Graph Diffusion Wasserstein Distances

Amélie Barbe, Marc Sebban, Paulo Gonçalves, Pierre Borgnat, Rémi Gribonval

Graph Diffusion Wasserstein Distances

A. Barbe, M. Sebban, P. Gonçalves, P. Borgnat and R. Gribonval

Abstract. Optimal Transport (OT) for structured data has received much attention in the machine learning community, especially for addressing graph classification or graph transfer learning tasks. In this paper, we present the Diffusion Wasserstein (`DW`) distance, as a generalization of the standard Wasserstein distance to undirected and connected graphs where nodes are described by feature vectors. `DW` is based on the Laplacian exponential kernel and benefits from the heat diffusion to catch both structural and feature information from the graphs. We further derive lower/upper bounds on `DW` and show that it can be directly plugged into the Fused Gromov Wasserstein (`FGW`) distance that has been recently proposed, leading - for free - to a DifFused Gromov Wasserstein distance (`DFGW`) that allows a significant performance boost when solving graph domain adaptation tasks.

Keywords: Optimal Transport · Graph Laplacian · Heat Diffusion.

1 Introduction

Many real-world problems in natural and social sciences take the form of structured data such as graphs which require efficient metrics for comparison. In this context, graphs kernels that take into account both structural and feature information have achieved a tremendous success during the past years to address graph classification tasks (see the most recent survey [12]), *e.g.* using Support Vector Machines.

Unlike standard graph kernel-based methods, we consider in this paper graphs as discrete distributions with the main objective of defining a distance on this space of probability measures, which is the main goal of Optimal Transport (OT) (see [15] for the original problem and [11] for a regularized version). OT has received much attention from the Machine Learning community from both theoretical and practical perspectives. It provides a natural geometry for probability measures and aims at moving a source distribution on the top of a target measure in an optimal way with respect to a cost matrix. If the latter is related to an actual distance on some geometric space, the solution of the OT problem defines a distance (the so-called p-Wasserstein distance `W`) on the corresponding space of probability measures.

Several recent works in OT have been devoted to the comparison of structured data such as undirected graphs. Following [16], the authors of [19] introduced the Gromov-Wasserstein (`GW`) distance allowing to compute a distance between two metric measures. `GW` can be used to catch and encode some structure of graphs, like the shortest path between two vertices. However, this distance is not able to

jointly take into account both features (or attributes) at the node level and more global structural information. To address this issue, the Fused-Gromov Wasserstein (FGW) distance was introduced in [24] as an interpolation between GW (over the structure) and the Wasserstein distance (over the features). In order to better take into account the global structure of the graphs for graph comparison and alignment, and for the transportation of signals between them, the authors of [13,14] introduced GOT, as a new Wasserstein distance between graph signal distributions that resorts to the graph Laplacian matrices. This approach, initially constrained by the fact that both graphs have the same number of vertices [13] was recently extended to graphs of different sizes in [14]. However, the graph alignment and the proposed distance are still based only on the structure, and do not use the features.

In this paper, we address the limitations of the aforementioned graph OT-based methods by leveraging the notions of heat kernel and heat diffusion that are widely used to capture topological information in graphs [3,23] or graph signals [21]. Inspired from the Graph Diffusion Distance (GDD) introduced in [9], and rooted in the Graph Signal Processing (GSP) approaches [17] of graphs with features (or "graph signals" in GSP), better known under the name of *attributed graphs* in machine learning, we present the Diffusion Wasserstein (DW) distance, as a generalization of the standard Wasserstein distance to attributed graphs. While GDD is limited to graphs of the same size, does not take into account features, and would not be directly usable in the OT setting, we leverage its definition to capture in an OT problem the graph structure combined with the smoothing of features along this structure. Leveraging the properties of the heat diffusion, we establish the asymptotic behavior of our new distance. We also provide a sufficient condition for the expected value of DW to be upper bounded by the Wasserstein distance. We further show that computing DW boils down to reweighting the original features by taking into account the heat diffusion in the graphs. For this reason, DW can be plugged into FGW in place of the Wasserstein distance to get for free a family of so-called DifFused Gromov Wasserstein distances (DFGW). We will show in the experiments that DFGW significantly outperforms FGW when addressing Domain Adaptation tasks with OT (see the seminal work of [5]) whose goal here is to transfer knowledge from a source graph to a different but related target graph. Interestingly, DW alone is shown to be very competitive while benefiting from a gain in computation time.

The rest of this paper is organized as follows. Section 2 is dedicated to the main background knowledge in Optimal Transport necessary for the rest of the paper; Section 3 is devoted to the presentation of our new heat diffusion-based distances and the derivation of properties giving some insight into their asymptotic behavior. We perform a large spectrum of experiments in Section 4 that give evidence of the efficiency of our distances to address domain adaptation tasks between attributed graphs. We conclude in Section 5.

2 Preliminary knowledge

We present in this section the main background knowledge in Optimal Transport as well as some definitions that will be necessary throughout this paper.

2.1 Optimal transport

Let us consider two empirical probability measures μ and ν, called *source* and *target* distributions, and supported on two sample sets $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$, respectively, lying in some feature space $\mathcal{X}$ and with weights $a = (a_i)_{i=1}^m$, $b = (b_j)_{j=1}^n$ such that $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{y_j}$, where δ is the Dirac function. If $\mathcal{X} = \mathbb{R}^r$ for some integer $r \geq 1$, a matrix representation of X (resp. of Y) is the matrix $\mathbf{X} \in \mathbb{R}^{m \times r}$ (resp. $\mathbf{Y} \in \mathbb{R}^{n \times r}$) which rows are $x_i^\top, 1 \leq i \leq m$ (resp. $y_j^\top, 1 \leq j \leq n$). Let $M = M(X,Y) \in \mathbb{R}_+^{m \times n}$ be a cost matrix, where $M_{ij} \stackrel{def}{=} [d(x_i, y_j)]_{ij}$ is the cost (w.r.t. to some distance function d) of moving x_i on top of y_j. Let $\Pi(a,b)$ be a transportation polytope defined as the set of admissible coupling matrices γ:

$$\Pi(a,b) = \{\gamma \in \mathbb{R}_+^{m \times n} \text{ s.t. } \gamma 1_n = a, \gamma^T 1_m = b\},$$

where γ_{ij} is the mass transported from x_i to y_j and 1_k is the vector of dimension k with all entries equal to one. The p-Wasserstein distance $\mathtt{W}_p^p(\mu,\nu)$ between the source and target distributions is defined as follows:

$$\mathtt{W}_p^p(\mu,\nu) = \min_{\gamma \in \Pi(a,b)} \langle \gamma, M^p(X,Y) \rangle_F, \tag{1}$$

where $\langle ., . \rangle_F$ is the Frobenius inner product and $M^p(X,Y) := (M_{ij}^p)_{ij}$ is the entrywise p-th power of $M(X,Y)$ with an exponent $p > 0$. With $p = 2$ and d the Euclidean distance, if μ and ν are uniform ($a_i = 1/m$, $b_j = 1/n$), then the barycentric projection $\hat{\mathbf{X}}$ of $\mathbf{X}$ can be defined in closed-form [5] as follows: $\hat{\mathbf{X}} = m\gamma^\star \mathbf{Y}$, where $\gamma^\star$ is the optimal coupling of Problem (1).

