ERGOT: ENTROPY-REGULARIZED GRAPH OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph comparison is a fundamental task in many scientific fields, which not only relates to graph matching, an NP-hard problem, but also has multitudinous applications in graph learning. We tackle this task by studying optimal graph representation and an entropy-regularized optimal transport problem between graphs (ErGOT). First, we analytically derive a family of Gaussian variables that optimally represent graph topology and node relation. Second, we realize graph comparison by encompassing ErGOT, a problem with low sample complexity, between represented topology information. Third, we control biases in the solution of ErGOT by a 2-Sinkhorn divergence, whose closed-form expression can be derived on the manifold of Gaussian variables. As the Gaussian geometry changes with entropy regularization magnitude, the ErGOT defined with 2-Sinkhorn divergence wanders between pure graph optimal transport and maximum mean graph discrepancy. We demonstrate that these statistically efficient, principally unbiased, and in-between properties ensure a fast convergence of our approach to higher performance than the state-of-art algorithms in graph alignment, classification, and sketching tasks.

1 INTRODUCTION

General backgrounds of graph comparison. Graph is a basic type of data structure with extensive applications in engineering Deo (2017), physics Newman (2003), chemistry Trinajstic (2018), and biology Mheich et al. (2020). Mainstream graph analyses (e.g., alignment) frequently begin with graph comparison, which deals with similarities and differences between graphs. However, graph comparison itself is a daunting challenge. On the one hand, researchers frequently lack *a priori* knowledge about node alignment relations, making graph matching an NP-hard problem Conte et al. (2004). On the other hand, a meaningful metric of graph similarities and differences remains elusive even when nodes are perfectly aligned Petric Maretic et al. (2019); Maretic et al. (2022).

Graph optimal transport for graph comparison. Recently, optimal transport theory Villani (2009); Peyré et al. (2019) has been introduced to realize graph comparison Garg & Jaakkola (2019); Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022); Dong & Sawin (2020). Among these existing approaches, the newly proposed GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022) benefit from the probabilistic representation of graph via graph signals distributed on nodes Ortega et al. (2018); Dong et al. (2016). The graph representation is discovered to simultaneously ensure an appropriate description of graph properties (e.g., topology, heterogeneity, and dynamics) and an analytic expression of the objective of optimal transport (e.g., the 2-Wasserstein distance Villani (2009); Peyré et al. (2019)). These properties have been demonstrated as computationally favorable Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022). Our research primarily focuses on this promising direction.

Remaining challenges in graph optimal transport. However, there remain numerous challenges in this emerging direction, among which, two critical problems are listed below:

(I) **General principles to choose graph representation remain elusive yet**. In existing graph-signal-based optimal transport frameworks Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022), a graph $\mathcal{G}(V, E)$ is represented as a Gaussian vari-

able $\mathcal{X} \sim \mathcal{N}\left(\mathbf{0}, \mathcal{Q}^2\left(L\right)\right)$, where $\mathcal{Q}\left(\cdot\right)$ denotes a function of graph Laplacian L Biyikoglu et al. (2007). In GOT Petric Maretic et al. (2019); Petric Maretic (2021), researchers follow the idea derived by factor analysis and low-rank models Dong et al. (2016); Kalofolias (2016) to define $\mathcal{Q}\left(\cdot\right)$. In fGOT Maretic et al. (2022), more definitions of $\mathcal{Q}\left(\cdot\right)$ are proposed from the engineering practice perspective. However, although different types of $\mathcal{X}$ have distinct effects in graph comparison Maretic et al. (2022), researchers lack general principles in defining $\mathcal{X}$ to capture target graph properties Ortega et al. (2018).

(II) **Graph-signal-based optimal transport is not computationally ideal yet**. Pure optimal transport between graphs requires solving a linear problem, which entails a critical burden in computation Genevay et al. (2017); Mena et al. (2017); Feydy et al. (2019). Although *Sinkhorn operator* has been applied to define an approximation of pure optimal transport that allows automatic differentiation Genevay et al. (2017) and supports graph optimal transport solutions Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022), there remains a lot of room for further improvement Cuturi (2013) because the current graph optimal transport still deals with a non-convex problem with high sample complexity (see Mallasto et al. (2021); Genevay et al. (2019) for explanations).

