ERGOT: ENTROPY-REGULARIZED GRAPH OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph comparison is a fundamental task, which not only relates to graph matching, an NP-hard problem, but also has various applications in graph learning. We tackle this task by studying optimal graph representation and the entropy-regularized optimal transport 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 formulating ErGOT, a framework with low sample complexity, on represented graph information. Third, we control biases in the solution by defining ErGOT with 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, ErGOT defined with 2-Sinkhorn divergence wanders between pure optimal transport and maximum mean discrepancy among graphs. We demonstrate that these statistically efficient, principally unbiased, and in-between properties ensure theoretically faster convergence of our approach to empirically higher performance than the state-of-art algorithms on graph alignment, sketching, and retrieval 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 analysis (e.g., alignment) frequently begins 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 similarity and difference 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 exploiting 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 variable $\mathcal{X} \sim \mathcal{N}\left(\mathbf{0}, \mathcal{Q}^2(L)\right)$, where $\mathcal{Q}(\cdot)$ 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 $\mathcal{Q}(\cdot)$. In fGOT Maretic et al. (2022), more definitions of $\mathcal{Q}(\cdot)$ are proposed from the perspective of engineering practice. 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 named as ErGOT. Our first contribution is to analytically derive a family of Gaussian variables that optimally represents 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, 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 objective of ErGOT with low biases. Driven by 2-Sinkhorn divergence, ErGOT can complement the advantages of 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)).

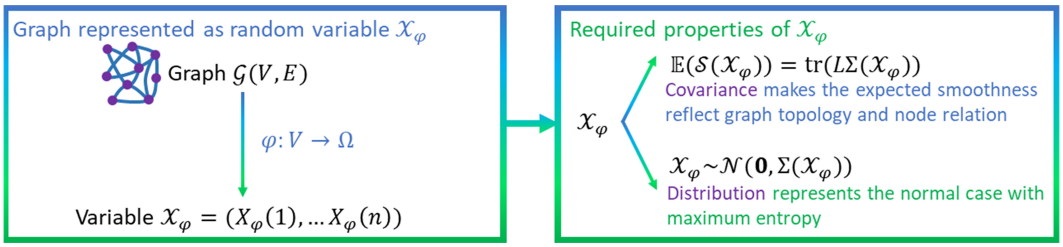

Figure 1: Conceptual illustrations of optimal graph representation.

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 $n = |V|$ (**Fig. 1**). 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 graph properties via the smoothness of ϕ and the distribution of X_ϕ (**Fig. 1**).

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 way, $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$ creates an expected smoothness index that is independent of graph topology and fully determined by graph size n (notion $\dagger$ denotes the Moore–Penrose pseudoinverse Barata & Hussein (2012)). 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	Expected smoothness index	Expressed graph properties
$\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$	$\mathbb{E}(\mathcal{S}(\phi)) = \mathrm{tr}\left(L^2\right)$	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.

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 $\sqrt{r}\langle \mathcal{X}_\phi \rangle \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\left(\mathcal{X}_\phi\right)\right)$ as $r \to \infty$. Consequently, one can directly define

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

to represent the normal case in application (**Fig. 1**). Because $\mathbb{E}\left(\mathcal{X}_\phi\right)$ and $\Sigma\left(\mathcal{X}_\phi\right)$ are finite, the Gaussian variable in Eq. (3) is also the random variable in $\mathbb{R}$ that has a maximum entropy Cover (1999) (**Fig. 1**). 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 $\mathcal{X}_\phi^a \sim \mathcal{N}\left(\mathbf{0}, \Sigma_a\right)$ and $\mathcal{X}_\phi^b \sim \mathcal{N}\left(\mathbf{0}, \Sigma_b\right)$, the representations of graphs $\mathcal{G}_a\left(V_a, E_a\right)$ and $\mathcal{G}_b\left(V_b, E_b\right)$, 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 APPROACH In GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022), researchers study the pure optimal transport (**Fig. 2**) between $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$, where the 2-Wasserstein distance in $\mathcal{P}\left(\Omega\right)$, the space of probability measures, is given as

$$\mathsf{OT}_2\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b\right) = \inf_{\gamma} \int_{\Omega\times\Omega} \|x-y\|_2^2 \mathsf{d}\gamma\left(x,y\right), \text{ s.t. } \gamma \in \Gamma_{ab}. \tag{4}$$

Notion Γ_{ab} is the set of joint probabilities γ such that $\int \gamma\left(x,y\right)\mathsf{d}y = \rho_a\left(x\right)$ and $\int \gamma\left(x,y\right)\mathsf{d}x = \rho_b\left(y\right)$, where $\rho_a\left(\cdot\right)$ and $\rho_b\left(\cdot\right)$ 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 (**Fig. 2**)

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

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).