2.2 Optimal transport on graphs

In order to be able to apply the OT setting on structured data, we need now to formally define the notion of probability measure on graphs and adapt the previous notations. Following [24], let us consider undirected and connected attributed graphs as tuples of the form $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{S})$, where $\mathcal{V}$ and $\mathcal{E}$ are the classic sets of vertices (also called nodes) and edges of the graph, respectively. $\mathcal{F} : \mathcal{V} \to \mathcal{X}$ is a function which assigns a feature vector $x_i \in \mathcal{X}$ (also called a graph signal in [17]) to each vertex v_i of the graph (given an arbitrary ordering of the vertices). $\mathcal{S} : \mathcal{V} \to \mathcal{Z}$ is a function which associates each vertex v_i with some structural representation $z_i \in \mathcal{Z}$, *e.g.* a local description of the

graph, the vertex and a list of its neighbors, etc. We can further define a cost function $C : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ which measures the dissimilarity $C(z, z')$ between two structural representations z, z'. Typically, $C(z, z')$ can capture the length of the shortest path between two nodes. Additionally, if the graph $\mathcal{G}$ is *labeled*, each vertex v_i is also assigned a label from some label space $\mathcal{L}$.

When each vertex of the graph is weighted according to its relative importance, the source and target graphs can be seen as probability distributions,

$$\mu = \sum_{i=1}^{m} a_i \delta_{(x_i, z_i)}, \quad \nu = \sum_{j=1}^{n} b_j \delta_{(y_j, z'_j)} \tag{2}$$

where x_i, z_i are the features / structural representations associated to the vertices of the source graph while y_j,z'_j are those associated to the target one. Equipped with these notations, we can now present the Fused Gromov-Wasserstein (FGW) distance introduced in [24] as the first attempt to define a distance that takes into account both structural and feature information in a OT problem.

Let $\mathcal{G}^s$ (resp. $\mathcal{G}^t$) be a source (resp. target) graph described by its discrete probability measure μ (resp. ν). Let $C^s \in \mathbb{R}^{m \times m}$ and $C^t \in \mathbb{R}^{n \times n}$ be the structure matrices associated with the source and target graphs respectively. The FGW distance is defined via the minimization of a convex combination between (i) the Wasserstein cost matrix which considers the features x_i, y_j associated with the nodes and (ii) the Gromov-Wasserstein cost matrix [19] which takes into account the structure of both graphs. More formally, for each $\alpha \in [0, 1]$, one can define

$$\mathtt{FGW}_p^p(\mu, \nu) = \min_{\gamma \in \Pi(a,b)} \left\{ \sum_{i,j,k,l} \left((1-\alpha) M_{ij}^p + \alpha |C_{ik}^s - C_{jl}^t|^p\right) \gamma_{ij} \gamma_{kl} \right\} \tag{3}$$

where the summation indices are $1 \leq i, k \leq m$ and $1 \leq j, l \leq n$, and the dependency on α is omitted from the notation $\mathtt{FGW}_p(\mu, \nu)$ for the sake of concision. Note that α can be seen as a hyper-parameter which will allow FGW, given the underlying task and data, to find a good compromise between the features and the structures of the graphs. In the special case $\alpha = 0$, we recover the Wasserstein distance $\mathtt{FGW}_p(\mu, \nu \mid \alpha = 0) = \mathtt{W}_p(\mu, \nu)$. By abuse of notation we denote $\mathtt{W}_p(\mu, \nu) = \mathtt{W}_p(\mu_{\mathcal{X}}, \nu_{\mathcal{X}})$, for μ, ν as in (2) and $\mu_{\mathcal{X}} := \sum_{i=1} a_i \delta_{x_i}$, $\nu_{\mathcal{X}} := \sum_j b_j \delta_{y_j}$ their marginals on the feature space $\mathcal{X}$. The case $\alpha = 1$ corresponds to the definition of the Gromov-Wasserstein distance $\mathtt{GW}_p(\mu, \nu) := \mathtt{GW}_p(\mu_{\mathcal{Z}}, \nu_{\mathcal{Z}}) = \mathtt{FGW}_p(\mu, \nu \mid \alpha = 1)$ with $\mu_{\mathcal{Z}}, \nu_{\mathcal{Z}}$ the marginals on the structure space. Roughly speaking, the optimal coupling matrix $\gamma^\star$ will tend to associate two source and target nodes if both their feature and structural representations are similar.

Despite the fact that FGW has been shown to be efficient to address graph classification and clustering tasks [24], we claim that it might face two limitations: (i) given an underlying task, in the presence of noisy features, the best value of the hyper-parameter α in Problem (3) might be close to 1, thus focusing

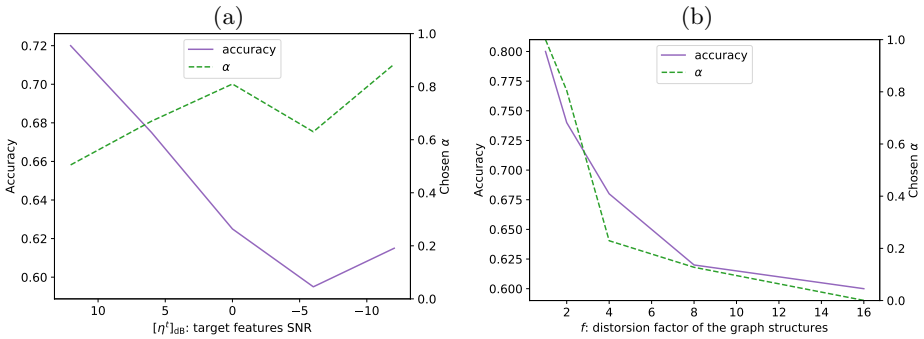


Fig. 1: DA by minimization of FGW (3). Plots display the evolutions of the classification accuracy of the target vertices (purple curves, left axis) along with the corresponding selected hyper-parameter α (dashed green curves, right axis) against two limiting regimes. (a) Features reliability. Source and target graphs structures are i.i.d. ($p_{11} = 0.02$, $p_{22} = 0.03$, $p_{12} = p_{21} = 0.01$) and $\sigma^s = 1$ (i.e. $(x_i = \pm 1 + \mathcal{N}(0,1))_{i=1,\dots,m}$). The signal to noise ratio $[\eta^t]_{\mathrm{dB}} = 20\log_{10}(\eta^t)$ of the target features varies along the X-axis. (b) Structure reliability. Source and target features are i.i.d. ($[\eta^s]_{\mathrm{dB}} = [\eta^t]_{\mathrm{dB}} = -6\,\mathrm{dB}$). The target graph structure progressively deviates from that of $\mathcal{G}^s$ by increasing the probability of inter-class connectivity $p^t_{12} = f \cdot p^s_{12}$, with $f \geq 1$ and (p^s_{uv}) as in (a). The structure distorsion factor f evolves on the X-axis. All plots display median values estimated over 50 i.i.d realisations. The graphs' size is $n = m = 100$.

mainly on the structures and "forgetting" the feature information; (ii) on the other hand, if $\mathcal{G}^s$ and $\mathcal{G}^t$ are structurally very different, the optimal coupling will be likely associated with a tuned parameter α close to 0 and thus skipping the structural information by only leveraging the features.