Our framework and contributions. In this paper, we attempt to resolve challenges (I-II) by proposing a new framework. Our first contribution is to analytically derive a family of Gaussian variables that optimally represent graph topology and node relation. Our second contribution is to generalize the graph-signal-based optimal transport proposed in Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022) to an entropy-regularized optimal transport problem on the Gaussian geometry (ErGOT), which is efficient in sampling. Beyond the original optimization target of entropy-regularized optimal transport, we further derive a closed-form expression of 2-Sinkhorn divergence to define an optimization target of ErGOT with low biases. Driven by 2-Sinkhorn divergence, our ErGOT framework can wander between pure optimal transport and maximum mean discrepancy between graphs according to entropy regularization magnitude.

2 Related works

Graph matching. Graph matching, such as exact De Santo et al. (2003) and inexact Gao et al. (2010) matching with and without edge-preserving properties, is a kind of NP-hard quadratic programming problem whose solution is constrained as a permutation matrix Conte et al. (2004); Cour et al. (2006); Jiang et al. (2017). Consequently, numerous relaxation approaches (e.g., continuous domain Yu et al. (2018), spectral clustering Caelli & Kosinov (2004), and semi-definite programming Schellewald & Schnörr (2005) relaxation) have been proposed to approximate the original problem. The metric for graph comparison in those works frequently lacks an analytic expression.

Graph kernel. Graph kernel approaches decompose a graph into multiple atomic substructures (e.g., graphlets Shervashidze et al. (2009), random walks Kashima et al. (2004), shortest paths Borgwardt & Kriegel (2005), and cycles Horváth et al. (2004)) to define the kernel value among these substructures (i.e., counting the number of shared substructures) Kriege et al. (2020); Cai et al. (2018) as a metric of graph comparison. The validity of graph comparison is determined by the capacity of these handcrafted substructures (i.e., extracted by certain manually defined functions Narayanan et al. (2017)) to reflect graph properties, which may be limited by the high-dimensional, sparse, and non-smooth graph representation in kernel spaces Yanardag & Vishwanathan (2015).

Graph optimal transport. Graph optimal transport is a natural idea to define the metric of graph comparison. Early back to Gu et al. (2015), the p-Wasserstein distance has been calculated on normalized graph Laplacian spectra. Meanwhile, a regularized Gromov-Wasserstein distance has been defined for optimal transport between structure data (e.g., irregular polygons) Peyré et al. (2016). More recently, optimal transport problem has been analyzed based on minimum cost flow on graphs Garg & Jaakkola (2019), a combination of structure and feature information of graphs Titouan et al. (2019), and a pair of simultaneous transport processes between nodes and Laplacian spectra Dong & Sawin (2020). Notably, the graph-signal-based optimal transport frameworks (e.g., GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022)) have been developed to consider optimal transport between random signals distributed on graphs. With an ideal

definition of these signals, fundamental properties of graphs can be captured in the optimal transport problem Maretic et al. (2022). However, such a definition remains elusive yet (see challenge (I)).

3 FRAMEWORK OF ERGOT

Probabilistic representation of graphs. Why should graphs be represented by certain Gaussian variables controlled by graph Laplacian as Dong et al. (2016); Kalofolias (2016); Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022) propose? How to enable the probabilistic representation to capture target graph properties? In this section, we explore a unified answer.

Let us consider a mapping $\phi : V \to \Omega$ from the node set V of a graph $\mathcal{G}(V, E)$ to a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with $\Omega = \mathbb{R}$, which defines a random variable $\mathcal{X}_\phi = (X_\phi(1), \ldots, X_\phi(n))$, where $X_\phi(i) = \phi(v_i)$ and $n = |V|$. An ideal mapping ϕ should reflect graph topology and node relation (e.g., homogeneity or heterogeneity) by some of its properties. In our research, we explore a way to represent target graph properties via the smoothness of ϕ and the distribution of X_ϕ.

In graph signal theories, the smoothness index of ϕ on $\mathcal{G}$ is determined by the topology information contained in graph Laplacian L Chung & Graham (1997); Shuman et al. (2013).

