Fused Gromov-Wasserstein Distance for Heterogeneous and Temporal Graphs

Silvia Beddar-Wiesing[1][0000−0002−2984−2119] and Dominik Köhler[2][0009−0007−8049−2648]

[1] University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany
[2] University of Paderborn, Warburgerstr. 100, 33098 Paderborn, Germany
s.beddarwiesing@uni-kassel.de, dominik.koehler@upb.de

Abstract. The Wasserstein distance, originating from optimal transport theory, is a powerful metric for comparing distributions and has been extensively adapted for various domains such as structured data and high-dimensional distributions. Its use in Machine Learning has advanced applications like graph embeddings, adversarial learning, variational autoencoders, and generative models. However, current applications on graphs are limited to static graphs with homogeneous attributes, restricting their utility for heterogeneous or dynamic graphs. This work in progress addresses these limitations by proposing the Attributed Fused Gromov-Wasserstein Distance (AFGWD) for graphs with diverse attribute spaces and the Temporal Fused Gromov-Wasserstein Distance (TFGWD) for discrete- and continuous-time dynamic heterogeneous graphs. Efficient computation strategies using existing approximation methods are discussed to tackle the computational challenges of Wasserstein distances. These advancements aim to broaden the applicability of Wasserstein distances in complex and dynamic graph data scenarios, paving the way for future research.

1 Introduction

The Wasserstein distance originates in optimal transport theory and represents a distance metric between two distributions [32]. It has already been mathematically analyzed thoroughly and adapted for various applications. Among them are extensions for distributions in different spaces (Gromov-Wasserstein distance [18]), structured data (Fused Gromov-Wasserstein distance [29]), Riemannian manifolds (Spectral Wasserstein distance [19]), high-dimensional distributions (Sliced Gromov-Wasserstein distance [26]), or other geometric domains (Convolutional Wasserstein distance [22]). In addition, various Machine Learning approaches already integrate Wasserstein distances to learn on different domains. For example, it is used as a regularizer for embedding attributed graphs [13], adversarial learning on knowledge graphs [7], adversarial transfer learning [3], training variational autoencoders [27] and generative adversarial networks [1], for determining barycenters of graphs [2] or time series [4], to reconstruct graphs [24], generate images [15, 20], or detect drifts [30]. Real-world applications are diverse, from EEG signal reconstruction with a Wasserstein GAN [16] to the planning of power systems [5] or intelligent fault diagnosis [3].

However, when applying Wasserstein distances on graphs, the extensions are limited to static graphs with attributes from the same space. This limitation excludes the applicability to a wide range of applications that use more complex data types, such as heterogeneous or dynamic graphs. To address this gap, we propose an **Attributed Fused Gromov-Wasserstein distance (AFGWD)** for graphs with different attribute spaces in this work in progress. This extension considers both the structural information and the different attribute spaces and directly serves as a distance metric for heterogeneous graphs. Furthermore, we introduce a **Temporal Fused Gromov-Wasserstein distance (TFGWD)** for dynamic graphs that are represented either as a sequence of graph snapshots (discrete-time representation) or as a sequence of events (continuous-time representation) and are potentially heterogeneous. Subsequently, we present a brief overview of efficient calculation strategies based on existing approximation methods, which address the high computation effort of Wasserstein distances. Finally, we discuss potential subsequent projects for future research.

2 Preliminaries

This section introduces graphs and the Fused-Gromov-Wasserstein Distance (FGWD). For this purpose, definitions of different types of graphs are taken from [25], followed by the different Wasserstein distances from [28] that are important for this work.

2.1 Graphs

Definition 2.1.01 (Graph) Let $\mathcal{V} \subset \mathbb{N}$ be a (finite) set of nodes,
$\mathcal{E} \subseteq \{\{u, v\} \mid u, v \in \mathcal{V}\}$ a set of undirected edges. An **(undirected) graph** is then defined as tuple $g = (\mathcal{V}, \mathcal{E})$ of the node and edge sets.

Graphs can be attributed, i.e., there exist additional information for the nodes and edges, respectively. These can include, e.g., vectorial, graphical, temporal, or textual information.

Definition 2.1.02 (Attributed Graph) Let $\mathcal{V}$ and $\mathcal{E}$ be sets of nodes and edges as above. Further, let $\omega : \mathcal{V} \to \mathcal{A}$ and $\theta : \mathcal{E} \to \mathcal{B}$ be mappings from the node and edge sets to attribute spaces $\mathcal{A} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{B} \subseteq \mathbb{R}^{d_2}$. Then, $g = (\mathcal{V}, \mathcal{E}, \omega, \theta)$ determines an **attributed graph**.

