1 - \beta)d_X(x, x') + \beta d(a, b) = (1 - \beta)d_Y(I(x), I(x')) + \beta d(a, b) \tag{15}$$

which can be rewritten as:

$$(1 - \beta)d \oplus \beta d_X((x, a), (x', b)) = (1 - \beta)d \oplus \beta d_Y(f(x, a), f(x', b)) \tag{16}$$

and so f is an isometry with respect to $(1 - \beta)d \oplus \beta d_X$ and $(1 - \beta)d \oplus \beta d_Y$. Because f is also measure preserving and surjective $(X \times \Omega, (1 - \beta)d_X \oplus \beta d, \mu)$ and $(Y \times \Omega, (1 - \beta)d_Y \oplus \beta d, \nu)$ are isomorphic. □

However, the converse is not necessarily true, as it is easy to cook up an example with the same structure but with permuted labels, so that objects are isomorphic but not strongly isomorphic. For example, in the tree example Figure 4, the structures are isomorphic and the distances between the features within each space are the same between each structured objects, so that $(X \times \Omega, (1 - \beta)d_X \oplus \beta d, \mu)$ and $(Y \times \Omega, (1 - \beta)d_Y \oplus \beta d, \nu)$ are isomorphic, yet not strongly isomorphic, as shown in the example since $FGW > 0$.

3.3. Convergence of Structured Objects

The metric property naturally endows the structured object space with a notion of convergence, as described in the next definition:

Definition 9. *Convergence of structured objects.*
Let $\left((X_n \times A_n, d_{X_n}, \mu_n)\right)_{n\in\mathbb{N}}$ *be a sequence of structured objects. It converges to* $(X \times \Omega, d_X, \mu)$ *in the Fused Gromov–Wasserstein sense if:*

$$\lim_{n\to\infty} d_{FGW,\alpha,p,1}(\mu_n, \mu) = 0 \tag{17}$$

Using Proposition 1, it is straightforward to see that if the sequence converges in the *FGW* sense, both the features and the structure converge respectively in the Wasserstein and Gromov–Wasserstein sense (see [17] for the definition of convergence in the Gromov–Wasserstein sense).

An interesting question arises from this definition. Let us consider a structured object $(X \times \Omega, d_X, \mu)$ and let us sample the joint distribution so as to consider $(\{(x_i, a_i)\}_{i\in\{1,..,n\}}, d_X, \mu_n)_{n\in\mathbb{N}}$ with $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i,a_i}$ where $(x_i, a_i) \in X \times \Omega$ are sampled from μ. Does this sequence converges to $(X \times \Omega, d_X, \mu)$ in the *FGW* sense and how fast is the convergence?

This question can be answered thanks to a notion of "size" of a probability measure. For the sake of conciseness, we will not exhaustively present the theory, but the reader can refer to [31] for more details. Given a measure μ on Ω, we denote as $dim_p^*(\mu)$ its upper Wasserstein dimension. It coincides with the intuitive notion of "dimension" when the measure is sufficiently well behaved. For example, for any absolutely continuous measure μ with respect to the Lebesgue measure on $[0,1]^d$, we have $dim_p^*(\mu) = d$ for any $p \in [1, \frac{d}{2}]$. Using this definition and the results presented in [31], we answer the question of convergence of finite sample in the following theorem (proof can be found in Section 7.3):

Theorem 2. *Convergence of finite samples and a concentration inequality*
Let $p \geq 1$. *We have:*

$$\lim_{n\to\infty} d_{FGW,\alpha,p,1}(\mu_n, \mu) = 0.$$

Moreover, suppose that $s > d_p^*(\mu)$. *Then there exists a constant C that does not depend on n such that:*

$$\mathbb{E}[d_{FGW,\alpha,p,1}(\mu_n, \mu)] \leq Cn^{-\frac{1}{s}}. \tag{18}$$

The expectation is taken over the i.i.d samples (x_i, a_i). *A particular case of this inequality is when* $\alpha = 1$ *so that we can use the result above to derive a concentration result for the Gromov-Wasserstein distance. More precisely, if* $\nu_n = \frac{1}{n}\sum_i \delta_{x_i}$ *denotes the empirical measure of* $\nu \in \mathcal{P}(X)$ *and if* $s' > d_p^*(\nu)$, *we have:*

$$\mathbb{E}[d_{GW,p}(\nu_n, \nu)] \leq C' n^{-\frac{1}{s'}}. \tag{19}$$

This result is a simple application of the convergence of finite sample properties of the Wasserstein distance, since in this case μ_n and μ are part of the same ground space so that (18) derive naturally from (11) and the properties of Wasserstein. In contrast to the Wasserstein distance case, this inequality is not necessarily sharp and future work will be dedicated to the study of its tightness.

3.4. Interpolation Properties between Wasserstein and Gromov-Wasserstein Distances

FGW distance is a generalization of both Wasserstein and Gromov–Wasserstein distances in the sense that it achieves an interpolation between them. More precisely, we have the following theorem:

Theorem 3. *Interpolation properties.*
As α *tends to zero, one recovers the Wasserstein distance between the features information and as* α *goes to one, one recovers the Gromov–Wasserstein distance between the structure information:*

$$\lim_{\alpha\to 0} d_{FGW,\alpha,p,q}(\mu, \nu) = (d_{W,qp}(\mu_A, \nu_B))^q \tag{20}$$

$$\lim_{\alpha \to 1} d_{FGW,\alpha,p,q}(\mu,\nu) = (d_{GW,qp}(\mu_X,\nu_Y))^q \tag{21}$$

Proof of this theorem can be found in Section 7.4.

This result shows that *FGW* can revert to one of the other distances. In machine learning, this allows for a validation of the α parameter to better fit the data properties (i.e., by tuning the relative importance of the feature *vs* structure information). One can also see the choice of α as a representation learning problem and its value can be found by optimizing a given criterion.

3.5. Geodesic Properties

One desirable property in OT is the underlying geodesics defined by the mass transfer between two probability distributions. These properties are useful in order to define the dynamic formulation of OT problems. This dynamic point of view is inspired by fluid dynamics and found its origin in the Wasserstein context with [32]. Various applications in machine learning can be derived from this formulation: interpolation along geodesic paths were used in computer graphics for color or illumination interpolations [33]. More recently, Ref. [34] used Wasserstein gradient flows in an optimization context, deriving global minima results for non-convex particles gradient descent. In [35], the authors used Wasserstein gradient flows in the context of reinforcement learning for policy optimization.

The main idea of this dynamic formulation is to describe the optimal transport problem between two measures as a curve in the space of measures minimizing its total length. We first describe some generality about geodesic spaces and recall classical results for dynamic formulation in both Wasserstein and Gromov–Wasserstein contexts. In a second part, we derive new geodesic properties in the *FGW* context.

Geodesic spaces. Let (X,d) be a metric space and x,y two points in X. We say that a curve $w : [0,1] \to X$ joining the *endpoints* x and y (i.e., with $w(0) = x$ and $w(1) = y$) is a constant speed geodesic if it satisfies $d(w(t),w(s)) \leq |t-s|d(w(0),w(1)) = |t-s|d(x,y)$ for $t,s \in [0,1]$. Moreover, if (X,d) is a length space (i.e., if the distance between two points of X is equal to the infimum of the lengths of the curves connecting these two points) then the converse is also true and a constant speed geodesic satisfies $d(w(t),w(s)) = |t-s|d(x,y)$. It is easy to compute distances along such curves, as they are directly embedded into $\mathbb{R}$.

In the Wasserstein context, if the ground space is a complete separable, locally compact length space, and if the endpoints of the geodesic are given, then there exists a geodesic curve. Moreover, if the transport between the endpoints is unique, then there is a unique displacement interpolation between the endpoints (see Corollary 7.22 and 7.23 in [15]). For example, if the ground space is $\mathbb{R}^d$ and the distance between the points is measured via the ℓ_2 norm, then geodesics exist and are uniquely determined (this can be generalized to strictly convex costs). In the Gromov–Wasserstein context, there always exists constant speed geodesics as long as the endpoints are given. These geodesics are unique modulo the isomorphism equivalence relation (see [16]).

The *FGW* case. In this paragraph, we suppose that $\Omega = \mathbb{R}^d$. We are interested in finding a geodesic curve in the space of structured objects i.e., a constant speed curve of structured objects joining two structured objects. As for Wasserstein and Gromov–Wasserstein, the structured object space endowed with the Fused Gromov–Wasserstein distance maintains some geodesic properties. The following result proves the existence of such a geodesic and characterizes it:

Theorem 4. *Constant speed geodesic.*

Let $p \geq 1$ and $(X \times \Omega, d_X, \mu_0)$ and $(Y \times \Omega, d_Y, \mu_1)$ in $\mathbb{S}_p(\mathbb{R}^d)$. Let π^* be an optimal coupling for the Fused Gromov–Wasserstein distance between μ_0, μ_1, and $t \in [0,1]$. We equip $\mathbb{R}^d$ with the ℓ_m norm for $m \geq 1$.

We define $\eta_t : X \times \Omega \times Y \times \Omega \to X \times Y \times \Omega$ *such that:*

$$\forall (x,a),(y,b) \in X \times \Omega \times Y \times \Omega,\ \eta_t(x,a,y,b) = (x,y,(1-t)a + tb) \tag{22}$$

Then:

$$(X \times Y \times \Omega, (1-t)d_X \oplus t d_Y, \mu_t = \eta_t \# \pi^*)_{t\in[0,1]} \tag{23}$$

is a constant speed geodesic connecting $(X \times \Omega, d_X, \mu_0)$ *and* $(Y \times \Omega, d_Y, \mu_1)$ *in the metric space* $\left(\mathbb{S}_p(\mathbb{R}^d), d_{FGW,\alpha,p,1}\right)$.

Proof of the previous theorem can be found in Section 7.5. In a sense, this result combines the geodesics in the Wasserstein space and in the space of all mm-spaces, since it suffices to interpolate the distances in the structure space and the features to construct a geodesic. The main interest is that it defines the minimum path between two structured objects. For example, when considering two discrete structured objects represented by the measures $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ and $\nu = \sum_{j=1}^m g_j \delta_{(y_j,b_j)}$, the interpolation path is given for $t \in [0,1]$ by the measure $\mu_t = \sum_{i=1}^n \sum_{j=1}^m \pi^*(i,j)\delta_{(x_i,y_j,(1-t)a_i+tb_j)}$ where π^* is an optimal coupling for the *FGW* distance. However this geodesic is difficult to handle in practice, since it requires the computation of the cartesian product $X \times Y$. To overcome this obstacle, an extension using theFrÃ©chet mean is defined in Section 4.2. The proper definition and properties of velocity fields associated to this geodesic is postponed to further works.

4. *FGW* in the Discrete Case

In practice, structured objects are often discrete and can be defined using the labeled graph formalism described previously. In this section, we discuss how to compute *FGW* efficiently. We also provide an algorithm for the computation of Fréchet means.