$$\mathcal{S}(\phi) = {\mathcal{X}_\phi}^T L \mathcal{X}_\phi. \tag{1}$$

A smaller $\mathcal{S}(\phi)$ corresponds to a higher smoothness of mapping ϕ. Please see Tian et al. (2022) for further explanations. Because $\mathcal{X}_\phi$ is a random variable, we primarily analyze $\mathbb{E}(\mathcal{S}(\phi))$, the expected smoothness index of ϕ. Such an expectation is derived in a quadratic form

$$\mathbb{E}(\mathcal{S}(\phi)) = \mathbb{E}(\mathcal{X}_\phi)^T L \mathbb{E}(\mathcal{X}_\phi) + \mathrm{tr}(L\Sigma(\mathcal{X}_\phi)), \tag{2}$$

where $\Sigma(\mathcal{X}_\phi) \in \mathbb{R}^{n\times n}$ denotes the covariance matrix of $\mathcal{X}_\phi$ and $\mathrm{tr}(\cdot)$ measures the trace. To avoid that $\mathcal{S}(\phi)$ diverges, we primarily analyze the case where the expectation $\mathbb{E}(\mathcal{X}_\phi)$ and the covariance $\Sigma(\mathcal{X}_\phi)$ of $\mathcal{X}_\phi$ are finite. For convenience, we further assume that $\mathbb{E}(\mathcal{X}_\phi) = \mathbf{0}$.

In a graph where node homogeneity and global structure are important, nodes are similar if they are connected by edges with higher weights (weights denote similarity). The smoothness of ϕ should be sufficiently high or invariant to express node homogeneity. In a graph where node heterogeneity and local structure matter, nodes are distinct if they are connected by edges with higher weights (weights denote difference). The smoothness of ϕ should be determined by graph topology and edge weights.

In **Table 1**, we suggest two prototypes of $\Sigma(\mathcal{X}_\phi)$. In general, $\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$ implies an expected smoothness index completely determined by graph topology, which is more applicable to node heterogeneity and local structure description. The nodes connected by an edge with a larger weight will behave inversely (with strongly negative covariance). In an opposite case, $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$ creates an expected smoothness index that is independent of graph topology and fully determined by graph size n. The nodes connected by an edge with a larger weight are more similar (with strongly positive partial correlation). In **Appendix A**, we present detailed proofs of all the results in **Table 1**.

Then, we analyze the distribution of variable $\mathcal{X}_\phi$. At the first glance, this seems to be an impossible question because $\mathcal{X}_\phi$ can have an arbitrary distribution as long as it keeps a zero mean and a covariance matrix in **Table 1**. However, no matter what kind of distribution $\mathcal{X}_\phi$ follows, what we face in real applications are its sampled observations. Therefore, we can study the distribution of an averaged observation $\langle\mathcal{X}_\phi\rangle = \frac{1}{r}\sum_{i=1}^{r}\mathcal{X}_\phi^i$, where each sample $\mathcal{X}_\phi^i$ is independently and identically distributed. Applying the multidimensional central limit theorem Van der Vaart (2000), we can readily derive $\sqrt{r}\langle\mathcal{X}_\phi\rangle \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma(\mathcal{X}_\phi))$ as $r \to \infty$. Consequently, one can directly define

$$\mathcal{X}_\phi \sim \mathcal{N}(\mathbf{0}, \Sigma(\mathcal{X}_\phi)) \tag{3}$$

Covariance matrix	Expected smoothness index	Expressed graph properties
$\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$	$\mathbb{E}(\mathcal{S}(\phi)) = \mathrm{tr}(L^2)$	Node heterogeneity and local structure
$\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$	$\mathbb{E}(\mathcal{S}(\phi)) = n - 1$	Node homogeneity and global structure

Table 1: Two prototypes of covariance matrix. Notion J denotes an all-one matrix.

to represent the normal case in application. Because $\mathbb{E}(\mathcal{X}_\phi)$ and $\Sigma(\mathcal{X}_\phi)$ are finite, the Gaussian variable in Eq. (3) is also the random variable in $\mathbb{R}$ that has a maximum entropy Cover (1999). Such a property is favorable in representing information (e.g., graph properties).

In sum, we can represent graph $\mathcal{G}$ as a Gaussian variable defined by a mapping ϕ whose smoothness is determined by graph topology and node relation (e.g., homogeneity and heterogeneity). The above derivations offer theoretical explanations of existing engineering experience Dong et al. (2016); Kalofolias (2016); Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022).

Entropy-regularized optimal transport problem on Gaussian geometry. Given $\mathcal{X}_\phi^a \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$ and $\mathcal{X}_\phi^b \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$, the representations of graphs $\mathcal{G}_a(V_a, E_a)$ and $\mathcal{G}_b(V_b, E_b)$, our task is to compare between $\mathcal{G}_a$ and $\mathcal{G}_b$ by solving optimal transport problem between $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$.