Edge-heterogeneous graphs, sometimes called multi-relational graphs, are defined differently in the literature. Here, node- and edge-heterogeneous graphs, heterogeneous graphs for short, are considered and need a definition slightly deviating from the literature to enable a straightforward adaptation of the concepts in this work.

Definition 2.1.03 (Heterogeneous Graph) Let $g = (\mathcal{V}, \mathcal{E}, \omega, \theta)$ be an attributed graph as above. Further, let $\mathcal{S} \subset \mathbb{N}$ be a set of node species and $\mathcal{R} \subset \mathbb{N}$ the set of relation types. The node and edge attribute mappings are then extended to $\bar{\omega} : \mathcal{V} \to \mathcal{A} \times \mathcal{S}$ and $\bar{\theta} : \mathcal{E} \to \mathcal{B} \times \mathcal{R}$. Then, $g = (\mathcal{V}, \mathcal{E}, \bar{\omega}, \bar{\theta})$ is called a **heterogeneous graph**.

Furthermore, graphs can change in their structure as well as their attributes over time. Based on the static graph definition from Def. 2.1.02, a dynamic graph can be represented in two ways.

Definition 2.1.04 (Dynamic Graph) A **dynamic graph** G is determined as a sequence of temporal changes in a graph which can either be represented in **discrete** or **continuous** fashion:

1. A dynamic graph in **discrete-time representation** is given as a set $G = \{g_1, \dots, g_k\}$ of static attributed graph snapshots $g_t = (\mathcal{V}_t, \mathcal{E}_t, \omega_t, \theta_t)$ at time steps $t = 1, \dots, k$.
2. The **continuous-time representation** of a dynamic graph is defined as a set $G = \{g_{t_0}, \mathbb{E}\}$ with initial static graph $g_{t_0} := (\mathcal{V}_{t_0}, \mathcal{E}_{t_0}, \omega_{t_0}, \theta_{t_0})$ at time stamp $t_0 \in \mathcal{T}$ and a set $\mathbb{E} = \{e_t, t \in \mathcal{T}\}$ of events encoding a structural $e_t \in \{add, delete\}$ or attribute change $e_t = attr\_change$ at time stamp $t > t_0 \in \mathcal{T}$.

We build on the definitions of various Wasserstein distances taken from [28] to determine a similarity measure on attributed, heterogeneous, and dynamic graphs. For this purpose, it is necessary to build a bridge between the definition of graphs and the Wasserstein metric, which is defined for probability measures. For this purpose, consider graphs in the more general setting as a structured object over a metric space as defined in [28]:

Definition 2.1.05 (Structured Object) A structured object over a metric space (Ω, d_Ω) is a triplet $(X \times \Omega, d_X, \mu)$, where (X, d_X) is a metric space and μ is a probability measure over $X \times \Omega$. The attribute space is denoted as(Ω, d_Ω), such that $d_\Omega : \Omega \times \Omega \to \mathbb{R}_+$ is the distance in the attribute 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. The structure and attribute marginals of μ are denoted μ_X and μ_Ω, respectively.

From the perspective of graph theory, the structure space (X, d_X) involves a distance metric d_X representing, e.g., the neighborhood information in form of the shortest path. Here, the probability measure μ of a graph can then be defined using associated node probabilities h_i for all nodes $i \in \mathcal{V}$ with structural information $x_i \in X$ and attributes $a_i \in \Sigma$ by

$$\mu = \sum_{i \in \mathcal{V}} h_i \delta_{(x_i, a_i)}, \text{ with marginals } \mu_X = \sum_{i \in \mathcal{V}} h_i \delta_{x_i} \text{ and } \mu_\Omega = \sum_{i \in \mathcal{V}} h_i \delta_{a_i},$$

and the dirac measure δ evaluated on the set $\mathcal{V}$. Such node probabilities could be, e.g., determined by the node degrees or other structural information.

2.2 Wasserstein Distances

Having established the formulation of probability measures on graphs, we can proceed with the definitions of the different Wasserstein distances taken from [28].

Definition 2.2.01 (Wasserstein Distance) Let (X, d_X) be a Polish space, $\mu, \nu \in \mathcal{P}(X)$ be two probability measures and $p \in \mathbb{N}$. Then, the **Wasserstein distance** is defined as

$$d_{W,p}(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int\limits_{X \times X} d_X(a,b)^p \, \mathrm{d}\pi(a,b) \right)^{\frac{1}{p}} .$$

Here, $\Pi : X \times X \to X$ is the set of all couplings between μ and ν, i.e., all joint distributions over $X \times X$ whose marginals are μ and ν. For detailed derivations and helpful illustrations we refer to [28]. For probability measures from different spaces, the Gromov-Wasserstein distance has been established. It additionally takes the structure of the input spaces into account.