Fig. 1 illustrates these two limitations in the context of a clustering task, which aims at classifying the nodes of a target graph $\mathcal{G}^t$ by minimizing its FGW distance to a fully labelled source graph $\mathcal{G}^s$. In these experiments, α in (3) is tuned using a circular validation procedure derived from [2]. Note that we use the same following model to generate $\mathcal{G}^s$ and $\mathcal{G}^t$. The graph structures are drawn from a two-class contextual stochastic block model [10] with symmetric connectivity matrix $(p_{uv})_{(u,v)\in\{1,2\}^2}$. The vertices' features are scalar random variables $X_i \sim l(i) + \sigma\mathcal{N}(0,1)$, with mean $l(i) = \pm 1$, depending on the class of the vertex i, and standard deviation σ. We define the signal to noise ratio $\eta = 2/\sigma$. In Fig. 1(a), as η^t decreases, the distribution of the target features smooths out and becomes gradually less representative of the clusters. Accordingly, the Wasserstein distance between the source and the target features' distributions increases in expression (3), thus penalising the classification accuracy of FGW. To cope with this loss of feature's reliability, the tuned value of the hyper-parameter α converges towards 1, yielding a FGW based clustering that mostly relies on the graphs' structures and that necessarily undergoes a drop in performance.

In Fig. 1(b), the features distributions are now similar ($\eta^s = \eta^t$) but the structure of $\mathcal{G}^t$ differs from that of $\mathcal{G}^s$ by densifying the inter-cluster connectivity:

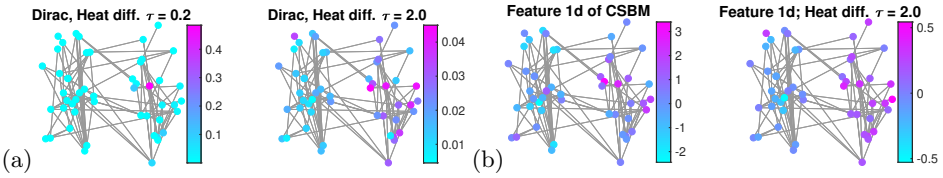


Fig. 2: Illustration of the heat diffusion of features on a graph drawn from a SBM with two blocks [10]. (a) Diffusion from a single vertex: one sees that the diffusion is first spread over the block of the vertex. (b) Diffusion of random features (drawn according to a Contextual SBM [7]): we see that the diffusion is also smoothing the features over each group, helping for their identification.

$p^t_{12} = f \cdot p^s_{12},\ f \geq 1$. As the distorsion factor f increases, the clusters in $\mathcal{G}^t$ tend to blur out. Accordingly, the distance C^t_{jl} between any two vertices y_j and y_l in different classes, statistically reduces in comparison with the same distance C^s_{ik} in $\mathcal{G}^s$. To lessen the penalty effect stemming from this structure mismatch, the hyper-parameter α in the distance (3) rapidly falls towards 0. Since now, the resulting `FGW` based clustering solely exploits the features reliability, it naturally undergoes a performance drop, as it was the case in Fig. 1-(a).

3 Diffusion and DifFused-Gromov Wasserstein distances

To overcome the aforementioned limitations, we suggest in the following to exploit the heat diffusion in graphs and the information provided by the Laplacian matrices to design a new family of Wasserstein distances more robust to local changes in the feature and structure representations.

3.1 From heat diffusion to the Diffusion Wasserstein (`DW`) distance

Let us consider an undirected graph $\mathcal{G}$ with m vertices and its (combinatorial) Laplacian operator $L \in \mathbb{R}^{m \times m}$. We consider a (real or complex valued) feature vector $x \in \mathbb{R}^m$ or $\mathbb{C}^m$ which contains a value for each node. A dynamical process analogous to a diffusion process is obtained by considering the heat equation on the graph: $dw(\tau)/d\tau = -Lw(\tau)$, with $\tau \in \mathbb{R}^+$ and $w(0) = x$. It admits a closed-form solution $w(\tau) = \exp(-\tau L)x$, which is the application of the heat kernel $\exp(-\tau L)$ on the initial feature vector x. Using the GSP interpretation [17], and the functional calculus on the Laplacian operator, the effect of this heat diffusion process is to smooth the features, and the larger τ, the smoother the result. The limit when $\tau \to +\infty$ is even that the solution is constant (on each connected component). It means that τ is a parameter both controlling the smoothing of features and defining a scale of analysis of the structure of the graph.

The properties of this process were used in [9] to introduce a distance between graphs (see also [23]). The original idea was to diffuse Dirac features (1 on a given node, 0, elsewhere), as illustrated in Fig. 2(a), and stack all the diffusing patterns

so that $\exp(-\tau L)$ is a matrix characterizing the graph at some scale τ. Then, to compare two graphs of the same size (m nodes), given their Laplacian L_1 and L_2, the authors of [9] propose to consider $||\exp(-\tau L_1) - \exp(-\tau L_2)||_F$ and keep the minimum value of this quantity over all the possible τ's. While they show that it is a distance, and that it captures well structural (dis)similarities between graphs, its shortcoming is that (i) it can only be used with graphs of the same size, (ii) it forgets about existing features on these graphs and (iii) it cannot be directly used in an OT setting.

To introduce our proposed Diffusion Wasserstein distance, we leverage the closed-form solution of the heat equation applied now to r features $\mathbf{X} \in \mathbb{R}^{m\times r}$ on the graph: $\exp(-\tau L)\mathbf{X}$. Each such term describes now the smoothing of all the features on the graph structure, at a specific characteristic scale τ, as seen in Fig. 2(b). Because it combines features and structure, this solution will be central in the following definition of our new distance between graphs with features.

Definition 1. *Consider a source graph $\mathcal{G}^s$, a target graph $\mathcal{G}^t$ represented through two discrete probability measures μ and ν (cf (2)) with weights vectors $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ and Laplacian matrices $L^s \in \mathbb{R}^{m\times m}$ and $L^t \in \mathbb{R}^{n\times n}$. Let $\mathbf{X} \in \mathbb{R}^{m\times r}$, $\mathbf{Y} \in \mathbb{R}^{n\times r}$ represent the sample sets associated to the features on their vertices.*

Given parameters $0 \leq \tau^s, \tau^t < \infty$, consider the diffused sample sets $\tilde{X}$, $\tilde{Y}$ represented by the matrices $\tilde{\mathbf{X}} = \exp(-\tau^s L^s)\mathbf{X} \in \mathbb{R}^{m\times r}$, $\tilde{\mathbf{Y}} = \exp(-\tau^t L^t)\mathbf{Y} \in \mathbb{R}^{n\times r}$ and define $\tilde{M}(\tau^s, \tau^t) := M(\tilde{X}, \tilde{Y}) \in \mathbb{R}^{m\times n}$, a cost matrix between features that takes into account the structure of the graphs through diffusion operators. We define the Diffusion Wasserstein distance ($\mathtt{DW}$*) between μ and ν as:*

$$\mathtt{DW}_p^p(\mu, \nu \mid \tau^s, \tau^t) = \min_{\gamma \in \Pi(a,b)} \langle \gamma, \tilde{M}^p \rangle. \tag{4}$$

Here again $\tilde{M}^p$ is the entrywise p-th power of $\tilde{M}$. The underlying distance is implicit in $M(\cdot,\cdot)$. For the sake of concision, the dependency on τ^s and τ^t will be omitted from the notation $\mathtt{DW}_p^p(\mu,\nu)$ if not specifically required.