4.1. FGW in the Discrete Case

We consider two structured objects $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ and $\nu = \sum_{j=1}^m g_j \delta_{(y_j,b_j)}$, as described previously. We note M_{AB} the matrix $M_{AB}(i,j) = d(a_i,b_j)$ and C_1, C_2 the matrices $C_1(i,k) = d_X(x_i,x_k)$, $C_2(j,l) = d_Y(y_j,y_l)$ (see Figure 9). The Fused Gromov–Wasserstein distance is defined as

$$d_{FGW,\alpha,p,q}(\mu,\nu) = \left(\min_{\pi \in \Pi(h,g)} E_{p,q,\alpha}(\pi) \right)^{\frac{1}{p}} \tag{24}$$

where:

$$E_{p,q,\alpha}(\pi) = \sum_{i,j,k,l} \left((1-\alpha)M_{AB}(i,j)^q + \alpha |C_1(i,k) - C_2(j,l)|^q\right)^p \pi_{i,j}\, \pi_{k,l}.$$

Solving the related Quadratic Optimization problem. Equation (24) is clearly a quadratic problem *w.r.t.* π. Note that, despite the apparent $\mathcal{O}(m^2n^2)$ complexity of computing the tensor product, one can simplify the sum to complexity $\mathcal{O}(mn^2 + m^2n)$ [26] when considering $p = 1, q = 2$. In this case, the *FGW* computation problem can be re-written as finding π^*, such that:

$$\pi^* = \underset{\pi \in \Pi(h,g)}{\arg\min} \quad \mathrm{vec}(\pi)^T Q(\alpha) \mathrm{vec}(\pi) + \mathrm{vec}(D(\alpha))^T \mathrm{vec}(\pi) \tag{25}$$

where $Q = -2\alpha C_2 \otimes_K C_1$ and $D(\alpha) = (1-\alpha)M_{AB}$. $\otimes_K$ denote the Kronecker product of two matrices, vec the column-stacking operator. With such form, the resulting optimal map can be seen as a quadratic regularized map from initial Wasserstein [36,37]. However, unlike these approaches, we have a

quadratic but probably non-convex term. The gradient G that arises from Equation (24) can be expressed with the following partial derivative *w.r.t.* π:

$$G = (1-\alpha)M_{AB}^{q} + 2\alpha \left(\sum_{k,l} |C_1(i,k) - C_2(j,l)|^q \pi_{k,l} \right)_{i,j} \tag{26}$$

Solving a large scale QP with a classical solver can be computationally expensive. In [36], the authors propose a solver for a graph regularized optimal transport problem whose resulting optimization problem is also a QP. We can then directly use their conditional gradient defined in Algorithmic 1 to solve our optimization problem. It only needs at each iteration to compute the gradient in Equation (26) and solve a classical OT problem for instance with a network flow algorithm. The line-search part is a constrained minimization of a second degree polynomial function that is adapted to the non convex loss in Algorithmic 2. While the problem is non convex, conditional gradient is known to converge to a local stationary point [38]. When C_1, C_2 are squared Euclidean distance matrices the problem reduces to a concave graph matching so that the line-search of the conditional gradient Algorithm 1 always leads to 1 [39,40].

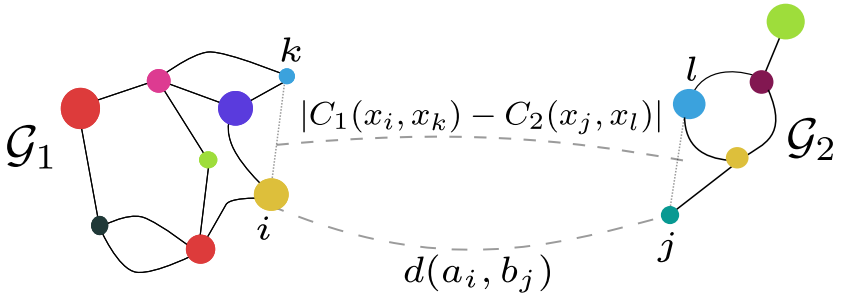

Figure 9. *FGW* loss for a coupling π depends on both a similarity between each feature of each node of each graph $(d(a_i, b_j))_{i,j}$ and between all intra-graph structure similarities $(|C_1(i,k) - C_2(j,l)|)_{i,j,k,l}$.

Algorithm 1 Conditional Gradient (CG) for *FGW*

$\pi^{(0)} \leftarrow \mu_X \mu_Y^\top$
for $i = 1, \ldots,$ **do**
  $G \leftarrow$ Gradient from Equation (26) *w.r.t.* $\pi^{(i-1)}$
  $\tilde{\pi}^{(i)} \leftarrow$ Solve OT with ground loss G
  $\tau^{(i)} \leftarrow$ Line-search for loss (24) with $\tau \in (0,1)$ using Algorithm 2
  $\pi^{(i)} \leftarrow (1-\tau^{(i)})\pi^{(i-1)} + \tau^{(i)}\tilde{\pi}^{(i)}$
end for

Algorithm 2 Line-search for CG ($q = 2$)

c_{C_1,C_2} from Equation (6) in [26]
$a = -2\alpha \langle C_1 \tilde{\pi}^{(i)} C_2, \tilde{\pi}^{(i)} \rangle$
$b = \langle (1-\alpha) M_{AB} + \alpha c_{C_1,C_2}, \tilde{\pi}^{(i)} \rangle - 2\alpha \left(\langle C_1 \tilde{\pi}^{(i)} C_2, \pi^{(i-1)} \rangle + \langle C_1 \pi^{(i-1)} C_2, \tilde{\pi}^{(i)} \rangle \right)$
if $a > 0$ **then**
  $\tau^{(i)} \leftarrow \min(1, \max(0, \frac{-b}{2a}))$
else
  $\tau^{(i)} \leftarrow 1$ if $a + b < 0$ else $\tau^{(i)} \leftarrow 0$
end if

4.2. Structured Optimal Transport Barycenter

An interesting use of the *FGW* distance is to define a barycenter of a set of structured data as a Fréchet mean. In that context, one seeks the structured object that minimizes the sum of the (weighted) *FGW* distances with a given set of objects. OT barycenters have many desirable properties and applications [26,41], yet no formulation can leverage both structural and feature information in the barycenter computation. Here, we propose to use the *FGW* distance to compute the barycenter of a set of K structured objects $(\mu_k)_k \in \mathbb{S}(\Omega)^K$ associated with structures $(C_k)_k$, features $(B_k)_k$ and base histograms $(h_k)_k$.

We suppose that the feature space is $\Omega = (\mathbb{R}^d, \ell_2)$ and $p = 1$. For simplicity, we assume that the base histograms and the histogram h associated to the barycenter are known and fixed. Note that it could be also included in the optimization process, as suggested in [26].

In this context, for a fixed $N \in \mathbb{N}$ and $(\lambda_k)_k$, such that $\sum_k \lambda_k = 1$, we aim to find:

$$\min_{\mu} \sum_k \lambda_k d_{FGW,\alpha,1,q}(\mu, \mu_k) = \min_{C,\ A \in \mathbb{R}^{N\times d}, (\pi_k)_k} \sum_k \lambda_k E_{1,q,\alpha}(M_{AB_k}, C, C_k, \pi_k) \tag{27}$$

Note that this problem is convex *w.r.t.* C and A, but not *w.r.t.* π_k. Intuitively, looking for a barycenter means finding feature values supported on a fixed size support, and the structure that relates them. Interestingly enough, there are several variants of this problem, where features or structure can be fixed for the barycenter. Solving the related simpler optimization problems extend straightforwardly.

Solving the barycenter problem with Block Coordinate Descent (BCD). We propose minimizing Equation (27) using a BCD algorithm, i.e., iteratively minimizing with respect to the couplings π_k, the structure metric C and the feature vector A.The minimization of this problem *w.r.t.* $(\pi_k)_k$ is equivalent to computing K independent Fused Gromov–Wasserstein distances using the algorithm presented above. We suppose that the feature space is $\Omega = (\mathbb{R}^d, \ell_2)$ and we consider $q = 2$. Minimization *w.r.t.* C in this case has a closed form (see Prop. 4 in [26]):

$$C \leftarrow \frac{1}{hh^T} \sum_k \lambda_k \pi_k^T C_k \pi_k \tag{28}$$

where h is the histogram of the barycenter and the division is computed pointwise. Minimization *w.r.t.* A can be computed with (Equation (8) in [42]):

$$A \leftarrow \sum_k \lambda_k B_k \pi_k^T \operatorname{diag}\left(\frac{1}{h}\right) . \tag{29}$$

5. Examples and Applications

In this section we derive some applications of *FGW* in graph contexts such as classification, clustering and coarsening of graphs.

5.1. Illustrations of FGW

In this section, we present several applications of *FGW* as a distance between structured objects and provide an interpretation of the OT matrix.

Example with one-dimensional (1D) features and structure spaces. Figure 10 illustrates the differences between Wasserstein, Gromov–Wasserstein, and Fused Gromov–Wasserstein couplings π^*. In this example, both the feature and structure space are one-dimensional (Figure 10 left). The feature space (vertical axis) denotes two clusters among the elements of both objects illustrated in the OT matrix M_{AB} , the structure space (horizontal axis) denotes a noisy temporal sequence along the indexes illustrated in the matrices C_1 and C_2 (Figure 10 center). Wasserstein respects the clustering, but forgets the temporal structure, Gromov–Wasserstein respects the structure, but do not take the clustering into account. Only FGW retrieves a transport matrix respecting both feature and structure.

Example on two simple images. We extract a 28×28 image from the MNIST dataset and generate a second one through translation or mirroring of the digit in the original image. We use pixel gray levels of the pixels as the features, and the structure is defined as the city-block distance on the pixel coordinate grid. We use equal weights for all of the pixels in the image. Figure 11 shows the different couplings obtained when only considering either the features, the structure only or both information. *FGW* aligns the pixels of the digits, recovering the correct order of the pixels, while both Wassertein and Gromov–Wasserstein distances fail at providing a meaningful transportation map. Note that in the Wasserstein and Gromov-Wasserstein case, the distances are equal to 0, whereas *FGW* manages to spot that the two images are different. Additionally, note that, in the *FGW* sense, the original digit and its mirrored version are also equivalent as there exists an isometry between their structure spaces, making *FGW* invariant to rotations or flips in the structure space in this case.

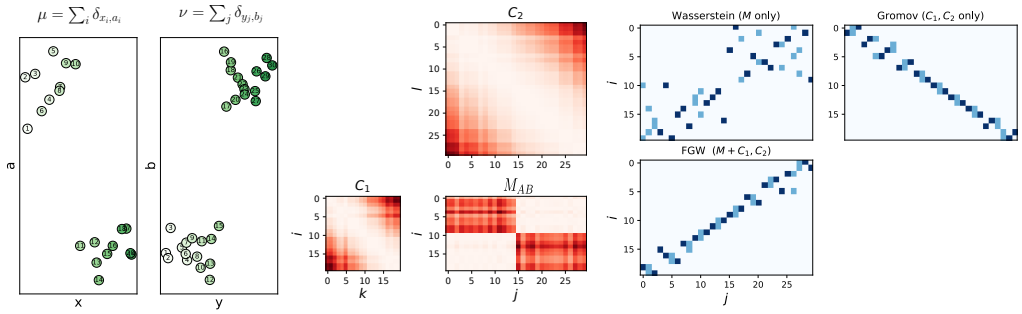

Figure 10. Illustration of the difference between *W*, *GW*, and *FGW* couplings. (**left**) Empirical distributions μ with 20 samples and ν with 30 samples which color is proportional to their index. (**middle**) Cost matrices in the feature (M_{AB}) and structure domains (C_1, C_2) with similar samples in white. (**right**) Solution for all methods. Dark blue indicates a non zero coefficient of the transportation map. Feature distances are large between points laying on the diagonal of M_{AB} such that Wasserstein maps is anti-diagonal but unstructured. Fused Gromov–Wasserstein incorporates both feature and structure maps in a single transport map.

Example on time series data. One of the main assets of *FGW* is that it can be used on a wide class of objects and time series are one more example of this. We consider here 25 monodimensional time series composed of two humps in $[0, 1]$ with random uniform height between 0 and 1. The signals are distributed according to two classes translated from each other with a fixed gap. The *FGW* distance is computed by considering d as the euclidean distance between the features of the signals (here the value of the signal in each point) and d_X and d_Y as the euclidean distance between timestamps.