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

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

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

where $\tau \in (0,\infty)$, the *Sinkhorn operator* satisfies that $\lim_{\tau\to 0^+}\varsigma\left(M/\tau\right)$ is a permutation matrix (see **Appendix B** for detailed explanations) Mena et al. (2017); Petric Maretic et al. (2019), and any $\varsigma\left(M/\tau\right)$ is a doubly stochastic matrix in the *Birkhoff polytope* Mena et al. (2017) (**Fig. 2**). 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. Please note that although the *Sinkhorn operator* implicitly creates entropy regularization during computation Eisenberger et al. (2022), the formal optimization objective defined in Eq. (6) still belongs to pure optimal transport problem. Please see **Appendix B** for explanations.

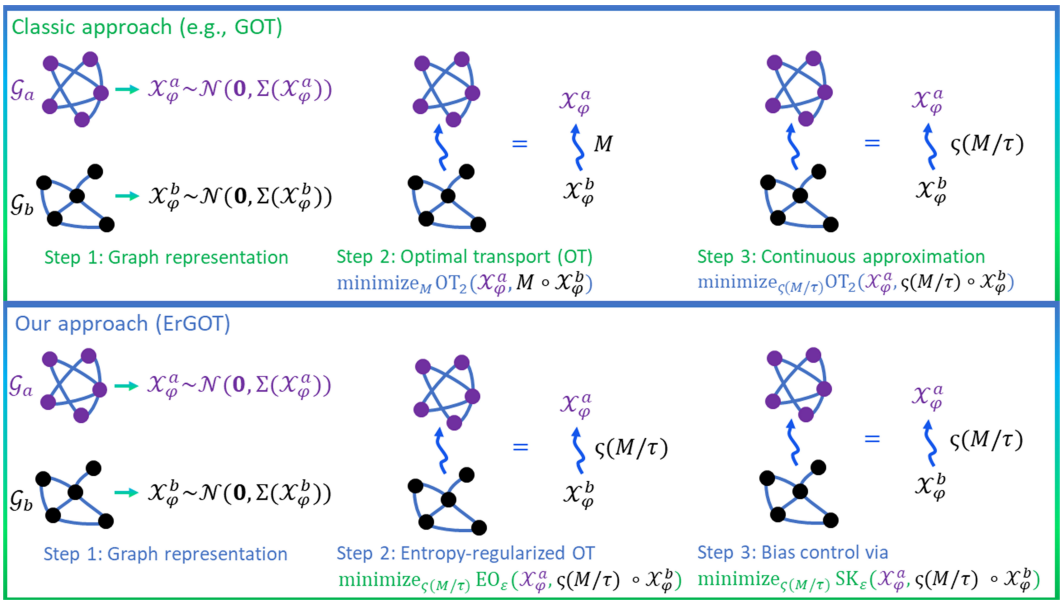

Figure 2: Conceptual illustrations of classic approach and our approach. Please note that "steps" are used to describe the progressive relation of mathematical ideas rather than real steps in algorithms.

OUR APPROACH (STEPS 1-2). Our approach differs from the classic one at the very beginning. We start by explicitly considering the entropy-regularized optimal transport between $\mathcal{X}_\phi^a$ and $\mathcal{X}_\phi^b$ Cuturi (2013)

$$\text{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b\right) = \inf_\gamma \left(\int_{\Omega \times \Omega} \|x - y\|_2^2 \mathrm{d}\gamma\left(x, y\right) + \varepsilon D_{\text{KL}}\left(\gamma \| \rho_a \otimes \rho_b\right) \right), \text{ s.t. } \gamma \in \Gamma_{ab} \tag{8}$$

where $D_{\text{KL}}\left(\cdot \| \cdot\right)$ denotes the Kullback-Leibler divergence

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

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* (**Fig. 2**). 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)

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

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 rate 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.