3.2 Role of the diffusion parameters on DW

Denote $D^s = \exp(-\tau^s L^s) \in \mathbb{R}^{m\times m}$, $D^t = \exp(-\tau^t L^t) \in \mathbb{R}^{n\times n}$ the diffusion matrices, which depend on the (symmetric) Laplacians $L^s \in \mathbb{R}^{m\times m}$, $L^t \in \mathbb{R}^{n\times n}$ and the diffusion parameters $0 \leq \tau^s, \tau^t < \infty$. Given $1 \leq i \leq m, 1 \leq j \leq n$ let $x_i, y_j \in \mathbb{R}^r$ be the features on nodes i on $\mathcal{G}^s$ and j on $\mathcal{G}^t$, i.e. respectively the i-th row of $\mathbf{X} \in \mathbb{R}^{m\times r}$ and the j-th row of $\mathbf{Y} \in \mathbb{R}^{n\times r}$, and similarly for $\tilde{x}_i, \tilde{y}_j \in \mathbb{R}^r$ built from $\tilde{\mathbf{X}} = D^s\mathbf{X}$ and $\tilde{\mathbf{Y}} = D^t\mathbf{Y}$. Observe that $\tilde{M}(\tau^s, \tau^t)$ and $\mathtt{DW}_p^p(\mu, \nu \mid \tau^s, \tau^t)$ depend on the diffusion parameters τ^s, τ^t. When $\tau^s = \tau^t = 0$, since $D^s = I_m$ and $D^t = I_n$ we have $\tilde{M}(0,0) = M$ hence

$$\mathtt{DW}_p^p(\mu, \nu \mid 0, 0) = \mathtt{W}_p^p(\mu, \nu), \tag{5}$$

i.e., DW generalizes the Wasserstein distance W.

From now on we focus on $\mathtt{DW}$ defined using a cost matrix $\tilde{M}$ based on the Euclidean distance and $p = 2$. Denote

$$\begin{cases} M_{ij}^2 = \|x_i - y_j\|_2^2 \\ \tilde{M}_{ij}^2 = \|\tilde{x}_i - \tilde{y}_j\|_2^2 \end{cases}$$

the squared entries of the cost matrices associated to the Wasserstein ($\mathtt{W}_2$) and Diffusion Wasserstein ($\mathtt{DW}_2$) distances. The next proposition establishes the asymptotic behavior of $\mathtt{DW}_2^2(\mu, \nu)$ with respect to τ^s and τ^t as well as an upper bound expressed in terms of a uniform coupling matrix. Denote $\bar{\gamma} \in \Pi(a, b) \subset \mathbb{R}_+^{m \times n}$ this (uniform) transport plan such that $\bar{\gamma}_{i,j} = 1/nm, \forall i, j$.

Proposition 1. *Consider Laplacians $L^s \in \mathbb{R}^{m \times m}$, $L^t \in \mathbb{R}^{n \times n}$ associated to two undirected connected graphs ($\mathcal{G}^s$ and $\mathcal{G}^t$) and two matrices $\mathbf{X} \in \mathbb{R}^{m \times r}, \mathbf{Y} \in \mathbb{R}^{n \times r}$ representing the sample sets $x_i \in \mathbb{R}^r, 1 \leq i \leq m$ and $y_j \in \mathbb{R}^r, 1 \leq j \leq n$ (associated to their vertices). Consider the associated measures μ, ν with flat weight vectors $a = 1_m/m$, $b = 1_n/n$. We have*

$$\lim_{\tau^s, \tau^t \to \infty} \mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t) = \|\frac{1}{m}\sum_i x_i - \frac{1}{n}\sum_j y_j\|_2^2. \tag{6}$$

Moreover, the function $(\tau^s, \tau^t) \mapsto \langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle$ is non-increasing with respect to τ^s and with respect to τ^t and also satisfies for each $0 \leq \tau^s, \tau^t < \infty$

$$\|\frac{1}{m}\sum_i x_i - \frac{1}{n}\sum_j y_j\|_2^2 \leq \mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t) \leq \langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle \leq \langle \bar{\gamma}, M^2\rangle \tag{7}$$

$$\lim_{\tau^s, \tau^t \to \infty} \langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle = \|\frac{1}{m}\sum_i x_i - \frac{1}{n}\sum_j y_j\|_2^2. \tag{8}$$

The proof is in the supplementary material.

Remark 1. The reader can check that in the proof we also establish that

$$\langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle = \quad \langle \bar{\gamma}, M^2\rangle + \left[\sum_{i=2}^{m} \left(e^{-2\tau^s \lambda_i^s} - 1\right) \|\widehat{x}_i\|_2^2 + \sum_{j=2}^{n} \left(e^{-2\tau^t \lambda_j^t} - 1\right) \|\widehat{y}_j\|_2^2\right].$$

Contrary to its non-increasing upper bound $\langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle$, the squared Diffusion Wasserstein distance $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$ may not behave monotonically with τ^s, τ^t. Even though $\mathtt{DW}_2^2(\mu, \nu \mid 0, 0) = \mathtt{W}_2^2(\mu, \nu)$ we may thus have $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t) > \mathtt{W}_2^2(\mu, \nu)$ for some values of τ^s, τ^t. The following gives a sufficient condition to ensure that (in expectation) $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$ does not exceed $\mathtt{W}_2^2(\mu, \nu)$.

Proposition 2. *Consider integers $m, n, r \geq 1$, $a \in \mathbb{R}_+^m$, $b \in \mathbb{R}_+^n$ such that $\sum_i a_i = 1 = \sum_j b_j$, two random Laplacians $L^s \in \mathbb{R}^{m \times m}$, $L^t \in \mathbb{R}^{n \times n}$ drawn independently according to possibly distinct probability distributions, two random feature matrices $\mathbf{X} \in \mathbb{R}^{m \times r}$, $\mathbf{Y} \in \mathbb{R}^{n \times r}$, and $0 \leq \tau^s, \tau^t < \infty$. If $\mathbb{E}\ \tilde{M}_{ij}^2(\tau^s, \tau^t) \leq M_{ij}^2\ \forall (i, j)$, then $\mathbb{E}\ \mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t) \leq \mathtt{W}_2^2(\mu, \nu)$.*

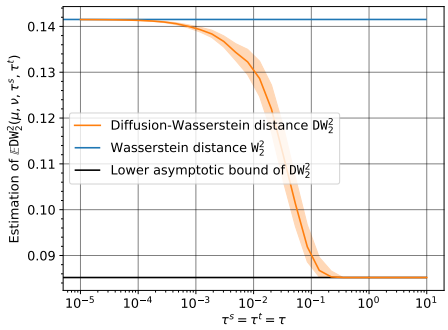


Fig. 3: Numerical illustration of Proposition 2, with distance $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$ defined in Eq. (4). $\mathbb{E}\,\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$ is empirically estimated from 2500 independent realisations of source and target graphs drawn from the same stochastic block model, with $p_{11} = 0.32$, $p_{22} = 0.32$, $p_{12} = p_{21} = 0.02$ and $n = m = 100$. The feature vectors $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$ are arbitrarily chosen and remain fixed across all realisations, to restrict randomness only to the structures. Empirical median (solid line) and quartiles 1 and 3 (strip) of $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s = \tau, \tau^t = \tau)$ are plotted against τ and compared to the Wasserstein distance $\mathtt{W}_2^2(\mu, \nu) = \mathtt{DW}_2^2(\mu, \nu \mid 0, 0)$ (upper bound) and to the asymptotic regime given in Eq. (6), when $\tau \to +\infty$ (lower plateau).