Definition 2.2.02 (Gromov-Wasserstein distance) Let $(X, d_X), (Y, d_Y)$ be two Polish spaces, $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ two probability measures and $p \in \mathbb{N}$. Then, the **Gromov-Wasserstein distance** is given by

$$d_{GW,p}(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int\limits_{(X \times Y)^2} L(x, y, x', y')^p \, \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y') \right)^{\frac{1}{p}}$$

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

The Gromov-Wasserstein distance is defined as the Wasserstein distance over all pairwise distances computed on the two Polish spaces separately to integrate structural information.

Combining both the Wasserstein and Gromov-Wasserstein distance, the Fused Gromov-Wasserstein distance (FGWD) determines a similarity measure on objects from different structured spaces with a shared attribute space as introduced in Def. 2.1.05. The FGWD is a convex combination of the Wasserstein distance in the attribute space and the Gromov-Wasserstein distance between the two structured spaces.

Definition 2.2.03 (Fused Gromov-Wasserstein distance) Let $\alpha \in [0, 1]$ and $p \in \mathbb{N}$. For two structured objects $(X \times \Omega, d_X, \mu)$ and $(Y \times \Omega, d_Y, \nu)$ with the shared attribute space (Ω, d_Ω), the **Fused Gromov-Wasserstein distance** is defined as

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

with

$$E_{p,q,\alpha}(\pi) = \int\limits_{(X\times\Omega\times Y\times\Omega)^2} D(\alpha)^p \, \mathrm{d}\pi\big((x,a),(y,b)\big)\, \mathrm{d}\pi\big((x',\cdot),(y',\cdot)\big)$$

and

$$\begin{aligned} D(\alpha) &= (1-\alpha)d_\Omega(a,b)^q + \alpha L(x,y,x',y')^q, \\ L(x,x',y,y') &= |d_X(x,x') - d_Y(y,y')|. \end{aligned} \tag{1}$$

Here, $d_\Omega(a,b)$ functions as feature and $L(x,y,x',y')$ as the structure cost.

3 Attributed and Heterogeneous Fused Gromov-Wasserstein Distance

The FGWD is constrained to two structured objects with the same attribute space and, thus, is severely restricted in its applicability to more complex objects. In this work, we first extend the FGWD to graphs with potentially heterogeneous attributes and structure. Afterward, we extend the distance to dynamic graphs in the next section.

Recapitulating Def. 2.1.03, heterogeneous graphs can be formulated as attributed graphs whose attribute spaces are extended by the node species and relation types. Consequently, we have to extend the FGWD to structured objects with potentially different attribute spaces to obtain a similarity measure based on the Wasserstein distance for heterogeneous graphs. For this purpose, we adapt the ideas of the Gromov-Wasserstein distance and consider both intra-space distances of the structure and attribute spaces, as illustrated in Fig. 1. In this way, the presented Attributed Fused Gromov-Wasserstein distance reflects both the characteristics of the structure and the attribute spaces.

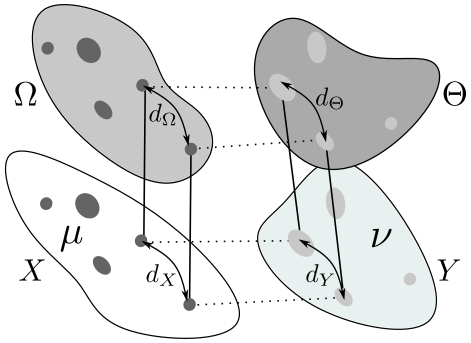

Fig. 1. The Attributed Fused Gromov-Wasserstein Distance for structured objects on different attribute spaces Ω, Θ with structure spaces X, Y.

Definition 3.0.01 (Attributed Fused Gromov-Wasserstein Distance) Let $(X \times \Omega, d_X, \mu)$ and $(Y \times \Theta, d_Y, \nu)$ be two structured objects with metric attribute spaces (Ω, d_Ω), (Θ, d_Θ), $\alpha \in [0,1]$ and p, $q \in \mathbb{N}$. Then, the **Attributed Fused Gromov-Wasserstein Distance (AFGWD)** is determined by

$$d_{AFGW,p,q,\alpha}(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} E_{p,q,\alpha}(\pi) \right)^{\frac{1}{p}}, \tag{2}$$

with

$$E_{p,q,\alpha}(\pi) = \int\limits_{(X\times\Omega\times Y\times\Theta)^2} D_{q,\alpha}(\pi)^p \,\mathrm{d}\pi((x,a),(y,b))\,\mathrm{d}\pi((x',a'),(y',b')), \tag{3}$$

and the convex combination of the attribute and structure costs

$$D_{q,\alpha}(\pi) = (1-\alpha)K(a,a',b,b')^q + \alpha L(x,y,x',y')^q$$
$$\text{with } K(a,a',b,b') = |d_\Omega(a,a') - d_\Theta(b,b')|,$$

and L as in Eq. (1).