OUR APPROACH (STEP 3). 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).

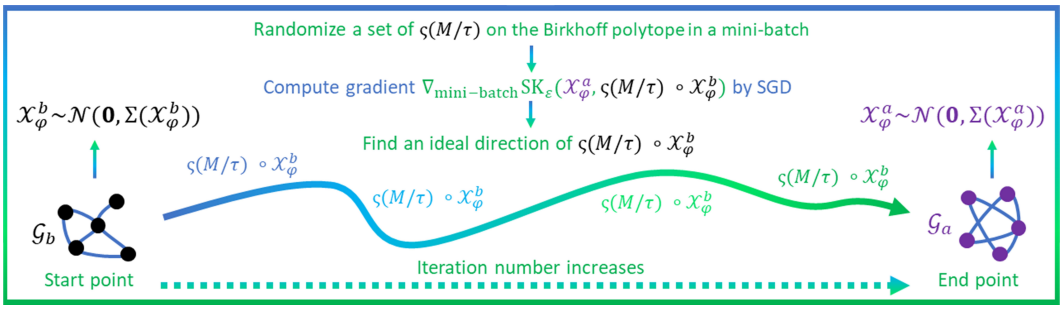

Figure 3: Conceptual illustrations of the computational implementation of ErGOT.

However, such a problem is not statistically ideal yet because the entropy-regularization implies a bias in Eq. (10). Specifically, minimizing $\text{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)$ 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 $\text{EO}_\varepsilon\left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a\right)$ and $\text{EO}_\varepsilon\left(\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)$ 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)

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

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

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

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)

$$\text{OT}_2\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) \xleftarrow{\varepsilon \to 0} \text{SK}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) \xrightarrow{\varepsilon \to \infty} \text{MMD}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right), \tag{13}$$

where $\text{MMD}_\varepsilon\left(\mathcal{X}_\phi^a, \varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right) = \|\mathbb{E}\left(\mathcal{X}_\phi^a\right) - \mathbb{E}\left(\varsigma\left(M/\tau\right) \circ \mathcal{X}_\phi^b\right)\|_2$ 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 (**Fig. 2**)

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

The algorithm for solving this problem is designed following the Bayesian exploration and re-parameterization introduced in Petric Maretic et al. (2019). Because gradient descent algorithm is not our research objective, here we do not elaborate it in the main text. Please see **Appendix E** for information of algorithm designs. In **Fig. 3**, we sketch the general mechanics of ErGOT algorithm.

4 Experiments

Experiment objectives. To comprehensively validate our ErGOT framework, we implement it on graph alignment, sketching, and retrieval tasks (**Fig. 4**) to compete with corresponding state-of-the-art approaches. Based on these experiments, we aim at verifying the following arguments:

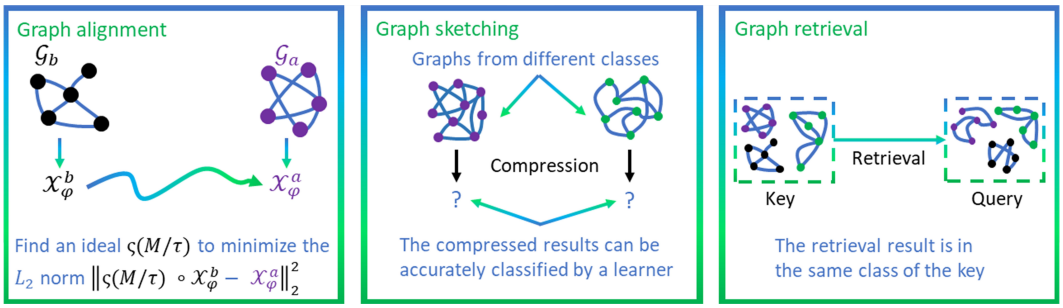

Figure 4: Conceptual illustrations of experiment objectives.

(A) ErGOT can robustly and effectively achieve high accuracy across different graph alignment conditions (e.g., graph sizes, graph types, covariance matrix designs, and sample sizes).

(B) Although ErGOT is proposed for graph alignment, it is also competitive on other tasks such as graph sketching and retrieval. Certainly, a more specialized algorithm may surpass ErGOT in a corner case. On average, ErGOT is expected to be optimal in most cases.