Remark 2. The case where the Laplacians and/or the features are deterministic is covered by considering probability distributions that are Diracs.

Proof. For brevity we omit the dependency on μ, ν.

$$\mathbb{E}\mathtt{DW}_2^2 = \mathbb{E} \inf_{\gamma \in \Pi(a,b)} \langle \tilde{M}^2, \gamma \rangle \leq \inf_{\gamma} \mathbb{E}\langle \tilde{M}^2, \gamma \rangle = \inf_{\gamma} \langle \mathbb{E}\tilde{M}^2, \gamma \rangle \leq \inf_{\gamma} \langle M^2, \gamma \rangle = \mathtt{W}_2^2. \qquad \square$$

Moreover, by [18, Remark 2.19] we have $\mathtt{W}_2^2(\mu, \nu) \geq \| \frac{1}{m} \sum_{i=1}^m x_i - \frac{1}{n} \sum_{j=1}^n y_j \|_2^2$. If $\mathbf{X}$ and $\mathbf{Y}$ are such that in fact $\mathtt{W}_2^2(\mu, \nu) > \| \frac{1}{m} \sum_{i=1}^m x_i - \frac{1}{n} \sum_{j=1}^n y_j \|_2^2$ then for sufficiently large τ^s, τ^t we must have $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t) < \mathtt{W}_2^2(\mu, \nu)$.

However we can find examples such that $\mathtt{DW}_2^2(\mu, \nu) > \mathtt{W}_2^2(\mu, \nu)$ and $\mathbb{E}\mathtt{DW}_2^2(\mu, \nu) > \mathtt{W}_2^2(\mu, \nu)$ for all $0 < \tau^s, \tau^t < \infty$. For this, it is sufficient to choose $\mathbf{X} = \mathbf{Y}$, so that $\mathtt{W}_2^2(\mu, \nu) = 0$, and deterministic or random graphs and parameters τ^s, τ^t such that $\exp(-\tau^s L^s)\mathbf{X}$ is not equal (even up to permutation) to $\exp(-\tau^t L^t)\mathbf{Y}$, so that (almost surely) $\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau_t) > 0$.

Figure 3 illustrates the results of Propositions 1 and 2, where we empirically estimated $\mathbb{E}\,\mathtt{DW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$, and plotted its evolution against $\tau = \tau^s = \tau^t$ (experimental conditions are detailed in the legend of Fig. 3). Trivially, we verify that $\mathtt{DW}_2^2(\mu, \nu \mid 0, 0) = \mathtt{W}_2^2(\mu, \nu)$. But, more importantly, we observe that $\mathbb{E}\,\mathtt{DW}_2^2$ systematically stands below $\mathtt{W}_2^2$, confirming thus the prediction of Proposition 2, and converges towards the theoretical bound given in Eq. (6) of Proposition 1, when $\tau \to \infty$. Interestingly also, although we know from the counter-example $\mathbf{X} = \mathbf{Y}$ above, that it is not true in general, the trend of $\mathbb{E}\,\mathtt{DW}_2^2$ in Fig. 3 seems to validate

the conjecture whereby it is often a non-increasing function of the diffusion scale τ. However, we still lack the theoretical conditions that warrant the result of Prop. 1 on $(\tau^s, \tau^t) \mapsto \langle \bar{\gamma}, \tilde{M}^2(\tau^s, \tau^t)\rangle$ to extend to $\min_{\gamma \in \Pi(a,b)} \langle \gamma, \tilde{M}^2(\tau^s, \tau^t)\rangle$.

From an algorithmic complexity perspective, notice that compared to `FGW`, our new distance `DW` allows us to get free from the costly term in $\mathcal{O}(m^2n^2)$ corresponding to the Gromov part of `FGW` (even though when $p = 2$ one can compute this term more efficiently in $\mathcal{O}(m^2n + n^2m)$ [19]), while still accounting for both the structure and the features of the graphs. Our study on the computational time of the state of the art methods in Section 4 will give evidence that `DW` is the cheapest way to compute a distance encompassing both sources of information.

3.3 DifFused Gromov Wasserstein (`DFGW`)

It is worth noticing that the heat diffusion operator, as defined in Section 3.1, can be seen as a reweighting scheme applied over the node features leading to the new cost matrix $\tilde{M}(\tau^s, \tau^t)$. Notice also that the latter can be precomputed during a preprocess. Therefore, by plugging $\tilde{M}(\tau^s, \tau^t)$ in place of M in `FGW`, we get for free a family (parameterized by $\alpha \in [0, 1]$) of so-called DifFused Gromov Wasserstein (`DFGW`) distances defined as follows:

Definition 2.

$$\texttt{DFGW}_p^p(\mu, \nu) = \min_{\gamma \in \Pi(a,b)} \left\{ \sum_{i,j,k,l} \left((1-\alpha)\tilde{M}_{ij}^p + \alpha |C_{ik}^s - C_{jl}^t|^p \right) \gamma_{ij}\gamma_{kl} \right\}, \quad (9)$$

where the summation indices are $1 \leq i, k \leq m$ and $1 \leq j, l \leq n$ for a source graph $\mathcal{G}^s$ (resp. a target graph $\mathcal{G}^t$) of size m (resp. n), and the dependency on τ^s,τ^t and the considered distance d is implicit in $\tilde{M}$.

A simple lower bound on `DFGW` holds with arguments similar to those of [24] leading to a lower bound on `FGW` in terms of `W` and `GW`.