CLASSIC PROBLEM SETTING. In GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022), researchers study the pure optimal transport between $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$, where the 2-Wasserstein distance in $\mathcal{P}(\Omega)$, the space of probability measures, is given as

$$\mathsf{OT}_2\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b\right) = \inf_{\gamma} \int_{\Omega\times\Omega} \|x-y\|_2 \mathsf{d}\gamma(x,y), \text{ s.t. } \gamma \in \Gamma_{ab}. \tag{4}$$

Notion Γ_{ab} is the set of joint probabilities γ such that $\int \gamma(x,y)\,\mathsf{d}y = \rho_a(x)$ and $\int \gamma(x,y)\,\mathsf{d}x = \rho_b(y)$, where $\rho_a(\cdot)$ and $\rho_b(\cdot)$ are the probability distributions of $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$, respectively. Notion $\|\cdot\|_2$ denotes the L_2 norm.

In Petric Maretic et al. (2019); Petric Maretic (2021); Maretic et al. (2022), the solution of this optimal transport problem is constrained as a permutation matrix $M \in \mathbb{R}^{|V_b|\times|V_a|}$ for practicability (see **Appendix B** for details). Therefore, the optimization objective is

$$\underset{M\in\mathbb{R}^{|V_b|\times|V_a|}}{\text{minimize}}\ \mathsf{OT}_2\left(\mathcal{X}_\phi^a, M\circ\mathcal{X}_\phi^b\right), \text{ s.t. } M \text{ is a permutation matrix}, \tag{5}$$

where $M\circ\mathcal{X}_\phi^b \sim \mathcal{N}\left(\mathbf{0}, M^T\Sigma_b M\right)$. Applying the Gaussian properties of $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$, the objective can be analytically derived Takatsu (2011).

$$\mathsf{OT}_2\left(\mathcal{X}_\phi^a, M\circ\mathcal{X}_\phi^b\right) = \mathrm{tr}\left(\Sigma_a + M^T\Sigma_b M - 2\sqrt{\Sigma_a^{\frac{1}{2}} M^T \Sigma_b^\dagger M \Sigma_a^{\frac{1}{2}}}\right). \tag{6}$$

The non-convex discrete problem in Eqs. (5-6) may have a factorial number of feasible solutions. To avoid this difficulty, researchers apply $\varsigma(\cdot)$, the *Sinkhorn operator* Mena et al. (2017), to transform the discrete problem into a differentiable one (see **Appendix B** for the definition of $\varsigma(\cdot)$ in detail)

$$\underset{\varsigma(M/\tau)\in\mathbb{R}^{|V_b|\times|V_a|}}{\text{minimize}}\ \mathsf{OT}_2\left(\mathcal{X}_\phi^a, \varsigma(M/\tau)\circ\mathcal{X}_\phi^b\right), \text{ s.t. } \varsigma(M/\tau) \text{ is a doubly stochastic matrix}, \tag{7}$$

where $\tau \in (0,\infty)$, the *Sinkhorn operator* satisfies that $\lim_{\tau\to 0^+}\varsigma(M/\tau)$ is a permutation matrix (see **Appendix B** for detailed explanations) Mena et al. (2017); Petric Maretic et al. (2019), and any $\varsigma(M/\tau)$ is a doubly stochastic matrix in the *Birkhoff polytope* Mena et al. (2017). The problem in Eq. (7) supports automatic differentiation Genevay et al. (2017) and converges to the original problem in Eq. (6) as $\tau \to 0^+$. Certainly, the above derivation implicitly requires that $|V_a| = |V_b|$ because any doubly stochastic matrix is a square matrix. This constraint is hold by GOT Petric Maretic et al. (2019); Petric Maretic (2021) and is relaxed in fGOT Maretic et al. (2022), where $|V_a| \neq |V_b|$ is allowed to compare between graphs with different sizes.

OUR PROBLEM SETTING. Our problem setting differs from the classic one at the very beginning, where we consider the entropy-regularized optimal transport between $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$ Cuturi (2013)

$$\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b\right) = \inf_{\gamma}\left(\int_{\Omega\times\Omega}\|x-y\|_2\mathsf{d}\gamma(x,y) + \varepsilon D_{\mathrm{KL}}(\gamma\|\rho_a\otimes\rho_b)\right), \text{ s.t. } \gamma\in\Gamma_{ab} \tag{8}$$

applying the Kullback-Leibler divergence $D_{\mathrm{KL}}(\cdot\|\cdot)$

$$D_{\mathrm{KL}}(\gamma\|\rho_a\otimes\rho_b) = \int_{\Omega\times\Omega}\log\left(\frac{\mathsf{d}\gamma}{\mathsf{d}\rho_a\mathsf{d}\rho_b}\right)\mathsf{d}\gamma. \tag{9}$$

Parameter $\varepsilon \in [0, \infty)$ in Eq. (8) is the entropy regularization magnitude.