The AFGWD considers the data structure in both attribute spaces by considering all pairwise distances of the node attributes. The attributes may be vectors, text, images, or others, depending on the application. The distance measures on these attribute spaces then have to fulfill the conditions of a metric to ensure that the AFGWD remains a metric.

Remark 3.0.02 (AFGWD for Heterogeneous Graphs) Heterogeneous graphs are defined in Def. 2.1.03 as attributed graphs whose node and edge attribute spaces are extended by the set of node and edge types. As a result, the AFGWD can be applied to heterogeneous graphs accordingly. For node types (or node species), we add an attribute to the attribute space defining the node type of each node. To integrate edge types (or relation types), we extend the attribute space by the space of all edge attributes on the structured object.

4 Temporal Fused-Gromov Wasserstein Distance

Besides additional node attributes or heterogeneity in many applications, graphs are often inherently dynamic. The challenges in processing dynamic graphs are the changing attributes and evolving node and edge sets. To the best of our knowledge, there are no approaches yet to compare such dynamic graphs with the Wasserstein distance. We propose a FGWD for dynamic graphs of the two most common representations to close this gap. On the one hand, many approaches for learning on dynamic graphs utilize the graph representation in the form of graph snapshot sequences (discrete-time representation). On the other hand, in some cases, the compact continuous-time representation of dynamic graphs is also used, which involves the start graph and a stream of graph events. Therefore, we define a distance measure for each representation that acts as a similarity measure between dynamic graphs as follows.

4.1 Discrete-Time Dynamic Graphs

In the case of dynamic graphs in discrete-time representation, we assume the input graphs to have the same time steps $\mathcal{T}$. Then, we can utilize the AFGWD from Def. 2 for each time step and define the Temporal Wasserstein distance for discrete-time graphs as the mean of the AFGWD per time step.

Definition 4.1.01 (Temporal Fused Gromov-Wasserstein Distance) Let $\{(X_t \times \Omega_t, d_{X_t}, d_{\Omega_t}, \mu_t)\}_{t\in\mathcal{T}}$ and $\{(Y_t \times \Theta_t, d_{Y_t}, d_{\Theta_t}, \nu_t))\}_{t\in\mathcal{T}}$ be two sequences of spaces with (possibly) changing structure (X_t, d_{X_t}), (Y_t, d_{Y_t}) and attribute spaces (Ω_t, d_{Ω_t}), (Θ_t, d_{Θ_t}). Let further μ_t, ν_t over the set of time steps $\mathcal{T} = \{1, \dots, T\}$ be the corresponding probability measures, and p, $q \in \mathbb{N}$. Then, the **Temporal Fused-Gromov Wasserstein Distance (TFGWD)** of the two probability measures $\mu_\mathcal{T} := \mu_1 \times \ldots \times \mu_T$ and $\nu_\mathcal{T} := \nu_1 \times \ldots \times \nu_T$ is defined as

$$d_{\mathrm{TFGW},\alpha,p,q}(\mu_\mathcal{T}, \nu_\mathcal{T}) := \frac{1}{T}\sum_{t\in\mathcal{T}} d_{AFGW\alpha,p,q}(\mu_t, \nu_t),$$

with

$$d_{AFGW\alpha,p,q}(\mu_t, \nu_t) = \inf_{\pi_t \in \Pi(\mu_t,\nu_t)} E_{p,q,\alpha}(\pi_t)$$

and the term $E_{p,q,\alpha}(\pi_t)$ of Eq. (3) from the Attributed Fused Gromov-Wasserstein distance.

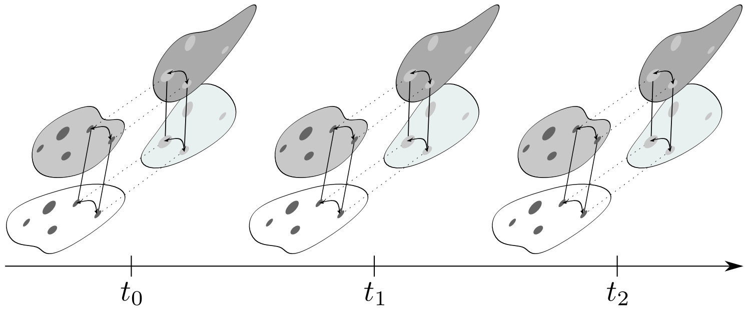

Fig. 2. For each time step, the AFGWD is calculated for two structure and attribute spaces each. The TFGWD then corresponds to the mean distance over all time steps.