Lemma 1. *Following [24], $\forall p$, $\texttt{DFGW}_p^p(\mu, \nu)$ is lower-bounded by the straightforward interpolation between $\texttt{DW}_p^p(\mu, \nu)$ and $\texttt{GW}_p^p(\mu, \nu)$:*

$$\texttt{DFGW}_p^p(\mu, \nu) \geq (1-\alpha)\texttt{DW}_p^p(\mu, \nu) + \alpha \texttt{GW}_p^p(\mu, \nu)$$

Proof. By definition, for $\gamma \in \Pi(a, b)$ we have $\texttt{GW}_p^p(\mu, \nu) \leq \sum_{i,j,k,l} |C_{ik}^s - C_{jl}^t|^p \gamma_{ij}\gamma_{kl}$. Similarly, since $\sum_{k,l} \gamma_{k,l} = 1$, we get $\texttt{DW}_p^p(\mu, \nu) \leq \sum_{i,j} \tilde{M}_{ij}^p \gamma_{ij} = \sum_{i,j,k,l} \tilde{M}_{ij}^p \gamma_{ij}\gamma_{k,l}$. As a result,

$$(1-\alpha)\texttt{DW}_p^p(\mu, \nu) + \alpha \texttt{GW}_p^p(\mu, \nu) \leq \sum_{i,j,k,l} \left((1-\alpha)\tilde{M}_{ij}^p + \alpha |C_{ik}^s - C_{jl}^t|^p \right) \gamma_{ij}\gamma_{kl}.$$

As this holds for every $\gamma \in \Pi(a, b)$ and $\texttt{DFGW}_p^p(\mu, \nu)$ is the infimum of the right hand side, this establishes the result. □

Just as the Diffusion Wasserstein distance, DFGW depends on the diffusion parameters τ^s, τ^t, and we have

$$\texttt{DFGW}_p^p(\mu, \nu \mid 0, 0) = \texttt{FGW}_p^p(\mu, \nu). \tag{10}$$

Therefore, DFGW generalizes the Fused Gromov Wasserstein distance. In the same spirit as Proposition 1, the next proposition establishes the asymptotic behavior of $\texttt{DFGW}_2^2(\mu, \nu \mid \tau^s, \tau^t)$ with respect to τ^s and τ^t.

Proposition 3. *With the notations and assumptions of Proposition 1 we have*

$$\texttt{DFGW}_2^2(\mu, \nu \mid \tau^s, \tau^t) \geq (1-\alpha)\|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{n}\sum_{j=1}^{n} y_j\|_2^2 + \alpha\texttt{GW}_2^2(\mu, \nu),\ \forall\, \tau^s, \tau^t$$

$$\lim_{\tau^s, \tau^t \to \infty} \texttt{DFGW}_2^2(\mu, \nu \mid \tau^s, \tau^t) = (1-\alpha)\|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{n}\sum_{j=1}^{n} y_j\|_2^2 + \alpha\texttt{GW}_2^2(\mu, \nu).$$

The proof is in the supplementary material.

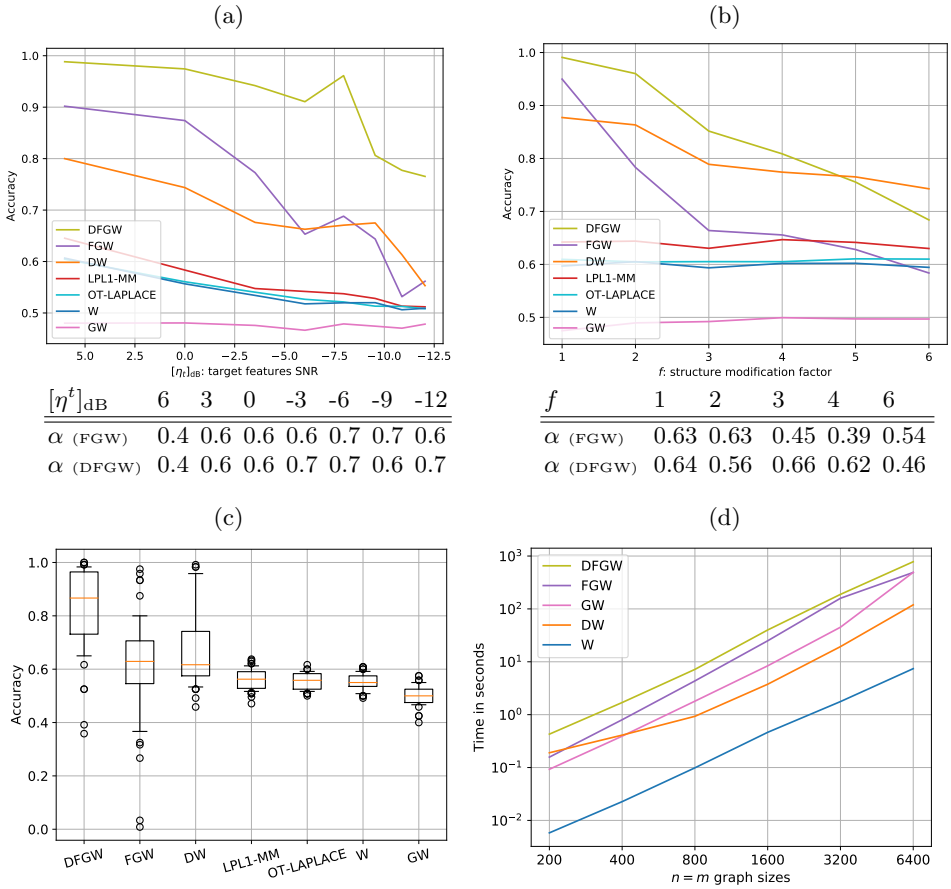
3.4 Metric properties of DW and DFGW

Recall that DW can be seen as a generalization of the Wasserstein distance W which leverages the diffusion operator over the features. Moreover, it is known that when the cost matrix $M_{ij} \overset{def}{=} [d(x_i, y_j)]_{ij}$ is associated to a distance d, W defines a metric. The next proposition shows that the diffusion does not change this metric property up to a natural condition.

Proposition 4. *For $p \in [1, \infty)$ and $0 \leq \tau^s, \tau^t < \infty$, the Diffusion Wasserstein $\texttt{DW}_p(\cdot, \cdot \mid \tau^s, \tau^t)$ defines a pseudo-metric: it satisfies all the axioms of a metric, except that $\texttt{DW}_p(\mu, \nu) = 0$ if, and only if, $\mathcal{T}(\mu) = \mathcal{T}(\nu)$. Here, $\mathcal{T}$ is the function which maps $\mu = \sum_{i=1}^m a_i \delta_{x_i, z_i}$ into $\tilde{\mu} = \mathcal{T}(\mu) = \sum_{i=1}^m a_i \delta_{\tilde{x}_i}$ where $\tilde{x}_i \in \mathbb{R}^k$ is built in a deterministic manner from the diffusion matrix D^s (which is itself a function of μ through the z_i's) and corresponds to the i-th row of $\tilde{\mathbf{X}} = D^s\mathbf{X}$.*

Proof. According to Def. 1, DW is defined between two probability measures $\mu = \sum_{i=1}^m a_i \delta_{(x_i, z_i)}$ and $\nu = \sum_{j=1}^n b_j \delta_{(y_j, z'_j)}$ with (x_i, z_i) and (y_j, z'_j) lying in some joint space $\mathcal{X} \times \mathcal{Z}$ encoding both the feature and the structure information of two source and target vertices, respectively. Since $\texttt{DW}_p(\mu, \nu) = \texttt{W}_p(\tilde{\mu}, \tilde{\nu}) = \texttt{W}_p(\mathcal{T}(\mu), \mathcal{T}(\nu))$, the proposition follows from the metric property of $\texttt{W}_p(\cdot, \cdot)$. □

On the other hand, it has been shown in [24] that when C^s and C^t are distance matrices the Fused Gromov Wasserstein $\texttt{FGW}_1$ defines a metric and that $\texttt{FGW}_p^p$ defines a semimetric (i.e., a relaxed version of the triangle inequality holds) when $p > 1$. Since DW is used in our DifFused Gromov Wasserstein distance in place of the Wasserstein counterpart in FGW, the same metric properties hold for DFGW.