A two-dimensional (2D) embedding is computed from a *FGW* distance matrix between a number of examples in this dataset with multidimensional scaling (MDS) [43] in Figure 12 (top). One can clearly see that the representation with a reasonable α value in the center is the most discriminant one. This can be better understood by looking as the OT matrices between the classes. Figure 12 (bottom) illustrates the behavior of *FGW* on one pair of examples when going from Wasserstein to Gromov–Wasserstein. The black line depicts the matching provided by the transport matrix and one can clearly see that while Wasserstein on the left assigns samples completely independently to their temporal position, the Gromov-Wasserstein on the right tends to align perfectly the samples (note that it could have reversed exactly the alignment with the same loss), but discards the values in the signal. Only the true *FGW* in the center finds a transport matrix that both respects the time sequences and aligns similar values in the signals.

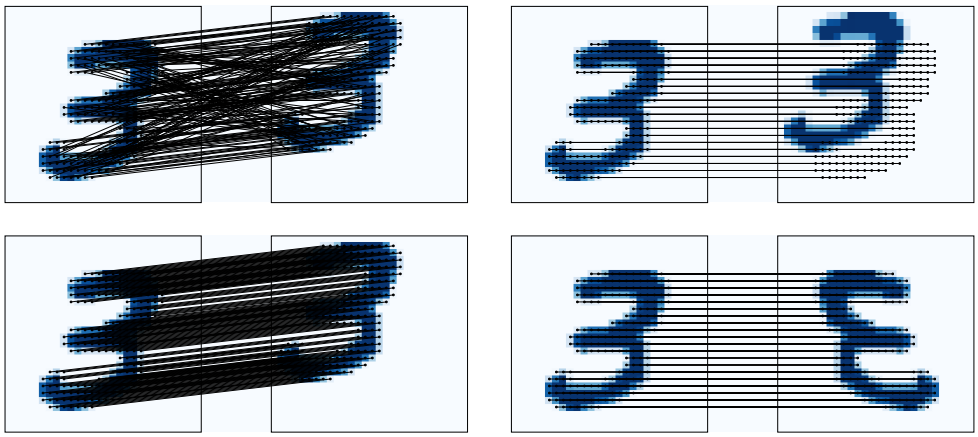

Figure 11. Couplings obtained when considering (**Top left**) the features only, where we have $d^{\Omega}_{W,1} = 0$ (**Top right**) the structure only, with $d_{GW,1} = 0$ (**Bottom left**, and **right**) both the features and the structure, with $d^{\Omega}_{FGW,0.1,1,2}$. For readibility issues, only the couplings starting from non white pixels on the left picture are depicted.

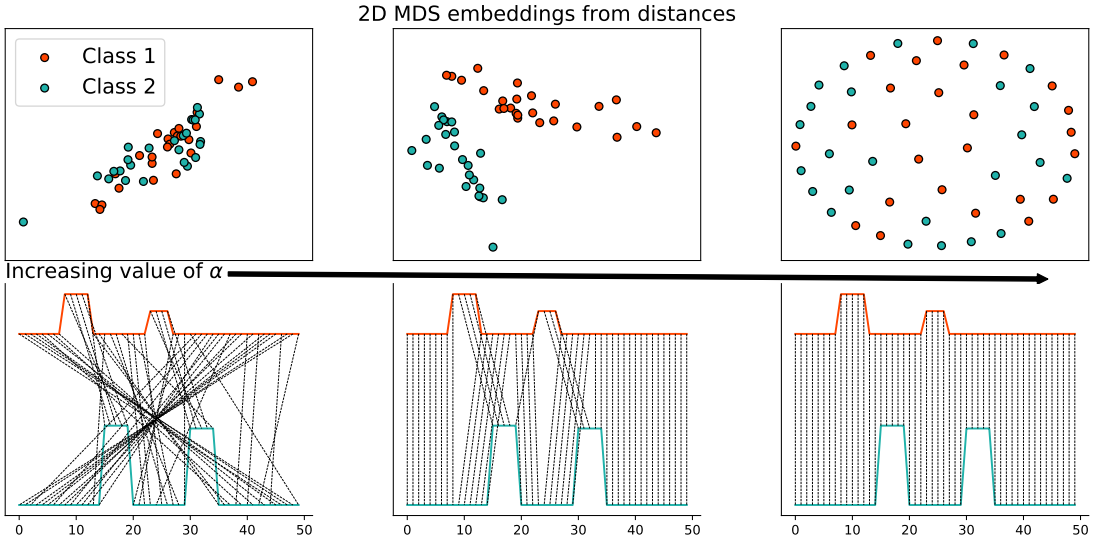

Figure 12. Behavior of trade-off parameter α on a toy time series classification problem. α is increasing from left ($\alpha = 0$: Wasserstein distance) to right ($\alpha = 1$: Gromov–Wasserstein distance). (**top row**) Two-dimensional (2D)-embedding is computed from the set of pairwise distances between samples with multidimensional scaling (MDS) (**bottom row**) illustration of couplings between two sample time series from opposite classes.

5.2. Graph-Structured Data Classification

In this section, we adress the question of training a classifier for graph data and evaluate the FGW distance used in a kernel with SVM.

Learning problem and datasets. We consider 12 widely used benchmark datasets divided into 3 groups. BZR, COX2 [44], PROTEINS, ENZYMES [45], CUNEIFORM [46] and SYNTHETIC [47] are vector attributed graphs. MUTAG [48], PTC-MR [49], and NCI1 [50] contain graphs with discrete attributes derived from small molecules. IMDB-B, IMDB-M [51] contain unlabeled graphs derived from social networks. All of the datasets are available in [52].

Experimental setup. Regarding the feature distance matrix M_{AB} between node features, when dealing with real valued vector attributed graphs, we consider the ℓ_2 distance between the labels of the vertices. In the case of graphs with discrete attributes, we consider two settings: in the first

one, we keep the original labels (denoted as RAW); we also consider a Weisfeiler-Lehman labeling (denoted as WL) by concatenating the labels of the neighbors. A vector of size H is created by repeating this procedure H times [49,53]. In both cases, we compute the feature distance matrix by using $d(a_i, b_j) = \sum_{k=0}^{H} \delta(\tau(a_i^k), \tau(b_j^k))$ where $\delta(x, y) = 1$ if $x \neq y$ else $\delta(x, y) = 0$ and $\tau(a_i^k)$ denotes the concatenated label at iteration k (for $k = 0$ original labels are used). Regarding the structure distances C, they are computed by considering a shortest path distance between the vertices.

For the classification task, we run a SVM using the indefinite kernel matrix $e^{-\gamma FGW}$, which is seen as a noisy observation of the true positive semidefinite kernel [54]. We compare classification accuracies with the following state-of-the-art graph kernel methods: (SPK) denotes the shortest path kernel [45], (RWK) the random walk kernel [55], (WLK) the Weisfeler Lehman kernel [53], and (GK) the graphlet count kernel [56]. For real valued vector attributes, we consider the HOPPER kernel (HOPPERK) [47] and the propagation kernel (PROPAK) [57]. We build upon the GraKel library [58] to construct the kernels and C-SVM to perform the classification. We also compare *FGW* with the PATCHY-SAN framework for CNN on graphs (PSCN) [9] building on our own implementation of the method.

To provide a comparison between the methods, most of the papers about graph classification usually perform a nested cross validation (using nine-fold for training, one for testing, and reporting the average accuracy of this experiment repeated 10 times), and report accuracies of the other methods taken from the original papers. However, these comparisons are not fair because of the high variance (on most datasets) *w.r.t.* the folds chosen for training and testing. This is why, in our experiments, the nested cross validation is performed on the same folds for training and testing for all methods [59]. In the result, Tables 1–3, we write in bold the best score for each dataset and we add a (*) when the best score does not yield to a significative improvement (based on a Wilcoxon signed rank test on the test scores) compared to the second best one. Note that, because of their small sizes, we repeat the experiments 50 times for MUTAG and PTC-MR datasets. For all methods using SVM, we cross validate the parameter $C \in \{10^{-7}, 10^{-6}, \ldots, 10^{7}\}$. The range of the WL parameter H is $\{0, 1 \ldots, 10\}$, and we also compute this kernel with H fixed at 2, 4. The decay factor λ for RWK $\{10^{-6}, 10^{-5} \ldots, 10^{-2}\}$, for the GK kernel we set the graphlet size $\kappa = 3$ and cross validate the precision level ϵ and the confidence δ as in the original paper [56]. The $t_{\max}$ parameter for PROPAK is chosen within $\{1, 3, 5, 8, 10, 15, 20\}$. For PSCN, we choose the normalized betweenness centrality as labeling procedure and cross validate the batch size in $\{10, 15, \ldots, 35\}$ and number of epochs in $\{10, 20, \ldots, 100\}$. Finally, for *FGW*, γ is cross validated within $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ and α is cross validated *via* a logspace search in $[0, 0.5]$ and symmetrically $[0.5, 1]$ (15 values are drawn).

Vector attributed graphs. The average accuracies reported in Table 1 show that FGW is a clear state-of-the-art method and performs best on four out of six datasets with performances in the error bars of the best methods on the other two datasets. The results for CUNEIFORM are significantly below those from the original paper [46], which can be explained by the fact that the method in this paper uses a graph convolutional approach specially designed for this dataset and that experiment settings are different. In comparison, the other competitive methods are less consistent, as they exhibit some good performances on some datasets only.

Discrete labeled graphs. We first note in Table 2 that *FGW* using WL attributes outperforms all competitive methods, including *FGW* with raw features. Indeed, the WL attributes allow for more finely encoding the neighborood of the vertices by stacking their attributes, whereas FGW with raw features only consider the shortest path distance between vertices, not their sequence of labels. This result calls for using meaningful feature and/or structure matrices in the FGW definition, which can be dataset-dependant, in order to enhance the performances. We also note that *FGW* with WL attributes outperforms the WL kernel method, highlighting the benefit of an optimal transport-based distance over a kernel-based similarity. Surprisingly, the results of PSCN are significantly lower than those from the original paper. We believe that it comes from the difference between the fold assignment for training and testing, which suggests that PSCN is difficult to tune.

Table 1. Average classification accuracy on the graph datasets with vector attributes.

Vector Attributes	BZR	COX2	CUNEIFORM	ENZYMES	PROTEIN	SYNTHETIC
FGW sp	**85.12 ± 4.15** *	77.23 ± 4.86	**76.67 ± 7.04**	71.00 ± 6.76	**74.55 ± 2.74**	**100.00 ± 0.00**
HOPPERK	84.15 ± 5.26	**79.57 ± 3.46**	32.59 ± 8.73	45.33 ± 4.00	71.96 ± 3.22	90.67 ± 4.67
PROPAK	79.51 ± 5.02	77.66 ± 3.95	12.59 ± 6.67	**71.67 ± 5.63** *	61.34 ± 4.38	64.67 ± 6.70
PSCN k = 10	80.00 ± 4.47	71.70 ± 3.57	25.19 ± 7.73	26.67 ± 4.77	67.95 ± 11.28	**100.00 ± 0.00**
PSCN k = 5	82.20 ± 4.23	71.91 ± 3.40	24.81 ± 7.23	27.33 ± 4.16	71.79 ± 3.39	**100.00 ± 0.00**

Non-attributed graphs. The particular case of the GW distance for graph classification is also illustrated on social datasets, which contain no labels on the vertices. Note that no artificial features were added to these datasets. Accuracies reported in Table 3 show that it greatly outperforms the SPK and GK graph kernel methods. This is, to the best of our knowledge, the first application of the Gromov–Wasserstein distance for social graph classification and it highlights the fact that *GW* is a good metric for comparing the graph structures.

Table 2. Average classification accuracy on the graph datasets with discrete attributes.