The measure retains the properties of a metric, as the triangle inequality holds for each metric in the sum separately. In Fig. 2, the distance of two sequences of different structure and attribute spaces is illustrated.

4.2 For Continuous-Time Dynamic Graphs

The definition for discrete dynamic graphs in 4.1.01 is restricted to graph sequences and is thus tied to the potentially cost-intensive processing of dynamic graphs. However, to have a Wasserstein distance for the more compact representation of dynamic graphs in continuous time, we define an FGWD in the following, which considers the start graphs and the event sequences individually. For this purpose, we assume that no simultaneous events are happening on the graph, and the event sequences are represented as Temporal Point Processes (TPP).

Definition 4.2.01 (Continuous-Time Fused Gromov-Wasserstein Distance) Let $(X \times \Omega, d_X, d_\Omega, \mu)$ and $(Y \times \Theta, d_Y, d_\Theta, \nu)$ be two structured spaces and $p, q \in \mathbb{N}$. Further, let $\phi = \{\phi_i = (t_i, e_i)\}_{i \leq n}$ and $\psi = \{\psi_j = (\tau_j, \epsilon_j)\}_{j \leq m}$ be two Temporal Point Processes with time stamps t_i, $\tau_j \in [0, T)$ and structural or attribute change events e_i, ϵ_j of potentially different lengths $n \leq m$. Then, the **Continuous-Time Fused Gromov-Wasserstein Distance (CTFGWD)** is a convex combination of the AFGWD of the start graphs and the Wasserstein distance of the event sequences:

$$d_{CTFGW\alpha,\gamma}(\mu, \nu, \phi, \psi) := (1-\gamma)\underbrace{d_{AFGW,p,q,\alpha}(\mu, \nu)}_{\text{start graphs}} + \gamma \underbrace{d_{W,p}(\phi, \psi)}_{\text{event sequences}}.$$

As distance function of the event time sequences in the Wasserstein distance d_W, we utilize the distance measure of time series $\|\cdot\|_*$ from [34] which is defined as:

$$\|\phi - \psi\|_* = \sum_{i=0}^{n} (|t_i - \tau_i| + (1 - \delta_{e_i, \epsilon_i})) + (m-n) \cdot T - \sum_{j=n+1}^{m} \tau_j.$$

The time series distance reflects the pairwise absolute distances of time stamps and incorporates period length differences. The convex combination of the start graph and the event sequence distances then determines the similarity of two dynamic graphs in continuous-time representation as illustrated in Fig. 3.

5 Computation

Computing the different Wasserstein distances is inherently cost-intensive because determining the optimal coupling between the two inputs is hard. Therefore, in this section, we adapt existing approximation algorithms from [28] to calculate the proposed Wasserstein distances.

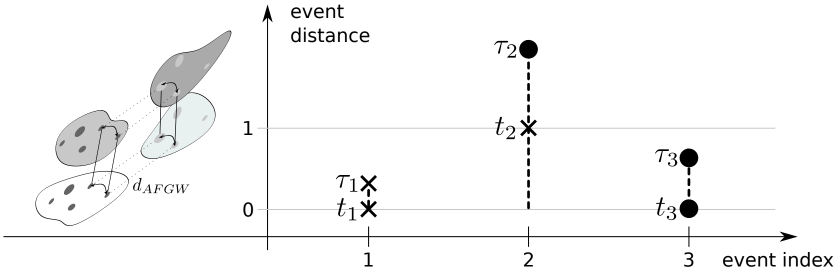

Fig. 3. The Wasserstein distance for two dynamic graphs in continuous-time is determined by the AFGWD d_{AFGW} between the two start graphs and the Wasserstein distance d_W of the Temporal Point Processes that characterize the event sequences. The event sequence heights represent time differences, and event differences are represented by different symbols (e.g. cross, bold point). The distance increases by 1 for each pair of different events.

5.1 FGWD for Graphs with Different Attribute Spaces

Let $\mu = \sum_{i \in \mathcal{V}} h_i \delta_{(x_i,a_i)}$ and $\nu = \sum_{j \in \mathcal{V}'} h'_j \delta_{(y_i,b_i)}$ be two attributed graphs. Analogous to [28], we set $p = 1$ and $q = 2$ so that we can rewrite the d_{AFGW} as quadratic optimization problem. For this purpose, let

$$\begin{aligned} M_\Omega &= (d_\Omega(a_i, a_k))_{a_i,a_k \in \Omega, i,k \in \mathcal{V}}\,, & M_\Theta &= (d_\Theta(b_j, b_l))_{b_j,b_l \in \Theta, j,l \in \mathcal{V}'} \\ C_X b &= (d_X(x_i, x_k))_{x_i,x_k \in X, i,k \in \mathcal{V}}\,, & C_Y &= (d_Y(y_j, y_l))_{y_j,y_l \in Y, j,l \in \mathcal{V}'} \end{aligned}$$