$[\eta^t]_{\mathrm{dB}}$	6	3	0	-3	-6	-9	-12
α (FGW)	0.4	0.6	0.6	0.6	0.7	0.7	0.6
α (DFGW)	0.4	0.6	0.6	0.7	0.7	0.6	0.7

f	1	2	3	4	6
α (FGW)	0.63	0.63	0.45	0.39	0.54
α (DFGW)	0.64	0.56	0.66	0.62	0.46

Fig. 4: Comparison of OT methods in a domain adaptation task between graphs. We consider attributed graphs whose structures follow a contextual stochastic block model and attributes a mixture Gaussian model. Y-axes of plots (a)–(b) represent the classification accuracies. Hyper-parameters and mean performance are determined from two distinct sets of 50 i.i.d. realisations each. (a) Structures of $\mathcal{G}^s$ and $\mathcal{G}^t$ are identical ($p_{11} = p_{22} = 0.4$, $p_{12} = p_{21} = 0.05$, $n = m = 250$). SNR of the source features is fixed ($[\eta^s]_{\mathrm{dB}} = 20\log_{10}(\eta^s) = 6\,\mathrm{dB}$) and σ^t of features $\mathbf{Y}_j \sim l(j) + \sigma^t \mathcal{N}(0,1)$ varies according to $[\eta^t]_{\mathrm{dB}}$ along the X-axis. (b) Features SNR $[\eta^s]_{\mathrm{dB}} = [\eta^t]_{\mathrm{dB}} = 6\,\mathrm{dB}$. The target graph follows a SBM with symmetric connectivity matrix $p^t_{12} = p^s_{12} = 0.05$, $p^t_{11} = p^s_{11} = 0.4$ and $p^t_{22} = p^s_{22}/f$ with $p^s_{22} = 0.4$ and f variable on the X-axis. Tables beneath the plots give the tuned hyper-parameters values for each case. (c) Performance when uncertainty bears on the features and on the structures simultaneously ($[\eta^t]_{\mathrm{dB}} = 0\,\mathrm{dB}$, $f = 3$).
(d) Computing times wrt the size of the graphs $n = m$ ($[\eta^s]_{\mathrm{dB}} = [\eta^t]_{\mathrm{dB}} = 0\,\mathrm{dB}$, $f = 1$).

4 Numerical experiments

In this section, we evaluate our diffusion distances on domain adaptation (DA) tasks between attributed graphs. We address the most complicated scenario where a source domain and a target domain are considered and the label information is only available in the former. Data from the two domains are supposedly drawn from different but related distributions and the goal is to reduce this distribution discrepancy while benefiting from the supervised information from the source [20]. Note that when dealing with a DA task between attributed graphs, the divergence can come from three situations: (i) a shift in the feature representation of the source/target nodes; (ii) a difference in the graph structures; (iii) both of them. In this section, we study these three settings.

Under different experimental conditions, we compare in these DA tasks the relevance of our diffusion distances `DW` defined in (4) and `DFGW` defined in (9), to state-of-the-art OT-based distances: `W`: Wasserstein (1); `GW`: Gromov-Wasserstein [19]; `FGW`: Fused Gromov-Wasserstein (3); `LPL1-MM`: that corresponds to the Wasserstein problem associated with a label regularization [4]; `OT-LAPLACE`: the same as the latter with a Laplacian regularization [5]. In the following experiments, we use the same Contextual Stochastic Block Model [10] and the same Gaussian mixture model as the ones described for experiments of Figure 1, to generate the graph structures and the nodes' features, respectively. Although both source and target graphs are labelled, for OT methods implying hyper-parameters, we tune them with the circular validation procedure derived in [2] that only uses the ground truth labels on the vertices of $\mathcal{G}^s$. As for the ground truth on the vertices of $\mathcal{G}^t$, they only serve to evaluate the classification performance (accuracy) of the methods. The procedure is that each target node inherits the label of the class from which it received most of its mass by the transport plan that is solution of the optimization problem. The tuning of the hyper-parameters ($\alpha \in [0,1]$, $\tau \in [10^{-3}, 10^{-0.5}]$ and the regularization parameters of `LPL1-MM` and `OT-LAPLACE`) and the performance evaluation are performed on two different sets of 50 i.i.d. realizations. Unless specified, we display the empirical mean values.

All codes are written in Python. Graphs are generated using the Pygsp [6] library; optimal transport methods use the implementation in POT [8].

Resilience to features' uncertainty. We start illustrating the effect of the heat diffusion when the target features are weakly representative of the underlying clusters, leading to a divergence between the source and the target domains. This is the case in Figure 4(a), where, as the target signal to noise ratio η^t decays, it smears out the modes of the features' distribution and makes the Wasserstein distance inefficient at discerning them. As a result, all transport methods relying uniquely on information from the features behave poorly and fail, in the limit of $\eta^t \to 0$, at inferring from $\mathcal{G}^s$, the two classes in $\mathcal{G}^t$. On the opposite, hybrid methods that also exploit similarity of the graphs' structure, naturally show better performance. Incidentally, we verified that the puzzling weak performance of Gromov-Wasserstein do not negate its capacity at clustering $\mathcal{G}^t$ correctly, but stems from the global labelling mismatch incurred by the

symmetry of the SBM connectivity matrices. Now, concentrating on our difFused Gromov Wasserstein distance, it systematically and significantly[1] outperforms FGW, whatever the value of η^t. As the diffusion operator D^t estimates the conditional mean value ($l(j) = \pm 1$) of each vertex j by locally averaging the features of its neighbouring nodes in the graph, DFGW turns out to be less sensitive to the noise amplitude σ_t. Notice though, that as $\tau^t \to \infty$, $D^t\mathbf{Y}$ converges to the global mean of $\mathbf{Y}$, we need to limit τ^t to the range $[10^{-3}, 10^{-0.5}]$, that is to say, before DW reaches the lower plateau in Fig. 3. As confirmed by the similar evolution of the optimized α for FGW and DFGW, the use of the Diffusion Wasserstein distance to define DFGW is responsible for the accuracy gain compared to FGW.

Interestingly, at low signal-to-noise ratio, the mere DW is able to compensate for the lack of an explicit metric between the graphs' structures, by retrieving it from the action of the diffusion operators D^s and D^t on the features.

Resilience to structures' uncertainty. Figure 4(b) illustrates the robustness of the diffusion distances with respect to inconsistencies in the structures of the source and target graphs. The plots display the DA performance achieved by the different OT methods, when source and target features follow the same distribution, but the graphs become less and less alike and hence, the structure information becomes less reliable in the context of an OT task.

As expected, the methods relying solely on the Wasserstein distance between the features perform constantly, with an accuracy level that is comparable to that of Fig. 4(a) at high SNRs. We also observe that GW continues to get poor performance for large values of the distorsion factor f, because now, it really is unable to infer the clusters from the graphs' structures. More remarkably, the performance of FGW rapidly degrades once source and target graphs start to slightly differ. This is confirmed by the corresponding trend of α, which overall decays with f, decreasing the contribution of the GW distance in (3). In comparison, although the accuracy obtained with DifFused Gromov-Wasserstein suffers from the growing inconsistency between graph structures too, it always remains above the curve for FGW. These results clearly demonstrate that the DW distance bears a structure information that the optimal solution of Eq. (9) is able to leverage and to combine with the Gromov-Wasserstein cost matrix.