Discrete Attr.	MUTAG	NCI1	PTC-MR
FGW raw sp	83.26 ± 10.30	72.82 ± 1.46	55.71 ± 6.74
FGW wl h = 2 sp	86.42 ± 7.81	85.82 ± 1.16	63.20 ± 7.68
FGW wl h = 4 sp	**88.42 ± 5.67**	**86.42 ± 1.63**	**65.31 ± 7.90**
GK k = 3	82.42 ± 8.40	60.78 ± 2.48	56.46 ± 8.03
RWK	79.47 ± 8.17	58.63 ± 2.44	55.09 ± 7.34
SPK	82.95 ± 8.19	74.26 ± 1.53	60.05 ± 7.39
WLK	86.21 ± 8.48	85.77 ± 1.07	62.86 ± 7.23
WLK h=2	86.21 ± 8.15	81.85 ± 2.28	61.60 ± 8.14
WLK h = 4	83.68 ± 9.13	85.13 ± 1.61	62.17 ± 7.80
PSCN k = 10	83.47 ± 10.26	70.65 ± 2.58	58.34 ± 7.71
PSCN k = 5	83.05 ± 10.80	69.85 ± 1.79	55.37 ± 8.28

Table 3. Average classification accuracy on the graph datasets with no attributes.

Without Attribute	IMDB-B	IMDB-M
GW sp	**63.80 ± 3.49**	**48.00 ± 3.22**
GK k = 3	56.00 ± 3.61	41.13 ± 4.68
SPK	55.80 ± 2.93	38.93 ± 5.12

Comparison between *FGW*, *W* and *GW*. During the validation step, the optimal value of α was consistently selected inside the $]0, 1[$ interval, i.e., excluding 0 and 1, suggesting that both structure and feature information are necessary.

5.3. Graph Barycenter and Compression

In this experiment, we use *FGW* to compute barycenters and approximations of toy graphs. In the first example, we generate graphs following either a circle or 8 symbol with 1D features following a sine and linear variation, respectively. For each example, the number of nodes is drawn randomly between 10 and 25, Gaussian noise is added to the features and a small noise is applied to the structure (some connections are randomly added). An example graph with no noise is provided for each class in the first column of Figure 13. One can see from there that the circle class has a feature varying smoothly (sine) along the graph, but the 8 has a sharp feature change at its center (so that low pass filtering would loose some information). Some examples of the generated graphs are provided in the 2nd-to-7th columns of Figure 13. We compute the *FGW* barycenter containing 10 samples while using the shortest path distance between the nodes as the structural information and the ℓ_2 distance

for the features. We recover an adjacency matrix by thresholding the similarity matrix C given by the barycenter. The threshold is tuned, so as to minimize the Frobenius norm between the original C matrix and the shortest path matrix constructed after thresholding C. Resulting barycenters are showed in Figure 13 for $n = 15$ and $n = 7$ nodes. First, one can see that the barycenters are denoised in both the feature space and structure space. Also note that the sharp change at the center of the 8 class is conserved in the barycenters, which is a nice result when compared to other divergences that tend to smooth-out their barycenters (ℓ_2, for instance). Finally, note that, by selecting the number of nodes in the barycenter, one can compress the graph or estimate a "high resolution" representation from all the samples. To the best of our knowledge, no other method can compute such graph barycenters. Finally, note that *FGW* is interpretable, because the resulting OT matrix provides correspondence between the nodes from the samples and those from the barycenter.

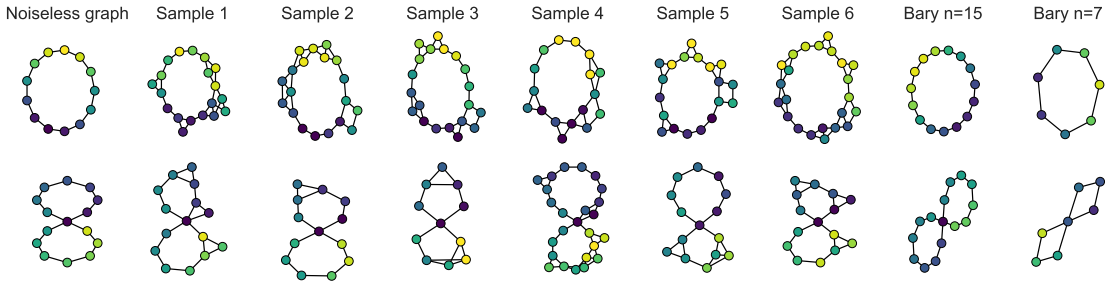

Figure 13. Illustration of *FGW* graph barycenter. The first column illustrates the original settings with the noiseless graphs, and columns 2 to 7 are noisy samples that constitute the datasets. Columns 8 and 9 show the barycenters for each setting, with different number of nodes. Blue nodes indicates a feature value close to -1, yellow nodes close to 1.

In the second experiment, we evaluate the ability of FGW to perform graph approximation and compression on a Stochastic Block Model graph [60,61]. The question is to see whether estimating an approximated graph can recover the relation between the blocks and perform simultaneously a community clustering on the original graph (using the OT matrix). We generate two community graphs illustrated in the left column of Figure 14. We can see that the relation between the blocks is sparse and has a "linear" structure, the example in the first line has features that follow the blocks (noisy but similar in each block) whereas the example in the second line has two modes per block. The first graph approximation (top line) is done with four nodes and we can recover both the blocks in the graph and the average feature on each blocks (colors on the nodes). The second problem is more complex due to the two modes per block but one can see that when approximating the graph with eight nodes we recover both the structure between the blocks and the sub-clusters in each block, which illustrates the strength of *FGW* that encodes both features and structures.

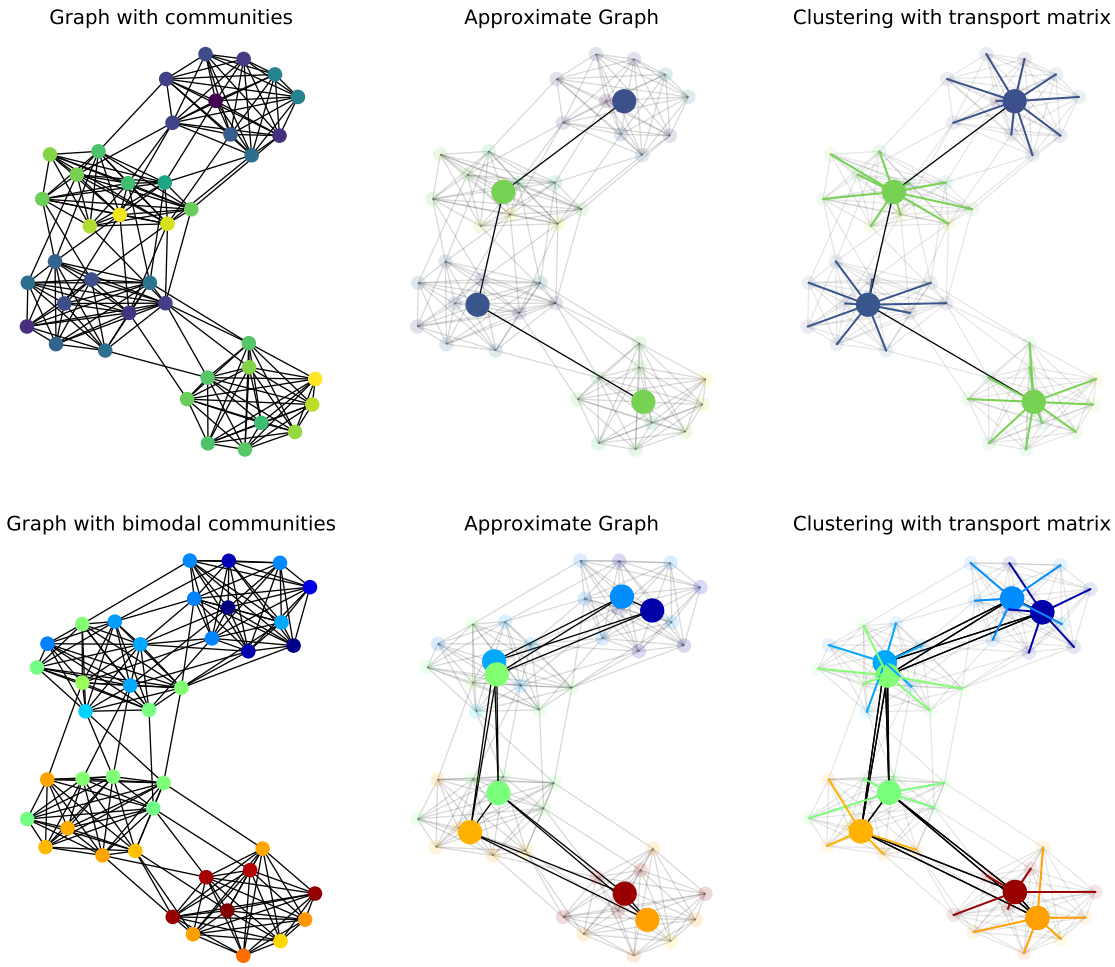

Figure 14. Example of community clustering on graphs using *FGW*. (**top**) Community clustering with four communities and uniform features per cluster. (**bottom**) Community clustering with four communities and bimodal features per cluster (and two nodes per cluster in the approximate graph).

5.4. Unsupervised Learning: Graphs Clustering

In the last experiment, we evaluate the ability of *FGW* to perform a clustering of multiple graphs and to retrieve meaningful barycenters of such clusters. To do so, we generate a dataset of four groups of community graphs. Each graph follows a simple Stochastic Block Model [60,61] and the groups are defined *w.r.t.* the number of communities inside each graph and the distribution of their labels. More precisely, labels are drawn from a 1D gaussian distribution specific to each community: each feature of a node is drawn from a law $\mathcal{N}(n_{c_1} + n_{c_2}, \sigma)$, where n_{c_1} is the index of the community and $n_{c_2} \in \{0, 1\}$ to create two modes per community. The dataset is composed of 40 graphs (10 graphs per group) and the number of nodes of each graph is drawn randomly from $\{20, 30, \ldots, 50\}$, as illustrated in Figure 15. We perform a k-means clustering using the *FGW* barycenter defined in Equation (27) as the centroid of the groups and the *FGW* distance for the cluster assignment. We set the number of nodes of each centroid to 30. We perform a thresholding on the pairwise similarity matrix C of the centroid at the end in order to obtain an adjacency matrix for visualization purposes. The threshold value is chosen, so as to minimize the distance that is induced by the Frobenius norm between the original matrix C and the shortest path matrix obtained from the adjacency matrix. The evolution of the barycenters along the iterations is reported in Figure 15. We can see that these centroids recover community structures and feature distributions that are representative of their cluster content. On this example, note that the clustering perfectly recovers the known groups in the dataset. To the best of our

knowledge, there exists no other method that is able to perform a clustering of graphs of arbitrary size and to retrieve the average graph in each cluster without having to solve a pre-image problem.