Graph alignment. Our graph alignment experiment is implemented on the stochastic block model (initialized with 2 communities, where edges within communities are removed with a probability of 0.3 and edges across communities are removed with a probability of 0.95) and the Erdős-Rényi graph (nodes are randomly connected according to a probability of 0.45). Once a graph is generated under an experiment condition, we randomly permute the order of nodes to obtain a graph to be aligned. The objective is to find a permutation matrix that best aligns the randomly permuted graph with the original one (**Fig. 4**).

We compare ErGOT ($\varepsilon = 500$, $\tau = 2$, and learning rate $\gamma = 0.5$) with Gromov-Wasserstein approach (GW, defined with an entropy regularization magnitude of 10^{-3} to avoid numerical problems) Peyré et al. (2016) and classic graph-signal-based optimal transport approaches (i.e., GOT Petric Maretic et al. (2019) and fGOT Maretic et al. (2022)). Because fGOT can be understood as a generalized version of GOT (e.g., generalized to more covariance matrices) and essentially solves the same question as GOT (i.e., the optimal transport defined in Eq. (7)), we uniformly implement GOT (default settings with $\tau = 2$ and $\gamma = 0.2$) with different covariance matrices rather than repeatedly deploy GOT and fGOT in our experiment. Specifically, we consider four kinds of covariance matrix designs, i.e., $\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$ (marked as Σ_1), $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$ (marked as Σ_2), $\Sigma(\mathcal{X}_\phi) = \exp(-rL)$ (marked as Σ_3), and $\Sigma(\mathcal{X}_\phi) = L^2$ (marked as Σ_4). Among them, Σ_1 and Σ_2 are our theoretical results in **Table 1** while Σ_3 and Σ_4 are adopted from fGOT Maretic et al. (2022).

The experiment is conducted in a CPU environment (Intel i7-8750H) with 32GB memory. To ensure the validity of our results, the experiment under each single condition is repeated for 20 times. Each time a graph is generated and randomly permuted to define the alignment task for ErGOT (1000 epochs), GW, and GOT (1000 epochs). Because we find that GOT may occasionally meet numerical problems (e.g., involves with inf or nan), we distinguish between two versions of alignment experiments, where we either apply a numerical trick (i.e., add $0.1I$, a scaled unit matrix, to the covariance matrix Petric Maretic et al. (2019); Maretic et al. (2022)) to resolve these problems (all results in **Fig. 5**) or simply leave them alone (all results in **Appendix F**). Please note that the Gromov-Wasserstein approach is accelerated by C++ Flamary et al. (2021) and, therefore, is much faster than the other two algorithms programmed by Python.

In **Fig. 5a** and **Fig. 5d**, we empirically demonstrate the low sample complexity of ErGOT by comparing it with GOT on different sample sizes (sample size $s \in \{5, 10, 15, 20, 25, 30, 35, 40\}$). After averaging the alignment errors (denoted by the L_2 distance between alignment targets and algorithm outputs) across all covariance matrix designs, we show that ErGOT achieves lower errors on all sample sizes. In **Fig. 5b** and **Fig. 5e**, we show the alignment accuracy of ErGOT ($s = 10$), GOT ($s = 10$), and GW as the graph size increases. Compared with GOT and GW, ErGOT is observed to maintain higher alignment precision across different graph sizes. In **Fig. 5c** and **Fig. 5f**, we measure the time cost of alignment by GW, ErGOT ($s = 10$), and GOT ($s = 10$), respectively, where ErGOT is shown as faster than GOT with the same epoch number (again, ErGOT and GOT