In our research, we also constrain the solution of our entropy-regularized optimal transport problem as a permutation matrix $M \in \mathbb{R}^{|V_b| \times |V_a|}$ and make it differentiable by using the *Sinkhorn operator*. Applying the properties of Gaussian geometry, we can follow Mallasto et al. (2021) to derive an closed-form expression of the optimization objective (see **Appendix C** for derivations)

$$\begin{aligned}&\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)\\ =&\operatorname{tr}\left(\Sigma_a + \varsigma\left(M/\tau\right)^T \Sigma_b \varsigma\left(M/\tau\right)\right) - \frac{\varepsilon}{2}\left(\operatorname{tr}\left(K_{ab}^\varepsilon\right) - \log\det\left(K_{ab}^\varepsilon\right) + |V_a|\log 2 - 2|V_a|\right), \quad (10)\end{aligned}$$

where $K_{ab}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2}\Sigma_a \varsigma\left(M/\tau\right)^T \Sigma_b \varsigma\left(M/\tau\right)}$, notion I denotes the unit matrix, and $\det\left(\cdot\right)$ denotes the determinant. Such a differentiable problem can be solved by gradient descent, whose algorithm will be introduced later.

Why should we consider an entropy-regularized optimal transport problem? The main reason lies in that entropy-regularized optimal transport has lower sample complexity (i.e., the convergence speed of a metric between a measure and its empirical counterpart as a function of sample size) than the pure one Weed & Bach (2019); Mallasto et al. (2021); Mena & Niles-Weed (2019) and, therefore, helps overcome challenge (II). This property is demonstrated as favorable in our experiments.

Bias control via 2-Sinkhorn divergence. To this point, we have proposed our entropy-regularized optimal transport problem between graphs, which is more computationally favorable than the original one studied by GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022). However, such a problem is not statistically ideal yet because the entropy-regularization implies a bias in Eq. (10). Specifically, minimizing $\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)$ with respect to $\varsigma\left(M/\tau\right)$ pushes $\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b$ towards a shrunk measure with a smaller support set than $\mathcal{X}_\phi^a$, the real target Feydy et al. (2019). This bias arises from the non-vanishing auto-correlation terms $\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a\right)$ and $\mathsf{EO}_\varepsilon\left(\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)$ when $\varepsilon > 0$ Feydy et al. (2019). To control this bias, we need to consider the 2-Sinkhorn divergence Feydy et al. (2019); Mallasto et al. (2021)

$$\begin{aligned}&\mathsf{SK}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)\\ =&\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) - \frac{1}{2}\mathsf{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a\right) - \frac{1}{2}\mathsf{EO}_\varepsilon\left(\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right), \quad (11)\end{aligned}$$

which has a closed-form expression on Gaussian geometry Mallasto et al. (2021)

$$\mathsf{SK}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) = \frac{\varepsilon}{4}\left(\operatorname{tr}\left(K_{aa}^\varepsilon - 2K_{ab}^\varepsilon + K_{bb}^\varepsilon\right) + \log\left(\frac{\det^2\left(K_{ab}^\varepsilon\right)}{\det\left(K_{aa}^\varepsilon\right)\det\left(K_{bb}^\varepsilon\right)}\right)\right). \quad (12)$$

We mark that $K_{aa}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2}\left[\Sigma_a\right]^2}$ and $K_{bb}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2}\left[\varsigma\left(M/\tau\right)^T \Sigma_b \varsigma\left(M/\tau\right)\right]^2}$. One can see **Appendix D** for the derivations of Eq. (12). Besides controlling bias Feydy et al. (2019), the 2-Sinkhorn divergence also enables our entropy-regularized optimal transport to wander between pure optimal transport (with more favorable geometric properties) and *maximum mean discrepancy* (with lower sample complexity) between graphs according to entropy regularization magnitude ε Feydy et al. (2019); Mallasto et al. (2021). Specifically, we have (see **Appendix D** for explanations)

$$\mathsf{OT}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) \xleftarrow{\varepsilon \to 0} \mathsf{SK}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) \xrightarrow{\varepsilon \to \infty} \mathsf{MMD}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right), \quad (13)$$