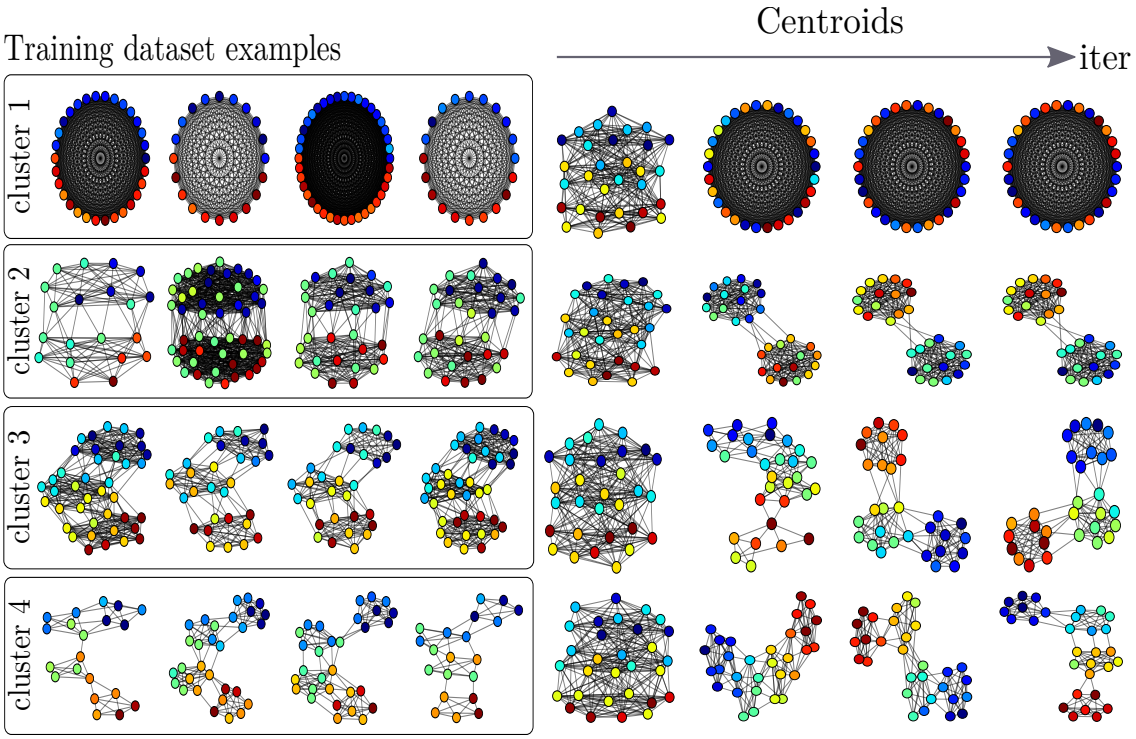

Figure 15. (**left**) Examples from the clustering dataset, color indicates the labels. (**right**) Evolution of the centroids of each cluster in the *k*-means clustering, from the random initialization until convergence to the final centroid.

6. Conclusions

This paper presents the Fused Gromov–Wasserstein (*FGW*) distance. Inspired by both Wasserstein and Gromov–Wasserstein distances, *FGW* can compare structured objects by including the inherent relations that exist between the elements of the objects, constituting their structure information, and their feature information, part of a common ground space between each structured objects. We have stated mathematical results about this new distance, such as metric, interpolation, and geodesic properties. We have also provided a concentration result for the convergence of finite samples. In the discrete case, algorithms to compute *FGW* itself and related FrÃ©chet means are provided. The use of this new distance is illustrated on problems involving structured objects, such as time series embedding, graph classification, graph barycenter computation, and graph clustering. Several questions are raised by this work. From a practical side, the FGW method is quite expensive to compute and further works will try to lower the computational complexity of the underlying optimization problem to ensure better scalability to very large graphs. Moreover, while we mostly consider in this work structure of graphs described by the shortest path, other choices could be made, such as the distances based on the Laplacian of the graph. Finally, from a theoretical point of view, it is often valuable that the geodesic path be unique, so as to defined properly objects, such as gradient flows. One interesting result would be, for example, to see if the geodesic is unique with respect to the strong isomorphism relation.

7. Proofs of the Mathematical Properties

This section presents all of the proofs of previous theorems and results. We note $P_i\#\mu$ the projection on the i-th marginal of μ. We will frequently use the following lemma:

Lemma 2. *Let $q \geq 1$. We have:*

$$\forall x, y \in \mathbb{R}_+,\ (x+y)^q \leq 2^{q-1}(x^q + y^q). \tag{30}$$

Proof. Indeed, if $q > 1$

$$(x+y)^q = \Big((\tfrac{1}{2^{q-1}})^{\frac{1}{q}} \tfrac{x}{(\frac{1}{2^{q-1}})^{\frac{1}{q}}} + (\tfrac{1}{2^{q-1}})^{\frac{1}{q}} \tfrac{y}{(\frac{1}{2^{q-1}})^{\frac{1}{q}}}\Big)^q \leq \big[(\tfrac{1}{2^{q-1}})^{\frac{1}{q-1}} + (\tfrac{1}{2^{q-1}})^{\frac{1}{q-1}}\big]^{q-1}\big(\tfrac{x^q}{\frac{1}{2^{q-1}}} + \tfrac{y^q}{\frac{1}{2^{q-1}}}\big)$$

$$= \tfrac{x^q}{\frac{1}{2^{q-1}}} + \tfrac{y^q}{\frac{1}{2^{q-1}}}.$$

Last inequality is a consequence of HÃlder inequality. The result remains valid for $q = 1$. □

7.1. Proof of Proposition 1—Comparison between FGW, GW and W

Proof. of the Proposition.

For the two inequalities (9) and (10), let π be an optimal coupling for the Fused Gromov-Wasserstein distance between μ and ν (assuming its existence for now). Clearly:

$$d_{FGW,\alpha,p,q}(\mu,\nu) = \Big(\int_{(X\times\Omega\times Y\times\Omega)^2} \big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\big)^p \, d\pi((x,a),(y,b))d\pi((x',a'),(y',b'))\Big)^{\frac{1}{p}}$$

$$\geq \Big(\int_{X\times\Omega\times Y\times\Omega} (1-\alpha)^p d(a,b)^{pq}\, d\pi((x,a),(y,b))\Big)^{\frac{1}{p}} = (1-\alpha)\Big(\int_{\Omega\times\Omega} d(a,b)^{pq} dP_{2,4}\#\pi(a,b)\Big)^{\frac{1}{p}}$$

Becasue $\pi \in \Pi(\mu,\nu)$ the coupling $P_{2,4}\#\pi$ is in $\Pi(\mu_A,\nu_B)$. So by suboptimality:

$$d_{FGW,\alpha,p,q}(\mu,\nu) \geq (1-\alpha)(d_{W,pq}(\mu_A,\nu_B))^q$$

which proves Equation (9). Same reasoning is used for Equation (10).

For the last inequality (11), let $\pi \in \Pi(\mu,\nu)$ be any admissible coupling. By suboptimality:

$$d_{FGW,\alpha,p,1}(\mu,\nu) \leq \Big(\int_{(X\times\Omega\times Y\times\Omega)^2} \big((1-\alpha)d(a,b) + \alpha|d_Z(x,x') - d_Z(y,y')|\big)^p d\pi((x,a),(y,b))d\pi((x',a'),(y',b'))\Big)^{\frac{1}{p}}$$

$$\overset{(*)}{\leq} \Big(\int_{(X\times\Omega\times Y\times\Omega)^2} \big((1-\alpha)d(a,b) + \alpha d_Z(x,y) + \alpha d_Z(x',y')\big)^p d\pi((x,a),(y,b))d\pi((x',a'),(y',b'))\Big)^{\frac{1}{p}}$$

$$\leq \Big(\int_{(X\times\Omega\times Y\times\Omega)^2} \big((1-\alpha)d(a,b) + \alpha d_Z(x,y) + (1-\alpha)d(a',b') + \alpha d_Z(x',y')\big)^p d\pi((x,a),(y,b))d\pi((x',a'),(y',b'))\Big)^{\frac{1}{p}}$$

$$\overset{(**)}{\leq} 2\Big(\int_{X\times\Omega\times Y\times\Omega} \big((1-\alpha)d(a,b) + \alpha d_Z(x,y)\big)^p d\pi((x,a),(y,b))\Big)^{\frac{1}{p}}$$

(*) is the triangle inequality of d_Z and (**) Minkowski inequality. Since this inequality is true for any admissible coupling π we can apply it with the optimal coupling for the Wasserstein distance defined in the proposition and the claim follows. □

7.2. Proof of Theorem 1—Metric Properties of FGW

We prove the theorem point by point: first the existence, then the equality relation and finally the triangle inequality statement. We first recall the definition of weak convergence of probability measure in a metric space [62]:

Definition 10 (Weak-convergence on a metric space)**.** *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on $(X, \mathcal{B}(d))$ where (X,d) is a completely metrizable topological space (Polish space) and $\mathcal{B}(d)$ is the Borel σ*

algebra. We say that $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly to* μ *in* $(X, \mathcal{B}(d))$ *if for all continuous and bounded functions* $f : X \to R$*:*

$$\int_X f d\mu_n \to \int_X f d\mu \tag{31}$$

We also recall the definition of semi continuity in a metric space:

Definition 11 (Lower-semi continuity)**.** *On a metric space* (X, d) *a function* $f : X \to \mathbb{R} \cup \{+\infty\}$ *is said to be lower semi-continuous (l.s.c.) if for every sequence* $x_n \to x$ *we have* $f(x) \leq \liminf f(x_n)$

To state the existence of a minimizer, we will rely on the following lemma:

Lemma 3. *Let* W *be a Polish space. If* $f : W \times W \to \mathbb{R}_+ \cup +\infty$ *is lower semi-continuous, then the functional* $J : \mathcal{P}(W) \to \mathbb{R} \cup +\infty$ *with* $J(\mu) = \int \int f(w, w')d\mu(w)d\mu(w')$ *is l.s.c. for the weak convergence of measures.*

Proof. f is l.s.c. and bounded from below by 0. We can consider $(f_k)_k$ a sequence of continuous and bounded functions converging increasingly to f (see e.g., [63]). By the monotone convergence theorem $J_k(\mu) \to J(\mu) \overset{def}{=} \sup_k J_k(\mu) = \sup_k \int \int f_k d\mu d\mu$.

Moreover every J_k is continuous for the weak convergence. Using Theorem 2.8 [62] on the Polish space $W \times W$ we know that if μ_n converges weakly to μ then the product measure $\mu_n \otimes \mu_n$ converges weakly to $\mu \otimes \mu$. In this way $\lim_{n\to\infty} J_k(\mu_n) = J_k(\mu)$ since f_k are continuous and bounded. In particular every J_k is l.s.c.

We can conclude that J is l.s.c. as the supremum of l.s.c. functionals on the metric space of $(\mathcal{P}(W), \delta)$ (see e.g., [63]). Here we equipped $\mathcal{P}(W)$ with a metric δ as *e.g.* $\delta(\mu, \nu) = \sum_{k=1}^{\infty} 2^{-k} | \int_W f_k d\mu - \int_W f_k d\nu |$ (see Remark 5.11 in [64]). □

Proposition 2. *Existence of an optimal coupling for the FGW distance. For* $p, q \geq 1$, $\pi \to E_{p,q,\alpha}(\pi)$ *always achieves a infimum* π^* *in* $\Pi(\mu, \nu)$ *such that* $d_{FGW,\alpha,p,q}(\mu, \nu) = E_{p,q,\alpha}(\pi^*) < +\infty$.

Proof. Since $X \times \Omega$ and $Y \times \Omega$ are Polish spaces we known that $\Pi(\mu, \nu) \subset \mathcal{P}(X \times \Omega \times Y \times \Omega)$ is compact (Theorem 1.7 in [63]), so by applying Weierstrass theorem we can conclude that the infimum is attained at some $\pi^* \in \Pi(\mu, \nu)$ if $\pi \to E_{p,q,\alpha}(\pi)$ is l.s.c.