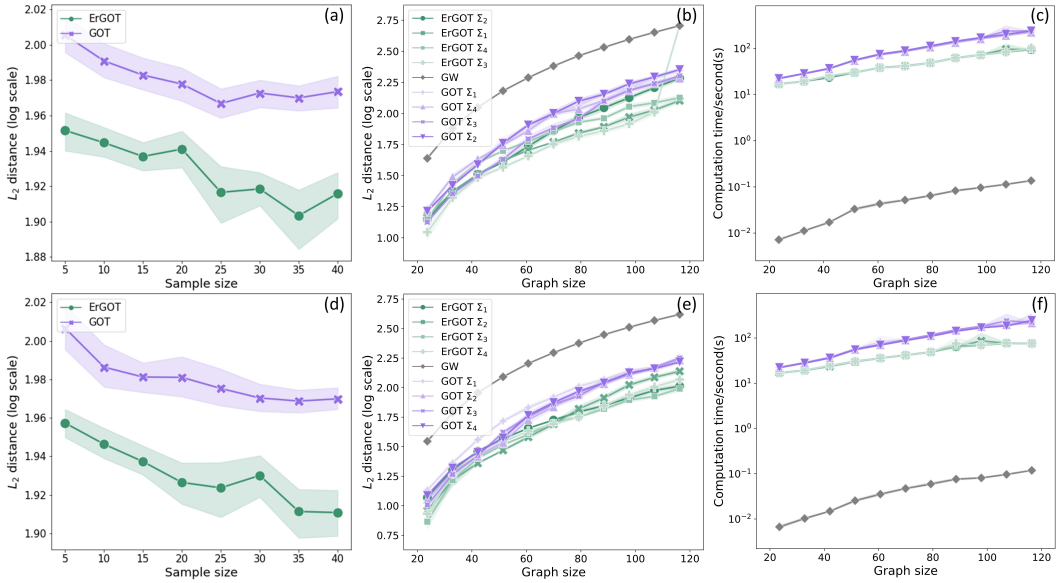

Figure 5: The results of graph alignment (with numerical tricks). (a-c) report the experiment on Erdős-Rényi graphs while (d-f) report the experiment on stochastic block models. (a) and (d) show the mean performance (averaged across different covariance matrix designs) of ErGOT and GOT as the sample size increases. (b) and (e) compare ErGOT with GOT and GW on different graph sizes. (c) and (f) measure the time cost of ErGOT (1000 epochs), GOT (1000 epochs), and GW.

can not be directly compared with GW because GW is accelerated by C++). These findings can be validated by the experiment without the numerical trick (see **Appendix A**) as well, where ErGOT surpasses GOT in a higher degree than our observations in **Fig. 5**. In sum, our results suggest ErGOT as a better choice in graph alignment tasks.

Graph sketching. In our graph sketching experiment, ErGOT ($\varepsilon = 500$ and $\tau = 2$) is compared with several state-of-the-art algorithms, including OTC Garg & Jaakkola (2019), COPT Dong & Sawin (2020), Algebraic Chen & Safro (2011); Ron et al. (2011), Variation Deo (2017), REC Gu et al. (2015), Affinity Livne & Brandt (2012), and HeavyE Dhillon et al. (2007) on representative data sets. In general, graph sketching is a process where we compress a graph $\mathcal{G}_a$ to a smaller graph $\mathcal{G}_b$ with a given size. The objective is to search an optimal candidate of $\mathcal{G}_b$ such that a certain distance between $\mathcal{G}_a$ and $\mathcal{G}_b$ can be minimized Dong & Sawin (2020). In our research, the graph sketching via ErGOT is realized based on the algorithm proposed by COPT Dong & Sawin (2020). Specifically, we modify the algorithm by replacing the COPT distance metric between the original graph and its potential sketching candidate by the 2-Sinkhorn divergence between $\mathcal{G}_a$ and $\mathcal{G}_b$.

We first compress all graphs by a give factor, after which we randomly select 70% of the compressed graphs to define the training set and train a SVM classifier with multi-scale Laplacian graph kernel Kondor & Pan (2016). Then, we evaluate the classification accuracy of SVM on the rest part of graphs. One can see a similar experiment in COPT Dong & Sawin (2020). Due to the limited space, our experiment results are subdivided into two parts, which are reported in **Table 2** and **Appendix F**, respectively.