where $\mathsf{MMD}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) = \left\|\mathbb{E}\left(\mathcal{X}_\phi^a\right) - \mathbb{E}\left(\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)\right\|_2$ denotes the maximum mean discrepancy Feydy et al. (2019); Mallasto et al. (2021). This property enables the 2-Sinkhorn divergence to take the advantages of both pure optimal transport and *maximum mean discrepancy* Feydy et al. (2019). Therefore, we use the 2-Sinkhorn divergence in Eq. (12) rather than the entropy-regularized 2-Wasserstein distance in Eq. (11) to define our optimization problem

$$\underset{\varsigma(M/\tau) \in \mathbb{R}^{|V_b| \times |V_a|}}{\text{minimize}} \ \mathsf{SK}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right), \ \text{s.t. } \varsigma\left(M/\tau\right) \text{ is a doubly stochastic matrix.} \quad (14)$$

The algorithm for solving this problem, designed following the Bayesian exploration and reparameterization introduced in Petric Maretic et al. (2019), is proposed in **Appendix E**.

4 EXPERIMENTS

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Türker Biyikoglu, Josef Leydold, and Peter F Stadler. *Laplacian eigenvectors of graphs: Perron-Frobenius and Faber-Krahn type theorems*. Springer, 2007.

Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pp. 8–pp. IEEE, 2005.

Terry Caelli and Serhiy Kosinov. An eigenspace projection clustering method for inexact graph matching. *IEEE transactions on pattern analysis and machine intelligence*, 26(4):515–519, 2004.

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18 (03):265–298, 2004.

Timothee Cour, Praveen Srinivasan, and Jianbo Shi. Balanced graph matching. *Advances in neural information processing systems*, 19, 2006.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Massimo De Santo, Pasquale Foggia, Carlo Sansone, and Mario Vento. A large database of graphs and its use for benchmarking graph isomorphism algorithms. *Pattern Recognition Letters*, 24(8): 1067–1079, 2003.

Narsingh Deo. *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2017.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016.

Yihe Dong and Will Sawin. Copt: Coordinated optimal transport on graphs. *Advances in Neural Information Processing Systems*, 33:19327–19338, 2020.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.

Vikas Garg and Tommi Jaakkola. Solving graph compression via optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Sinkhorn-autodiff: Tractable wasserstein learning of generative models. *arXiv preprint arXiv:1706.00292*, 7(8), 2017.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Jiao Gu, Bobo Hua, and Shiping Liu. Spectral distances on graphs. *Discrete Applied Mathematics*, 190:56–74, 2015.

Tamás Horváth, Thomas Gärtner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 158–167, 2004.

Bo Jiang, Jin Tang, Chris Ding, Yihong Gong, and Bin Luo. Graph matching via multiplicative update algorithm. *Advances in neural information processing systems*, 30, 2017.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pp. 920–929. PMLR, 2016.

Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Kernels for graphs. In *Kernel methods in computational biology*, pp. 155–170. MIT Press, 2004.

Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.

Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-wasserstein distance between gaussian measures. *Information Geometry*, pp. 1–35, 2021.

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Fgot: Graph distances based on filters and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7710–7718, 2022.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

Gonzalo Mena, David Belanger, Gonzalo Munoz, and Jasper Snoek. Sinkhorn networks: Using optimal transport techniques to learn permutations. In *NIPS Workshop in Optimal Transport and Machine Learning*, volume 3, 2017.

Ahmad Mheich, Fabrice Wendling, and Mahmoud Hassan. Brain network similarity: methods and applications. *Network Neuroscience*, 4(3):507–527, 2020.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018.

Hermina Petric Maretic. Representing graphs through data with learning and optimal transport. Technical report, EPFL, 2021.

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Got: an optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672. PMLR, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Christian Schellewald and Christoph Schnörr. Probabilistic subgraph matching based on convex relaxation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 171–186. Springer, 2005.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pp. 488–495. PMLR, 2009.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011.

Yang Tian, Hedong Hou, Guangzheng Xu, Yaoyuan Wang, Ziyang Zhang, and Pei Sun. Analytic relations between complex networks: encoding, decoding, and causality. *arXiv e-prints*, pp. arXiv–2207, 2022.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Nenad Trinajstic. *Chemical graph theory*. CRC press, 2018.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1365–1374, 2015.

Tianshu Yu, Junchi Yan, Yilin Wang, Wei Liu, et al. Generalizing graph matching beyond quadratic assignment model. *Advances in neural information processing systems*, 31, 2018.