We will use Lemma 3 to prove that the functionnal is l.s.c. on $\Pi(\mu, \nu)$. If we consider $W = X \times \Omega \times Y \times \Omega$ which is a a metric space endowed with the distance $d_X \oplus d \oplus d_Y \oplus d$ and $f((w = (x, a, y, b), w' = (x', a', y', b')) = ((1 - \alpha)d(a, b)^q + \alpha L(x, y, x', y')^q)^p$ then f is l.s.c. by continuity of d, d_X and d_Y. With the previous reasoning we can conclude that the infimum is attained. Finally finiteness come from:

$$\begin{aligned} &\int_{(X\times\Omega\times Y\times\Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\right)^p d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b')) \\ &\overset{*}{\leq} \int 2^{p-1}(1-\alpha)d(a,b)^{qp} d\mu_A(a)d\nu(b) + \int 2^{p-1}\alpha L(x,y,x',y')^{qp} d\mu_X(x)d\mu_X(x')d\nu_Y(y')d\nu_Y(y') \\ &\overset{**}{<} +\infty \end{aligned} \tag{32}$$

where in (*) we used Lemma 30 and in (**) that μ, ν are in $\mathbb{S}_{pq}(\Omega)$. □

Proposition 3. *Equality relation. For* $\alpha \in]0, 1[$, $d_{FGW,\alpha,p,q}(\mu, \nu) = 0$ *if an only if there exists a bijective function* $f = (f_1, f_2) : supp(\mu) \to supp(\nu)$ *such that:*

$$f\#\mu = \nu \tag{33}$$

$$\forall (x, a) \in supp(\mu)\ , f_2(x, a) = a \tag{34}$$

$$\forall (x,a),(x',a') \in supp(\mu)^2,\ d_X(x,x') = d_Y(f_1(x,a),f_1(x',a')) \tag{35}$$

Moreover if (μ,ν) *are generalized labeled graphs then* $d_{FGW,\alpha,p,q}(\mu,\nu) = 0$ *if and only if* (X,d_X,μ) *and* (Y,d_Y,ν) *are strongly isomorphic.*

Proof. For the first point, let us assume that there exists a function f verifying (33)–(35). We consider the map $\pi = (I_d \times f)\#\mu \in \Pi(\mu,\nu)$. We note $f = (f_1,f_2)$. Then:

$$\begin{aligned} E_{p,q,\alpha}(\pi) &= \int_{(X\times\Omega\times Y\times\Omega)^2} \Big((1-\alpha)d(a,b)^q + \alpha L((x,y,x',y')^q\Big)^p d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\ &= \int_{(X\times\Omega)^2} \Big((1-\alpha)d(a,f_2(x,a))^q + \alpha L\big((x,f_1(x,a),x',f_1(x',a'))\big)^q\Big)^p d\mu(x,a)d\mu(x',a') \\ &= \int_{(X\times\Omega)^2} \Big((1-\alpha)d(a,f_2(x,a))^q + \alpha |d_X(x,x') - d_Y(f_1(x,a),f_1(x',a'))|^q\Big)^p d\mu(x,a)d\mu(x',a') \\ &= 0 \end{aligned} \tag{36}$$

Conversely, suppose that $d_{FGW,\alpha,p,q}(\mu,\nu) = 0$. To prove the existence of a map $f : \text{supp}(\mu) \to \text{supp}(\nu)$ verifying (33)–(35) we will use the Gromov-Wasserstein properties. We are looking for a vanishing Gromov-Wassersein distance between the spaces $X \times \Omega$ and $Y \times \Omega$ equipped with our two measures μ and ν.

More precisely, we define for $\big((x,a),(y,b),(x',a'),(y',b')\big) \in (X\times\Omega\times Y\times\Omega)^2$ and $\beta \in]0,1[$:

$$d_{X\times\Omega}\big((x,a),(x',a')\big) = (1-\beta)d_X(x,x') + \beta d(a,a')$$

and

$$d_{Y\times\Omega}\big((y,b),(y',b')\big) = (1-\beta)d_Y(y,y') + \beta d(b,b')$$

We will prove that $d_{GW,p}(d_{X\times\Omega},d_{Y\times\Omega},\mu,\nu) = 0$. To show that we will bound the Gromov cost with the metrics $d_{X\times\Omega},d_{Y\times\Omega}$ by the Gromov cost with the metrics d_X,d_Y and a Wasserstein cost.

Let $\pi \in \Pi(\mu,\nu)$ be any admissible transportation plan. Subsequently, for $n \geq 1$:

$$\begin{aligned} &J_n(d_{X\times\Omega},d_{Y\times\Omega},\pi) \overset{def}{=} \int_{(X\times\Omega\times Y\times\Omega)^2} L(x,y,x',y')^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\ &= \int_{(X\times\Omega\times Y\times\Omega)^2} |(1-\beta)(d_X(x,x') - d_Y(y,y')) + \beta(d(a,a') - d(b,b'))|^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\ &\leq \int_{(X\times\Omega\times Y\times\Omega)^2} (1-\beta)|d_X(x,x') - d_Y(y,y')|^n d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b')) \\ &+ \int_{(X\times\Omega\times Y\times\Omega)^2} \beta|d(a,a') - d(b,b')|^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \end{aligned}$$

using Jensen inequality with convexity of $t \to t^n$ and subadditivity of $|.|$. We note $(*)$ the first term above and $(**)$ the second term above. By the triangle inequality property of d we have:

$(**) \leq \beta \int_{(X\times\Omega\times Y\times\Omega)^2} \big(d(a,b) + d(a',b')\big)^n d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b')) \overset{def}{=} \beta M_n(\pi)$ such that we have shown:

$$\forall \pi \in \Pi(\mu,\nu), \forall n \geq 1, J_n(d_{X\times\Omega},d_{Y\times\Omega},\pi) \leq (1-\beta)J_n(d_X,d_Y,\pi) + \beta M_n(\pi) \tag{37}$$

Now let π_* be an optimal coupling for $d_{FGW,\alpha,p,q}$ between μ and ν. By hypothesis $d_{FGW,\alpha,p,q}(\mu,\nu)=0$ so that:

$$J_{qp}(d_X, d_Y, \pi*) = 0 \tag{38}$$

and:

$$H_{qp}(\pi^*) = 0. \tag{39}$$

Subsequently, $\int_{(X\times\Omega\times Y\times\Omega)} d(a,b)^{qp} d\pi^*((x,a),(y,b))=0$ which implies that d is zero π^* *a.e.* so that $\int_{(X\times\Omega\times Y\times\Omega)} d(a,b)^m d\pi^*((x,a),(y,b)) = 0$ for any $m \in \mathbb{N}^*$. In this way:

$$M_{qp}(\pi^*) = \beta \int_{(X\times\Omega\times Y\times\Omega)^2} \sum_h \binom{qp}{h} d(a,b)^h d(a',b')^{qp-h} d\pi^*((x,a),(y,b)) d\pi^*((x',a'),(y',b'))$$
$$= \beta \sum_h \binom{qp}{h} \Big(\int_{(X\times\Omega\times Y\times\Omega)} d(a,b)^h d\pi^*((x,a),(y,b)) \Big) \Big(\int_{(X\times\Omega\times Y\times\Omega)} d(a',b')^{qp-h} d\pi^*((x',a'),(y',b')) \Big) = 0$$

Using Equation (37), we have shown

$$J_{qp}(d_{X\times\Omega}, d_{Y\times\Omega}, \pi_*) = 0$$

which implies that $d_{GW,p}(d_{X\times\Omega}, d_{Y\times\Omega}, \mu, \nu) = 0$ for the coupling π^*.

Thanks to the Gromov–Wasserstein properties (see [17]) this states the existence of an isometry between $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\nu)$. So there exists a surjective function $f = (f_1, f_2) : \mathrm{supp}(\mu) \to \mathrm{supp}(\nu)$, which verifies **P.1** and:

$$\forall((x,a),(x',a')) \in (\mathrm{supp}(\mu))^2,\ d_{X\times\Omega}((x,a),(x',a')) = d_{Y\times\Omega}(f(x,a), f(x',a')) \tag{40}$$

or equivalently:

$$\forall((x,a),(x',a')) \in (\mathrm{supp}(\mu))^2,\ (1-\beta) d_X(x,x') + \beta d(a,a') = (1-\beta) d_Y(f_1(x,a), f_1(x',a')) + \beta d(f_2(x,a), f_2(x',a')) \tag{41}$$

In particular, π^* is concentrated on $\{(x,y) = f(x,y)\}$ or equivalently $\pi^* = (I_d \times f)\#\mu$. Injecting π^* in (39) leads to:

$$H_{qp}(\pi^*) = \int_{(X\times\Omega\times Y\times\Omega)} d(a,b)^{qp} d\pi^*((x,a),(y,b)) = \int_{X\times\Omega} d(a, f_2(x,a))^{qp} d\mu(x,a) = 0 \tag{42}$$

Which implies:

$$\forall(x,a) \in \mathrm{supp}(\mu)\ , f_2(x,a) = a \tag{43}$$

Moreover, using the equality (41), we can conclude that:

$$\forall(x,a)(x',a') \in \mathrm{supp}(\mu)^2,\ d_X(x,x') = d_Y(f_1(x,a), f_1(x',a')) \tag{44}$$

In this way, f verifies all of the properties (33)–(35).

Moreover, suppose that μ and ν are generalized labeled graphs. In this case there exists $\ell_f : X \to A$ surjective such that $\mu = (I \times \ell_f)\#\mu_X$. Afterwards, (44) implies that:

$$\forall(x,x') \in \mathrm{supp}(\mu_X)^2,\ d_X(x,x') = d_Y(f_1(x, \ell_f(x)), f_1(x', \ell_f(x'))) \tag{45}$$

We define $I : \mathrm{supp}(\mu_X) \to \mathrm{supp}(\mu_Y)$ such that $I(x) = f_1(x, \ell_f(x))$. Then we have by (45) $d_X(x,x') = d_Y(I(x), I(x'))$ for $(x,x') \in \mathrm{supp}(\mu_X)^2$. Overall we have $f(x,a) = (I(x), a)$ for all $(x,a) \in \mathrm{supp}(\mu)$. Also since $f\#\mu = \nu$ we have $I\#\mu_X = \nu_Y$.

Moreover, I is a surjective function. Indeed let $y \in \text{supp}(\nu_Y)$. Let $b \in \text{supp}(\nu_B)$ such that $(y,b) \in \text{supp}(\nu)$. By surjectivity of f there exists $(x,a) \in \text{supp}(\mu)$ such that $(y,b) = f(x,a) = (I(x),a)$ so that $y = I(x)$.

Overall, f satisfies all **P.1**, **P.2** and **P.3** if μ and ν are generalized labeled graphs. The converse is also true using the reasoning in (36). □

Proposition 4. *Symmetry and triangle inequality.*

$d_{FGW,\alpha,p,q}$ is symmetric and for $q = 1$ satisfies the triangle inequality. For $q \geq 2$ the triangle inequality is relaxed by a factor 2^{q-1}

To prove this result we will use the following lemma:

Lemma 4. *Let $(X \times \Omega, d_X, \mu), (Y \times \Omega, d_Y, \beta), (Z \times \Omega, d_Z, \nu) \in \mathbb{S}(\Omega)^3$. For $(x,a),(x',a') \in (X \times \Omega)^2$, $(y,b),(y',b') \in (Y \times \Omega)^2$ and $(z,c),(z',c') \in (Z \times \Omega)^2$ we have:*

$$L(x,z,x',z')^q \leq 2^{q-1}(L(x,y,x',y')^q + L(y,z,y',z')^q) \tag{46}$$

$$d(a,c)^q \leq 2^{q-1}(d(a,b)^q + d(b,c)^q) \tag{47}$$

Proof. Direct consequence of (30) and triangle inequalities of d, d_X, d_Y, d_Z. □

Proof of Proposition 4. To prove the triangle inequality of $d_{FGW,\alpha,p,q}$ distance for arbitrary measures we will use the Gluing lemma which stresses the existence of couplings with a prescribed structure. Let $(X \times \Omega, d_X, \mu), (Y \times \Omega, d_Y, \beta), (Z \times \Omega, d_Z, \nu) \in \mathbb{S}(\Omega)^3$.