For an ideal graph sketching approach, compressed graphs are expected to contain sufficient information of the original ones such that a classifier can accurately classify them. On the three data sets reported in **Table 2**, ErGOT achieves optimal (being the best) or competitive (being the second best) performance in 2-fold and 4-fold compression tasks. Although some specialized algorithms for graph compression can achieve higher performance than ErGOT on other graph data sets (see **Appendix F**), ErGOT, an approach that is not initially proposed for graph compression, can keep the performance gap within a reasonable range. These results are acceptable because it is impossible for a single algorithm to achieve the best performance in every learning tasks implemented on every data sets (i.e., none of the considered state-of-the-art methods can do so). Therefore, we suggest that

	2X compression			4X compression		
	BZR_MD	MSRC_9	Proteins	BZR_MD	MSRC_9	Proteins
OTC	60.7 ± 4.0	80.9 ± 4.5	$72.8 \pm .8$	64.3 ± 2.7	84.8 ± 6.7	66.7 ± 1.8
HeavyE	61.7 ± 4.8	79.7 ± 6.3	72.3 ± 3.3	55.0 ± 4.7	76.1 ± 7.9	72.2 ± 2.7
Variation	60.2 ± 4.4	75.5 ± 2.7	72.1 ± 1.2	59.3 ± 3.2	78.5 ± 3.8	$72.4 \pm .75$
Algebraic	57.4 ± 5.2	77.0 ± 8.9	70.1 ± 2.7	53.4 ± 2.5	75.2 ± 6.9	69.1 ± 1.8
Affinity	58.5 ± 5.0	80.1 ± 3.0	71.2 ± 2.5	53.4 ± 3.5	75.8 ± 6.2	70.9 ± 2.3
REC	60.9 ± 7.3	82.4 ± 1.9	71.1 ± 1.5	54.5 ± 2.7	77.9 ± 3.7	71.5 ± 1.0
COPT	67.6 ± 4.0	86.3 ± 1.3	$\mathbf{74.0 \pm 1.3}$	68.4 ± 5.0	81.2 ± 4.8	73.7 ± 1.5
ErGOT	$\mathbf{68.6 \pm 4.0}$	$\mathbf{86.9 \pm 7.99}$	73.1 ± 3.4	$\mathbf{70.0 \pm 6.56}$	$\mathbf{86.2 \pm 8.66}$	$\mathbf{75.2 \pm 2.4}$

Table 2: Graph sketching experiment results on 3 data sets. The best model is marked in bold while the second best one is marked by the underline.

Algorithm	MSRC_9	PROTEINS	BZR	MUTAG	MSRC_21C	DHFR	COX2_MD
Affinity	14.95 ± 3.87	70.80 ± 5.09	76.0 ± 4.40	56.65 ± 9.42	10.3 ± 3.23	55.5 ± 6.14	41.9 ± 5.03
Heavy	13.65 ± 2.67	70.30 ± 4.42	78.6 ± 4.55	70.65 ± 7.70	7.25 ± 3.88	$53.5 \pm 4.6\ 9$	49.25 ± 4.86
Algebraic	11.60 ± 3.27	69.35 ± 4.17	69.7 ± 7.41	66.0 ± 6.50	10.45 ± 3.64	54.0 ± 5.33	36.5 ± 6.08
Variation	14.20 ± 4.01	$\mathbf{77.6 \pm 4.26}$	75.85 ± 4.97	70.6 ± 4.39	9.85 ± 2.90	55.25 ± 3.59	37.4 ± 5.74
REC	$\mathbf{16.75 \pm 3.25}$	71.45 ± 3.58	$\mathbf{79.5 \pm 3.71}$	60.2 ± 5.08	11.3 ± 2.88	51.75 ± 4.97	49.25 ± 4.86
OTC	9.50 ± 2.71	54.6 ± 4.91	$72{,}35 \pm 7.21$	68.1 ± 5.74	12.25 ± 3.45	$\mathbf{59.5 \pm 6.67}$	49.4 ± 4.71
COPT	13.6 ± 2.86	54.1 ± 3.58	68.25 ± 3.88	54.5 ± 4.65	10.45 ± 3.50	51.4 ± 5.04	50.15 ± 5.28
ErGOT	15.7 ± 2.53	64.5 ± 5.83	74.45 ± 4.26	$\mathbf{77.65 \pm 4.05}$	$\mathbf{14.35 \pm 3.77}$	57.05 ± 4.75	$\mathbf{50.5 \pm 5.46}$

Table 3: Graph retrieval experiment results on 7 data sets. The best model is marked in bold while the second best one is marked by the underline. Note that two algorithms, Heavy and REC, happen to achieve the same performance on COX2_MD data set.

ErGOT has the potential to be further explored in terms of graph sketching, yet more specialized improvements are necessary.