But the certainly most striking result of Fig. 4(b) comes from the performance of the transport plan resulting from minimizing the Diffusion Wasserstein distance alone. Although its accuracy never outperforms that of DFGW, it not only surpasses FGW once $f > 1$, but it also degrades at a slower rate than the two competing methods. One possible explanation, is that the circular procedure [2] that we used to determine α, has its own limits when $\mathcal{G}^s$ and $\mathcal{G}^t$ are drastically different, and it certainly does not yield the best compromise possible between the features and the structure modalities. DW does not entail to tune this trade-off, as it naturally embeds both modalities in its definition (4).

[1] A paired Student's t-test rejects the null hypothesis with p-value equal to 2.10^{-11}.

Resilience to domain divergence. The box plots of Fig. 4(c) compare the methods in a more realistic (and tricky) task of domain adaptation, when both the features and the structures in source and target spaces, differ. This scenario is a combination of the experimental settings of Fig. 4 (a) and (b). Once again, it highlights the advantage of diffusion based distances at optimizing a transport plan between the labels' nodes of attributed graphs, when these latter share some common ground but remain highly variable with regard to each other. A situation that is most likely to occur in real world applications.

Computing time. One major drawback of `FGW` is the complexity in $\mathcal{O}(m^2n^2)$ (going down to $\mathcal{O}(m^2n + n^2m)$ for $p = 2$ [19]) due to the quadratic problem optimization with respect to $\gamma_{ij}\gamma_{kl}$ in its definition (3). Naturally, the same pitfall holds for `DFGW`. It gets even worse since in addition, it encompasses the calculation of the diffusion operator $D^s\mathbf{X}$. On the other hand, we know from [1], that the calculation cost of $D^s\mathbf{X}$, hence of the `DW` distance, can be done in $\mathcal{O}(m^2r)$ for $\mathbf{X} \in \mathbb{R}^{m\times r}$. Indeed, in the experiments reported in Fig. 4(d), the computing time of the Diffusion Wasserstein distance remains one order of magnitude below that of `DFGW` and `FGW`. Then, in `DFGW`, the cost for computing `GW` always prevails over that of `DW`, inducing in (9), a limited overhead cost as compared to that of `FGW`. In this last series of experiment, our goal was not to address a domain adaptation task anymore, that is why we deemed irrelevant to compare with the computing times of `LPL1-MM` and `OT-LAPLACE` methods.

5 Conclusion

We exploit in this paper the heat diffusion in attributed graphs to design a family of Wasserstein-like distances that take into account both the feature and the structural information of the graphs. We study their asymptotic behavior and prove that they satisfy metric properties. This paper opens the door to several future lines of work. While `DW` benefits from a cheaper computational cost, `DFGW` systematically surpasses `FGW` but still suffers from the same limited scalability to large graphs. A promising solution would consist in relaxing the quartic constraint of the Gromov counterpart by resorting to random filtering on graphs [22] allowing for an approximation of the cost matrix. In such a case, the conservation of the metric and asymptotic properties will have to be studied. One could also use Chebyshev polynomials to approximate and speed up the heat kernel computing. Finally, other applications in graph classification or in clustering exploiting Diffusion Wasserstein barycenters can be envisioned.

References

1. Al-Mohy, A., Higham, N.: Computing the action of the matrix exponential, with an application to exponential integrators. SIAM J. Sc. Comp. **33**(2), 488–511 (2011)
2. Bruzzone, L., Marconcini, M.: Domain adaptation problems: A dasvm classification technique and a circular validation strategy. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 770–787 (2010)

3. Chung, F.: The heat kernel as the pagerank of a graph. Proceedings of the National Academy of Sciences **104**(50), 19735–19740 (2007)
4. Courty, N., Flamary, R., Tuia, D.: Domain adaptation with regularized optimal transport. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 274–289. Springer (2014)
5. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(9), 1853–1865 (2017)
6. Defferrard, M., Martin, L., Pena, R., Perraudin, N.: Pygsp: Graph signal processing in python (Oct 2017). https://doi.org/10.5281/zenodo.1003158
7. Deshpande, Y., Sen, S., Montanari, A., Mossel, E.: Contextual stochastic block models. In: NeurIPS 2018, Montréal, Canada. pp. 8590–8602 (2018)
8. Flamary, R., Courty, N.: Pot python optimal transport library (2017), `https://github.com/rflamary/POT`
9. Hammond, D.K., Gur, Y., Johnson, C.R.: Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. In: GlobalSIP. pp. 419–422. IEEE (2013)
10. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: First steps. Social networks **5**(2), 109–137 (1983)
11. Kantorovich, L.: On the translocation of masses. Doklady of the Academy of Sciences of the USSR **37**, 199–201 (1942)
12. Kriege, N.M., Johansson, F.D., Morris, C.: A survey on graph kernels. Applied Network Science **5**(1), 6 (2020)
13. Maretic, H.P., Gheche, M.E., Chierchia, G., Frossard, P.: GOT: an optimal transport framework for graph comparison. CoRR **abs/1906.02085** (2019)
14. Maretic, H.P., Gheche, M.E., Minder, M., Chierchia, G., Frossard, P.: Wasserstein-based graph alignment. cs.LG **abs/2003.06048** (2020)
15. Monge, G.: Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie royale des sciences de Paris (1781)
16. Mémoli, F.: Gromov-wasserstein distances and the metric approach to object matching. Found. Comput. Math. **11**(4), 417–487 (2011)
17. Ortega, A., Frossard, P., Kovačević, J., Moura, J., Vandergheynst, P.: Graph signal processing: Overview, challenges, and applications. Proceedings of the IEEE **106**(5), 808–828 (2018)
18. Peyré, G., Cuturi, M.: Computational Optimal Transport. arXiv (Mar 2018)
19. Peyré, G., Cuturi, M., Solomon, J.: Gromov-wasserstein averaging of kernel and distance matrices. In: Int. Conf. on Machine Learning. vol. 48, pp. 2664–2672 (2016)
20. Redko, I., Morvant, E., Habrard, A., Sebban, M., Bennani, Y.: Advances in Domain Adaptation Theory. Elsevier, ISBN 9781785482366, p. 187 (Aug 2019)
21. Thanou, D., Dong, X., Kressner, D., Frossard, P.: Learning heat diffusion graphs. IEEE Trans. on Sig. and Info. Proc. over Networks **3**(3), 484–499 (2017)
22. Tremblay, N., Puy, G., Gribonval, R., Vandergheynst, P.: Compressive Spectral Clustering. In: 33rd Int. Conf. on Machine Learning. New York, USA (Jun 2016)
23. Tsitsulin, A., Mottin, D., Karras, P., Bronstein, A., Müller, E.: Netlsd: Hearing the shape of a graph. In: ACM Int. Conf. on Knowledge Discovery & Data Mining, London (UK). p. 2347–2356. New York, USA (2018)
24. Vayer, T., Courty, N., Tavenard, R., Chapel, L., Flamary, R.: Optimal transport for structured data with application on graphs. In: Int. Conf. on Machine Learning. vol. 97, pp. 6275–6284 (2019)