Let $\pi_1 \in \Pi(\mu,\beta)$ and $\pi_2 \in \Pi(\beta,\nu)$ be optimal transportation plans for the Fused Gromov-Wasserstein distance between μ, β and β, ν respectively. By the Gluing Lemma (see [15] and Lemma 5.3.2 in [65]) there exists a probability measure $\pi \in P\big((X \times \Omega) \times (Y \times \Omega) \times (Z \times \Omega)\big)$ with marginals π_1 on $(X \times \Omega) \times (Y \times \Omega)$ and π_2 on $(Y \times \Omega) \times (Z \times \Omega)$. Let π_3 be the marginal of π on $(X \times \Omega) \times (Z \times \Omega)$. By construction $\pi_3 \in \Pi(\mu,\nu)$. So by suboptimality of π_3:

$$\begin{aligned}
&d_{FGW,\alpha,p,q}(d_X,d_Z,\mu,\nu) \leq \Big(\int_{(X\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(a,c)^q + \alpha L(x,z,x',z')^q)^p d\pi_3((x,a),(z,c))d\pi_3((x',a'),(z',c'))\Big)^{\frac{1}{p}} \\
&= \Big(\int_{(X\times\Omega\times Y\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(a,c)^q + \alpha L(x,z,x',z')^q)^p d\pi((x,a),(y,b),(z,c))d\pi((x',a'),(y',b'),(z',c'))\Big)^{\frac{1}{p}} \\
&\overset{(*)}{\leq} 2^{q-1}\Big(\int_{(X\times\Omega\times Y\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(a,b)^q + (1-\alpha)d(b,c)^q + \alpha L(x,y,x',y')^q + \alpha L(y,z,y',z')^q)^p \\
&d\pi((x,a),(y,b),(z,c))d\pi((x',a'),(y',b'),(z',c'))\Big)^{\frac{1}{p}} \\
&\overset{(**)}{\leq} 2^{q-1}\Big(\Big(\int_{(X\times\Omega\times Y\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q)^p d\pi((x,a),(y,b),(z,c))d\pi((x',a'),(y',b'),(z',c'))\Big)^{\frac{1}{p}} \\
&+ \Big(\int_{(X\times\Omega\times Y\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(b,c)^q + \alpha L(y,z,y',z')^q)^p d\pi((x,a),(y,b),(z,c))d\pi((x',a'),(y',b'),(z',c'))\Big)^{\frac{1}{p}}\Big) \\
&= 2^{q-1}\Big(\Big(\int_{(X\times\Omega\times Y\times\Omega)^2} ((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q)^p d\pi_1((x,a),(y,b))d\pi_1((x',a'),(y',b'))\Big)^{\frac{1}{p}} \\
&+ \Big(\int_{(Y\times\Omega\times Z\times\Omega)^2} ((1-\alpha)d(b,c)^q + \alpha L(y,z,y',z')^q)^p d\pi_2((y,b),(z,c))\, d\pi_2((y',b'),(z',c'))\Big)^{\frac{1}{p}}\Big) \\
&= 2^{q-1}(d_{FGW,\alpha,p,q}(\mu,\beta) + d_{FGW,\alpha,p,q}(\beta,\nu))
\end{aligned}$$

with (*) comes from (46) and (47) and (**) is Minkowski inequality. So when $q = 1$, $d_{FGW,\alpha,p,q}$ satisfies the triangle inequality and when $q > 1$, $d_{FGW,\alpha,p,q}$ satisfies a relaxed triangle inequality so that it defines a semi-metric as described previously. □

7.3. Proof of Theorem 2—Convergence and Concentration Inequality.

Proof. The proof of the convergence in FGW directly stems from the weak convergence of the empirical measure and Lemma 3. Moreover, since μ_n and μ are both in the same ground space, we have:

$$d_{FGW,\alpha,p,1}(\mu_n,\mu) \leq 2d_{W,p}(\mu_n,\mu) \implies \mathbb{E}[d_{FGW,\alpha,p,1}(\mu_n,\mu)] \leq 2\mathbb{E}[d_{W,p}(\mu_n,\mu)].$$

We can directly apply Theorem 1 in [31] to state the inequality. □

7.4. Proof of Theorem 3—Interpolation Properties between GW and W

Proof. Let $\pi_{OT} \in \Pi(\mu_A,\nu_B)$ be an optimal coupling for the pq-Wasserstein distance between μ_A and ν_B. We can use the same Gluing lemma (Lemma 5.3.2 in [65]) to construct:

$$\rho \in P(\overbrace{X \times \underbrace{\Omega}_{\pi_{OT}}}^{\mu} \times \overbrace{\Omega \times Y}^{\nu})$$

such that $\rho \in \Pi(\mu,\nu)$ and $P_{2,3}\#\rho = \pi_{OT}$.

Moreover we have:

$$\int_{\Omega\times\Omega} d(a,b)^{pq} d\pi_{OT}(a,b) = \int_{X\times\Omega\times\Omega\times Y} d(a,b)^{pq} d\rho(x,a,b,y) \tag{48}$$

Let $\alpha \geq 0$ and π_α optimal plan for the Fused Gromov-Wasserstein distance between μ and ν. We can deduce that:

$$\begin{aligned}
&d_{FGW,\alpha,p,q}(\mu,\nu)^p - (1-\alpha)^p d_{W,pq}(\mu_A,\nu_B)^{pq} \\
&= \int_{(X\times\Omega\times Y\times\Omega)^2} \Big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\Big)^p d\pi_\alpha((x,a),(y,b)) d\pi_\alpha((x',a'),(y',b')) - \int_{\Omega\times\Omega} (1-\alpha)^p d(a,b)^{pq} d\pi_{OT}(a,b) \\
&\overset{(*)}{\leq} \int_{(X\times\Omega\times Y\times\Omega)^2} \Big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\Big)^p d\rho(x,a,b,y) d\rho(x',a',b',y') - \int_{X\times\Omega\times Y\times\Omega} (1-\alpha)^p d(a,b)^{pq} d\rho(x,a,b,y) \\
&= (1-\alpha)^p \int_{(X\times\Omega\times Y\times\Omega)^2} d(a,b)^{pq} d\rho(x,a,b,y) d\rho(x',a',b',y') - (1-\alpha)^p \int_{X\times\Omega\times Y\times\Omega} d(a,b)^{pq} d\rho(x,a,b,y) \\
&+ \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} \int_{(X\times\Omega\times Y\times\Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y') \\
&= \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} \int_{(X\times\Omega\times Y\times\Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y').
\end{aligned}$$

We note $H_k = \int_{(X\times\Omega\times Y\times\Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y')$.

Using (9) we have shown that:

$$(1-\alpha)(d_{W,pq}(\mu_A,\nu_B))^q \leq d_{FGW,\alpha,p,q}(\mu,\nu) \leq \left((1-\alpha)^p (d_{W,pq}(\mu_A,\nu_B))^{pq} + \textstyle\sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} H_k\right)^{\frac{1}{p}}$$

Accordingly, $\lim_{\alpha\to 0} d_{FGW,\alpha,p,q}(\mu,\nu) = (d_{W,pq}(\mu_A,\nu_B))^q$.

For the case $\alpha \to 1$ we rather consider $\pi_{GW} \in \Pi(\mu_X,\nu_Y)$ an optimal coupling for the pq-Gromov–Wasserstein distance between μ_X and ν_Y and we construct

$$\gamma \in P(\overbrace{\Omega \times \underbrace{X}_{\pi_{GW}}}^{\mu} \times \overbrace{Y \times \Omega}^{\nu})$$

such that $\gamma \in \Pi(\mu,\nu)$ and $P_{2,3}\#\rho = \pi_{GW}$. In the same way as previous reasoning we can derive:

$$\alpha(d_{GW,pq}(\mu_X,\nu_Y))^q \leq d_{FGW,\alpha,p,q}(\mu,\nu) \leq \left(\alpha^p(d_{GW,pq}(\mu_X,\nu_Y))^{pq} + \sum_{k=0}^{p-1}\binom{p}{k}(1-\alpha)^{p-k}\alpha^k J_k\right)^{\frac{1}{p}} \quad (49)$$

with $J_k = \int d(a,b)^{q(p-k)}L(x,y,x',y')^{qk}d\rho(x,a,b,y)d\rho(x',a',b',y')$. In this way $\lim_{\alpha\to 1} d_{FGW,\alpha,p,q}(\mu,\nu) = (d_{GW,pq}(\mu_X,\nu_Y))^q$. □

7.5. Proof of Theorem 4—Constant Speed Geodesic

Proof. Let $t,s \in [0,1]$. Recalling:

$$\forall (x,a),(y,b) \in X\times\Omega\times Y\times\Omega,\ \eta_t(x,a,y,b) = (x,y,(1-t)a+tb) \quad (50)$$

We note $\mathbb{S}_t = (X\times Y\times\Omega,(1-t)d_X \oplus td_Y,\mu_t = \eta_t\#\pi^*)_{t\in[0,1]}$ and $d_t = (1-t)d_X \oplus td_Y$. Let $\|.\|$ be any ℓ_m norm for $m \geq 1$. It suffices to prove:

$$d_{FGW,\alpha,p,1}(\mu_t,\mu_s) \leq |t-s| d_{FGW,\alpha,p,1}(\mu_0,\mu_1) \quad (51)$$

To do so, we consider $\Delta_s^t \in P(X\times Y\times\Omega\times X\times Y\times\Omega)$ defined by $\Delta_s^t = (\eta_t\times\eta_s)\#\pi^* \in \Pi(\mu_t,\mu_s)$ and the following "diagonal" coupling:

$$d\gamma_s^t((x,y),a,(x'',y''),b) = d\Delta_s^t((x,y),a,(x'',y''),b)d\delta_{(x_0,x_1)}(x_0'',x_1'') \quad (52)$$

Subsquently, $\gamma_s^t \in P(X\times Y\times\Omega\times X\times Y\times\Omega)$ and since $\Delta_s^t \in \Pi(\mu_t,\mu_s)$ then $\gamma_s^t \in \Pi(\mu_t,\mu_s)$ So by suboptimality:

$$\begin{aligned}
&d_{FGW,\alpha,p,1}(\mu_t,\mu_s)^p \leq \int_{(X\times Y\times\Omega\times X\times Y\times\Omega)^2} \Big((1-\alpha)d(a,b) + \alpha|d_t[(x,y),(x',y')] - d_s[(x'',y''),(x''',y''')]|\Big)^p \\
&d\gamma_s^t(x,y,a,x'',y'',b)d\gamma_s^t(x',y',a',x''',y''',b') \\
&= \int_{(X\times Y\times\Omega\times X\times Y\times\Omega)^2} \Big((1-\alpha)d(a,b) + \alpha|d_t[(x,y),(x',y')] - d_s[(x,y),(x',y')]|\Big)^p \\
&d\Delta_s^t(x,y,a,x,y,b)d\Delta_s^t(x',y',a',x',y',b') \\
&= \int_{(X\times\Omega\times Y\times\Omega)^2} \left((1-\alpha)\|(1-t)a+tb-(1-s)a-sb\| + \alpha|(1-t)d_X(x,x')+td_Y(y,y') - (1-s)d_X(x,x')+sd_Y(y,y')|\right)^p \\
&d\pi^*(x,a,y,b)d\pi^*(x',a',y',b') \\
&= |t-s|^p \int_{(X\times\Omega\times Y\times\Omega)^2} \Big((1-\alpha)\|a-b\| + \alpha|d_X(x,x') - d_Y(y,y')|\Big)^p d\pi^*(x,a,y,b)d\pi^*(x',a',y',b')
\end{aligned}$$

So $d_{FGW,\alpha,p,1}(\mu_t,\mu_s) \leq |t-s| d_{FGW,\alpha,p,1}(d_0,d_1,\mu_0,\mu_1)$. □