Graph retrieval. Our graph retrieval experiment begins with a graph compression step, where we compress graphs in every data set to a certain number of nodes (see **Appendix F** for the settings of node number on each data set). Similar to our previous experiment, the graph compression by ErGOT is still realized by a modified COPT algorithm, where distance metric is set as the 2-Sinkhorn divergence ($\varepsilon = 500$ and $\tau = 2$) between the original graph and the compressed one. In each sketched data set, we randomly select 100 graphs as keys and use the rest part of graphs as queries. For every key graph, we search its nearest neighbor (searched according to L_2 distance between graph Laplacian matrices) in query graphs. Then, the label of the nearest neighbor is used as the predicted label to evaluate prediction accuracy (i.e., if the predicted label is same as the label of key graph). Our experiment is repeated for 20 times on each data set with different random seeds.

The test accuracy on 7 real-world data sets is reported in **Table 3**. After counting the number of times for an algorithm to be optimal on a data set (e.g., being the best or the second best), ErGOT is observed to be optimal for 5 times while its strongest competitors become optimal for 2 or 3 times at most. On average, ErGOT is suggested as an optimal choice for graph retrieval tasks though there do exist some specialized algorithms that are better than ErGOT on specific data sets.

5 Conclusion

In this paper, we formalize graph comparison task as an entropy-regularized optimal transport problem between the optimal representation of graphs. We propose ErGOT, an approach with comprehensive theoretical foundation and promising properties (e.g., low sample complexity), to solve the problem. In general, the ErGOT framework defined with a 2-Sinkhorn divergence is demonstrated as not only optimal in graph alignment tasks but also efficient in the tasks for which ErGOT is not initially proposed (e.g., graph sketching and retrieval). Although being optimal on average, ErGOT may be surpassed by specific specialized algorithms in corner cases. Therefore, a meaningful direction for future studies may be to develop task-specific variants of ErGOT. The source code of ErGOT will be released once the double blind review finishes.

REFERENCES

João Carlos Alves Barata and Mahir Saleh Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165, 2012.

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.

Jonathan M Borwein, Adrian Stephen Lewis, and Roger D Nussbaum. Entropy minimization, dad problems, and doubly stochastic kernels. *Journal of Functional Analysis*, 123(2):264–307, 1994.

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.

Pavel Chebotarev and Elena Shamis. On proximity measures for graph vertices. *arXiv preprint math/0602073*, 2006.

Jie Chen and Ilya Safro. Algebraic distance on graphs. *SIAM Journal on Scientific Computing*, 33 (6):3468–3490, 2011.

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.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

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.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11): 1944–1957, 2007.

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.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Florian Bernard, and Daniel Cremers. A unified framework for implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 509–518, 2022.

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.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

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.

Ivan Gutman and W Xiao. Generalized inverse of the laplacian matrix and some applications. *Bulletin (Académie serbe des sciences et des arts. Classe des sciences mathématiques et naturelles. Sciences mathématiques)*, pp. 15–23, 2004.

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.

Mohammad Emtiyaz Khan, Wu Lin, Voot Tangkaratt, Zuozhu Liu, and Didrik Nielsen. Variational adaptive-newton method for explorative learning. *arXiv preprint arXiv:1711.05560*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Risi Kondor and Horace Pan. The multiscale laplacian graph kernel. *Advances in neural information processing systems*, 29, 2016.

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

Oren E Livne and Achi Brandt. Lean algebraic multigrid (lamg): Fast graph laplacian linear solver. *SIAM Journal on Scientific Computing*, 34(4):B499–B522, 2012.

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.

Simone Di Marino and Augusto Gerolin. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):1–28, 2020.

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.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

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.

Dorit Ron, Ilya Safro, and Achi Brandt. Relaxation-based coarsening and multiscale graph organization. *Multiscale Modeling & Simulation*, 9(1):407–423, 2011.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

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.

Piet Van Mieghem, Karel Devriendt, and H Cetinay. Pseudoinverse of the laplacian and best spreader node in a network. *Physical Review E*, 96(3):032311, 2017.

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.

Wenjun Xiao and Ivan Gutman. Resistance distance and laplacian spectrum. *Theoretical chemistry accounts*, 110(4):284–289, 2003.

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.