be the attribute distance matrices and intra-graph structure similarities, respectively. Then, d_{AFGW} can be rewritten as

$$d_{AFGW\alpha}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} E_\alpha(\pi)$$

with

$$E_\alpha(\pi) = \sum_{\substack{i,k \in \mathcal{V} \\ j,l \in \mathcal{V}'}} \Big[(1-\alpha)\,|M_\Omega(i,k) - M_\Theta(j,l)|^2 + \alpha\,|C_X(i,k) - C_Y(j,l)|^2\Big]\pi_{i,j}\pi_{k,l}.$$

Then, the related quadratic optimization problem is given by

$$\pi^* = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min}\ \mathrm{vec}(\pi)^\top N(\alpha)\mathrm{vec}(\pi) + \mathrm{vec}(\pi)^\top B(\alpha)\mathrm{vec}(\pi), \tag{4}$$

where $N(\alpha) = -2\alpha M_\Theta \otimes_K M_\Omega$, $B(\alpha) = (2\alpha - 2)C_2 \otimes_K C_1$, vec() denotes the column-wise vector-stacking operator and $\otimes_K$ is the Kronecker product of matrices. It can then be solved with an extension of the algorithms provided in [28] using the partial derivative of Eq. (4):

$$\nabla E_\alpha(\pi) = \tilde{C} - 4C_X \Pi C_Y + \tilde{M} - 4M_\Omega \Pi M_\Theta$$

with

$$\begin{aligned}
\tilde{C} &= 2\Big[(C_X \circ C_X)\mathrm{vec}(\pi_X)\mathbf{1}_Y^\top + \Big((C_Y \circ C_Y)\mathrm{vec}(\pi_Y)\mathbf{1}_X^\top\Big)^\top\Big], \\
\tilde{M} &= 2\Big[(M_\Omega \circ M_\Omega)\mathrm{vec}(\pi_X)\mathbf{1}_Y^\top + \Big((M_\Theta \circ M_\Theta)\mathrm{vec}(\pi_Y)^\top\mathbf{1}_X^\top\Big)^\top\Big], \\
\mathrm{vec}(\pi_X)_i &= \sum_{j\in Y}\pi_{i,j} \text{ and } \mathrm{vec}(\pi_Y)_j = \sum_{i\in X}\pi_{i,j}, \\
\Pi &= (\pi_{i,j})_{i\in X,\, j\in Y}.
\end{aligned}$$

and the Hadamard product $\circ$ of matrices. Then, the optimal coupling can be calculated using the conditional gradient algorithm along with the line search given in [28].

5.2 Dynamic FGWD

The TFGWD for discrete dynamic graphs can be reformulated analogously to apply the conditional gradient and line search described in [28] for each time step accordingly. For this purpose, the TFGWD is rewritten for two temporal structured spaces as in Def. 4.1.01 as

$$d_{TFGW\alpha}(\mu_\mathcal{T}, \nu_\mathcal{T}) = \frac{1}{T}\sum_{t\in\mathcal{T}} \inf_{\pi_t\in\Pi(\mu_t,\nu_t)} E_\alpha(\pi_t),$$

with

$$E_\alpha(\pi_t) = \sum_{\substack{i,k\in\mathcal{V}_t\\ j,l\in\mathcal{V}'_t}} \Big[(1-\alpha)\,|M_{\Omega_t}(i,k) - M_{\Theta_t}(j,l)|^2 + \alpha\,|C_{X_t}(i,k) - C_{Y_t}(j,l)|^2\Big]\pi_{t,i,j}\pi_{t,k,l}.$$

Using this adapted representation, we can now calculate the TFGWD stepwise. However, assuming that the graph changes smoothly, we use the optimal coupling π^*_{t-1} found in the previous time step $t-1$ as initialization for calculating the coupling at time t in the conditional gradient approach from [28].

For continuous-time dynamic graphs, the CTFGWD consists of the AFGWD between the start graphs and the adapted Wasserstein distance for the event sequences. As a result, the CTFGWD can be calculated by executing the algorithm from Sec. 5.1 on the start graphs together with minimizing the Wasserstein distance on the event sequences ϕ, ψ.

6 Discussion and Future Work

Time Shift Adaptation. In the definition of the discrete temporal FGWD, we assume that both sequences of attributed graphs have the same length of time steps. There are already developed Wasserstein distances to compare time series of different lengths, as for event sequences that adapt the ideas of Dynamic Time Warping as in [10] and [23]. A temporal shift could be considered for dynamic graphs by evaluating the pairwise Wasserstein distances accordingly.