Author Contributions: Conceptualization, T.V., L.C., N.C., R.T. and R.F.; methodology, T.V., L.C., N.C., R.T. and R.F.; software, T.V., R.F., L.C. and N.C.; formal analysis, T.V., L.C., N.C., R.T. and R.F.; investigation, T.V., L.C., N.C., R.T. and R.F.; resources, T.V., L.C., N.C., R.T. and R.F.; data curation, T.V., R.F., L.C. and N.C.; writing—original draft preparation, T.V.; writing—review and editing, T.V., L.C., N.C., R.T. and R.F.; visualization, T.V., L.C. and R.F.; supervision, L.C., N.C., R.T. and R.F.; project administration, L.C., N.C., R.T. and R.F.; funding acquisition, L.C., N.C., R.T. and R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially funded through the projects OATMIL ANR-17-CE23-0012, 3IA Cote dÁzur Investments ANR-19-P3IA-0002 and MATS ANR-18-CE23-0006 of the French National Research Agency (ANR).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv* **2018**, arXiv:cs.LG/1806.01261.
2. Pearl, J. Fusion, Propagation, and Structuring in Belief Networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]
3. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2009.
4. Džeroski, S.; De Raedt, L.; Driessens, K. Relational Reinforcement Learning. *Mach. Learn.* **2001**, *43*, 7–52. [CrossRef]
5. Hjort, N.; Holmes, C.; Mueller, P.; Walker, S. *Bayesian Nonparametrics: Principles and Practice*; Cambridge University Press: Cambridge, UK, 2010.
6. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
7. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
8. Shervashidze, N.; Schweitzer, P.; van Leeuwen, E.J.; Mehlhorn, K.; Borgwardt, K.M. Weisfeiler-Lehman Graph Kernels. *J. Mach. Learn. Res.* **2011**, *12*, 2539–2561.
9. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In Proceedings of the International Conference on Machine Learning Research, New York, NY, USA, 20–22 June 2016; Balcan, M.F., Weinberger, K.Q., Eds.; PMLR: New York, NY, USA, 2016; Volume 48, pp. 2014–2023.
10. Bakir, G.H.; Hofmann, T.; Schölkopf, B.; Smola, A.J.; Taskar, B.; Vishwanathan, S.V.N. *Predicting Structured Data (Neural Information Processing)*; The MIT Press: Cambridge, MA, USA, 2007.
11. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 43–49. [CrossRef]
12. Cuturi, M.; Blondel, M. Soft-DTW: A Differentiable Loss Function for Time-Series. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017; PMLR: Stockholm, Sweden, 2017; Volume 70, pp. 894–903.
13. Nowozin, S.; Gehler, P.V.; Jancsary, J.; Lampert, C.H. *Advanced Structured Prediction*; The MIT Press: Cambridge, MA, USA, 2014.
14. Niculae, V.; Martins, A.; Blondel, M.; Cardie, C. SparseMAP: Differentiable Sparse Structured Inference. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; PMLR: Stockholm, Sweden, 2018; Volume 80, pp. 3796–3805.
15. Villani, C. *Optimal Transport: Old and New*, 2009th ed.; Grundlehren der mathematischen Wissenschaften; Springer: Berlin/Heidelberg, Germany, 2008.
16. Sturm, K.T. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv* **2012**, arXiv:1208.0434.
17. Memoli, F. Gromov Wasserstein Distances and the Metric Approach to Object Matching. *Found. Comput. Math.* **2011**, 1–71. doi:10.1007/s10208-011-9093-5. [CrossRef]
18. Vayer, T.; Courty, N.; Tavenard, R.; Chapel, L.; Flamary, R. Optimal Transport for structured data with application on graphs. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Long Beach, CA, USA, 2019; Volume 97, pp. 6275–6284.
19. Verma, S.; Zhang, Z.L. Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.; Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 88–98.
20. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
21. Peyré, G.; Cuturi, M. Computational Optimal Transport. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
22. Solomon, J.; Peyré, G.; Kim, V.G.; Sra, S. Entropic Metric Alignment for Correspondence Problems. *ACM Trans. Graph.* **2016**, *35*, 72:1–72:13. [CrossRef]
23. Ezuz, D.; Solomon, J.; Kim, V.G.; Ben-Chen, M. GWCNN: A Metric Alignment Layer for Deep Shape Analysis. *Comput. Graph. Forum* **2017**, *36*, 49–57. [CrossRef]

24. Bunne, C.; Alvarez-Melis, D.; Krause, A.; Jegelka, S. Learning Generative Models across Incomparable Spaces. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 851–861.
25. Demetci, P.; Santorella, R.; Sandstede, B.; Noble, W.S.; Singh, R. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv* **2020**. [CrossRef]
26. Peyré, G.; Cuturi, M.; Solomon, J. Gromov-Wasserstein averaging of kernel and distance matrices. In Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York, NY, USA, 19–24 June 2016; pp. 2664–2672.
27. Haasdonk, B.; Bahlmann, C. Learning with Distance Substitution Kernels. In *Pattern Recognition*; Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A., Eds.; Springer: Berlin, Heidelberg, 2004; pp. 220–227.
28. Borg, I.; Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2005.
29. Bachem, O.; Lucic, M.; Krause, A. Practical Coreset Constructions for Machine Learning. *arXiv* **2017**, arXiv:stat.ML/1703.06476.
30. Thorpe, M.; Park, S.; Kolouri, S.; Rohde, G.K.; Slepčev, D. A Transportation L^p Distance for Signal Analysis. *J. Math. Imaging Vis.* **2017**, *59*, 187–210. [CrossRef]
31. Jonathan Weed, F.B. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv* **2017**, arXiv:1707.00087.
32. Benamou, J.D.; Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **2000**, *84*, 375–393. [CrossRef]
33. Bonneel, N.; van de Panne, M.; Paris, S.; Heidrich, W. Displacement Interpolation Using Lagrangian Mass Transport. In Proceedings of the 2011 SIGGRAPH Asia Conference, Hong Kong, China, 11–15 December 2011; ACM: New York, NY, USA, 2011; pp. 158:1–158:12. [CrossRef]
34. Chizat, L.; Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018.
35. Zhang, R.; Chen, C.; Li, C.; Duke, L.C. Policy Optimization as Wasserstein Gradient Flows. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; PMLR: Stockholm, Sweden, 2018; Volume 80, pp. 5741–5750.
36. Ferradans, S.; Papadakis, N.; Peyré, G.; Aujol, J.F. Regularized discrete optimal transport. *SIAM J. Imaging Sci.* **2014**, *7*, 1853–1882. [CrossRef]
37. Flamary, R.; Courty, N.; Tuia, D.; Rakotomamonjy, A. Optimal transport with Laplacian regularization: Applications to domain adaptation and shape matching. In *NIPS Workshop on Optimal Transport and Machine Learning*; OTML: Montreal, QC, Canada, 2014.
38. Lacoste-Julien, S. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv* **2016**, arXiv:1607.00345.
39. Maron, H.; Lipman, Y. (Probably) Concave Graph Matching. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 408–418.
40. Redko, I.; Vayer, T.; Flamary, R.; Courty, N. CO-Optimal Transport. *arXiv* **2020**, arXiv:stat.ML/2002.03731.
41. Agueh, M.; Carlier, G. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **2011**, *43*, 904–924. [CrossRef]
42. Cuturi, M.; Doucet, A. Fast Computation of Wasserstein Barycenters. In Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22–24 June 2014; Xing, E.P., Jebara, T., Eds.; PMLR: Bejing, China, 2014; Volume 32, pp. 685–693.
43. Kruskal, J.B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **1964**, *29*, 115–129. [CrossRef]
44. Sutherland, J.J.; O'brien, L.A.; Weaver, D.F. Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906–1915. [CrossRef] [PubMed]
45. Borgwardt, K.M.; Kriegel, H.P. Shortest-Path Kernels on Graphs. In Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, TX, USA, 27–30 November 2005; IEEE Computer Society: Washington, DC, USA, 2005; pp. 74–81, [CrossRef]

46. Kriege, N.; Fey, M.; Fisseler, D.; Mutzel, P.; Weichert, F. Recognizing Cuneiform Signs Using Graph Based Methods. In Proceedings of the International Workshop on Cost-Sensitive Learning (COST), San Diego, California, USA, 3–5 May 2018.
47. Feragen, A.; Kasenburg, N.; Petersen, J.; de Bruijne, M.; Borgwardt, K. Scalable kernels for graphs with continuous attributes. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 216–224.
48. Debnath, A.K.; Lopez de Compadre, R.L.; Debnath, G.; Shusterman, A.J.; Hansch, C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.* **1991**, *34*, 786–797. [CrossRef] [PubMed]
49. Kriege, N.M.; Giscard, P.; Wilson, R.C. On Valid Optimal Assignment Kernels and Applications to Graph Classification. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
50. Wale, N.; Watson, I.A.; Karypis, G. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowl. Inf. Syst.* **2008**, *14*, 347–375. [CrossRef]
51. Yanardag, P.; Vishwanathan, S. Deep Graph Kernels. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; ACM: New York, NY, USA, 2015; pp. 1365–1374.
52. Kersting, K.; Kriege, N.M.; Morris, C.; Mutzel, P.; Neumann, M. Benchmark Data Sets for Graph Kernels. 2016. Available online: https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets (accessed on 26 August 2020).
53. Vishwanathan, S.V.N.; Schraudolph, N.N.; Kondor, R.; Borgwardt, K.M. Graph Kernels. *J. Mach. Learn. Res.* **2010**, *11*, 1201–1242.
54. Luss, R.; d'Aspremont, A. Support Vector Machine Classification with Indefinite Kernels. In Proceedings of the 20th International Conference on Neural Information Processing Systems, Kitakyushu, Japan, 3 December 2007; pp. 953–960.
55. Gärtner, T.; Flach, P.; Wrobel, S. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 129–143.
56. Shervashidze, N.; Vishwanathan, S.V.N.; Petri, T.H.; Mehlhorn, K.; Borgwardt, K. Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*; Hilton Clearwater Beach Resort: Clearwater Beach, FL, USA 2009.
57. Neumann, M.; Garnett, R.; Bauckhage, C.; Kersting, K. Propagation kernels: efficient graph kernels from propagated information. *Mach. Learn.* **2016**, *102*, 209–245. [CrossRef]
58. Siglidis, G.; Nikolentzos, G.; Limnios, S.; Giatsidis, C.; Skianis, K.; Vazirgianis, M. GraKeL: A Graph Kernel Library in Python. *arXiv* **2018**, arXiv:1806.02193.
59. Shchur, O.; Mumme, M.; Bojchevski, A.; Günnemann, S. Pitfalls of Graph Neural Network Evaluation. *arXiv* **2018**, arXiv:1811.05868.2018.
60. Wang, Y.J.; Wong, G.Y. Stochastic blockmodels for directed graphs. *J. Am. Stat. Assoc.* **1987**, *82*, 8–19. [CrossRef]
61. Nowicki, K.; Snijders, T.A.B. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* **2001**, *96*, 1077–1087. [CrossRef]
62. Billingsley, P. *Convergence of Probability Measures*, 2nd ed.; Wiley Series in Probability and Statistics: Probability and Statistics; A Wiley-Interscience Publication; John Wiley & Sons Inc.: New York, NY, USA, 1999.
63. Santambrogio, F. *Optimal Transport for Applied Mathematicians*; Birkäuser: Basel, Switzerland, 2015.
64. Ambrosio, L.; Gigli, N.; Savare, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005. [CrossRef]
65. Ambrosio, L.; Gigli, N.; Savare, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*; Lectures in Mathematics; ETH Zürich, Birkhäuser: Basel, Switzerland, 2005.