Computation and Runtime. Our current work does not yet include the implementation of distances, which will be realized in the future to show their applicability. We proposed in Sec. 5 to use Sinkhorn distances to obtain a quadratic optimization problem and apply a conditional gradient descent algorithm. Other Authors have proposed greedy algorithms [9] to approximate the Wasserstein distance. Further research is needed to determine trade-offs between the runtime and the accuracy of the algorithms. In addition, it would be reasonable to analyze the computational time dependent on the size of the input graphs.

Graph Generation. Distances between probability measures are widely used for (homogenous) graph generation purposes, such as in generative adversarial networks (GANs) [8], variational autoencoders (VAEs) [12], and diffusion models [6, 31]. Although there exist approaches generating heterogeneous [14, 37] and dynamic graphs [11, 36, 17], these techniques have not been fully exploited for these domains. Future work could investigate the application of the proposed Wasserstein distances in the context of GANs, VAEs, and diffusion models and their impact on training stability and performance. Further, generating heterogeneous and dynamic graphs utilizing the corresponding Wasserstein distances appears promising due to the explicit integration of structure, complex attributes, and temporal evolution.

Molecular Data. Learning on molecular data is a research area in which graphs are commonly used to represent data. Most approaches use vectors or homogeneous graphs as molecule representations. However, there are also promising approaches using heterogeneous graphs [21, 35], techniques for generating heterogeneous graphs [14], that also utilize VAEs [37]. Nonetheless, the research field has yet to exploit all the state-of-the-art techniques. Consequently, using dynamic graphs and the proposed Wasserstein distances could provide new advances in future work.

Graph Neural Network Architecture. The Graph Neural Network model proposed in [33], e.g., utilizes the Fused Gromov-Wasserstein distance to incorporate distances from the input graph to pre-defined template graphs. This approach serves as a global pooling procedure and could be extended to heterogeneous and dynamic graphs accordingly.

7 Conclusion

In this work in progress, we addressed the limitations of applying Wasserstein distances to graphs, particularly the constraints on static graphs with homogeneous attributes. We introduced two novel extensions: the Attributed Fused Gromov-Wasserstein Distance (AFGWD) for heterogeneous graphs with different attribute spaces and the Temporal Fused Gromov-Wasserstein Distance (TFGWD) for dynamic graphs. These extensions incorporate structural information and diverse attribute spaces, providing a more versatile distance metric suitable for a broader range of graph-based applications. Furthermore, we outlined efficient computation strategies using existing approximation methods to mitigate the high computational costs associated with Wasserstein distances. These advancements not only enhance the applicability of Wasserstein distances in complex graph data scenarios but also open new avenues for research in heterogeneous and dynamic graph analysis. Future work will focus on refining these methods and exploring their practical implications in various real-world applications.

Contribution

Silvia Beddar-Wiesing Conceptualization (TFGWD, CTFGWD), Formal Analysis, Writing - Original draft preparation, Writing - Reviewing and Editing.
Dominik Köhler Conceptualization (AFGWD), Formal Analysis, Writing - Reviewing and Editing.

Acknowledgement

We would like to thank Nicolas Posocco for the inspiration for this work and the stimulating discussions.

Bibliography

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. page 214–223, 2017.

[2] L. Brogat-Motte, R. Flamary, C. Brouard, J. Rousu, and F. d'Alché-Buc. Learning to predict graphs with fused gromov-wasserstein barycenters. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pages 2321–2335. PMLR, 2022.

[3] C. Cheng, B. Zhou, G. Ma, D. Wu, and Y. Yuan. Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. Neurocomputing, 409:35–45, 2020.

[4] K. Cheng, S. Aeron, M. C. Hughes, and E. L. Miller. Dynamical wasserstein barycenters for time-series modeling. Advances in Neural Information Processing Systems, 34:27991–28003, 2021.

[5] L. Condeixa, F. Oliveira, and A. S. Siddiqui. Wasserstein-distance-based temporal clustering for capacity-expansion planning in power systems. In 2020 International Conference on Smart Energy Systems and Technologies (SEST), pages 1–6, 2020.

[6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[7] Y. Dai, W. Guo, and C. Eickhoff. Wasserstein adversarial learning based temporal knowledge graph embedding. arXiv preprint arXiv:2205.01873, 2022.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. CoRR, abs/1406.2661, 2014.

[9] G. Houry, H. Bao, H. Zhao, and M. Yamada. Fast 1-wasserstein distance approximations using greedy strategies. In International Conference on Artificial Intelligence and Statistics, pages 325–333. PMLR, 2024.

[10] H. Janati, M. Cuturi, and A. Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In S. Chiappa and R. Calandra, editors, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 1695–1704. PMLR, 26–28 Aug 2020.

[11] S. M. Kazemi, R. Goel, K. Jain, I. Kobyzev, A. Sethi, P. Forsyth, and P. Poupart. Representation learning for dynamic graphs: A survey. J. Mach. Learn. Res., 21:70:1–70:73, 2020.

[12] T. N. Kipf and M. Welling. Variational graph auto-encoders. CoRR, abs/1611.07308, 2016.

[13] S. Kolouri, N. Naderializadeh, G. K. Rohde, and H. Hoffmann. Wasserstein embedding for graph learning. arXiv preprint arXiv:2006.09430, 2020.

[14] C. Ling, C. Yang, and L. Zhao. Deep generation of heterogeneous networks. In ICDM, pages 379–388. IEEE, 2021.

[15] Y. Liu, Z. Qin, T. Wan, and Z. Luo. Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. Neurocomputing, 311:78–87, 2018.

[16] T.-j. Luo, Y. Fan, L. Chen, G. Guo, and C. Zhou. Eeg signal reconstruction using a generative adversarial network with wasserstein distance and temporal-spatial-frequency loss. Frontiers in Neuroinformatics, 14, 2020.

[17] S. Mahdavi, S. Khoshraftar, and A. An. Dynamic joint variational graph autoencoders. In PKDD/ECML Workshops (1), volume 1167 of Communications in Computer and Information Science, pages 385–401. Springer, 2019.

[18] F. Mémoli. Gromov-wasserstein distances and the metric approach to object matching. Found. Comput. Math., 11(4):417–487, 2011.

[19] F. Mémoli. Spectral gromov-wasserstein distances for shape matching. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 256–263, 2009.

[20] O. Ryu and B. Lee. Style transfer using optimal transport via wasserstein distance. In 2022 IEEE International Conference on Image Processing (ICIP), pages 2681–2685. IEEE, 2022.

[21] Z. Shui and G. Karypis. Heterogeneous molecular graph neural networks for predicting molecule properties. In ICDM, pages 492–500. IEEE, 2020.

[22] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. ACM Transactions on Graphics (ToG), 34(4):1–11, 2015.

[23] B. Su and G. Hua. Order-preserving wasserstein distance for sequence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[24] M. Tang, C. Yang, and P. Li. Graph auto-encoder via neighborhood wasserstein reconstruction. arXiv preprint arXiv:2202.09025, 2022.

[25] Thomas, Josephine and Moallemy-Oureh, Alice and Beddar-Wiesing, Silvia and Holzhüter, Clara. Graph Neural Networks Designed for Different Graph Types: A Survey. Transactions on Machine Learning Research, 2022.

[26] V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel. Sliced gromov-wasserstein. Advances in Neural Information Processing Systems, 32, 2019.

[27] I. O. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein autoencoders. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018.

[28] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Fused gromov-wasserstein distance for structured objects. Algorithms, 13(9):212, 2020.

[29] T. Vayer, N. Courty, R. Tavenard, L. Chapel, and R. Flamary. Optimal transport for structured data with application on graphs. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International

Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 6275–6284. PMLR, 2019.

[30] T. Viehmann. Partial wasserstein and maximum mean discrepancy distances for bridging the gap between outlier detection and drift detection. arXiv preprint arXiv:2106.12893, 2021.

[31] C. Vignac, I. Krawczuk, A. Siraudin, B. Wang, V. Cevher, and P. Frossard. Digress: Discrete denoising diffusion for graph generation. In ICLR. OpenReview.net, 2023.

[32] C. Villani et al. Optimal transport: old and new, volume 338. Springer, 2009.

[33] C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer, and N. Courty. Template based graph neural network with optimal transport distances. In NeurIPS, 2022.

[34] S. Xiao, M. Farajtabar, X. Ye, J. Yan, L. Song, and H. Zha. Wasserstein Learning of Deep Generative Point Process Models. Advances in neural information processing systems, 30, 2017.

[35] Z. Yu and H. Gao. Molecular representation learning via heterogeneous motif graph neural networks. In ICML, volume 162 of Proceedings of Machine Learning Research, pages 25581–25594. PMLR, 2022.

[36] L. Zhang, L. Zhao, S. Qin, D. Pfoser, and C. Ling. Tg-gan: Continuous-time temporal graph deep generative models with time-validity constraints. In Proceedings of the Web Conference 2021, pages 2104–2116, 2021.

[37] Y. Zhao, J. Yu, Y. Cheng, C. Yu, Y. Liu, X. Li, and S. Wang. Self-supervised heterogeneous graph variational autoencoders. CoRR, abs/2311.07929, 2023.