Exploiting Edge Features in Graph-based Learning with Fused Network Gromov-Wasserstein Distance

Junjie Yang [1] Matthieu Labeau [1] Florence d'Alché-Buc [1]

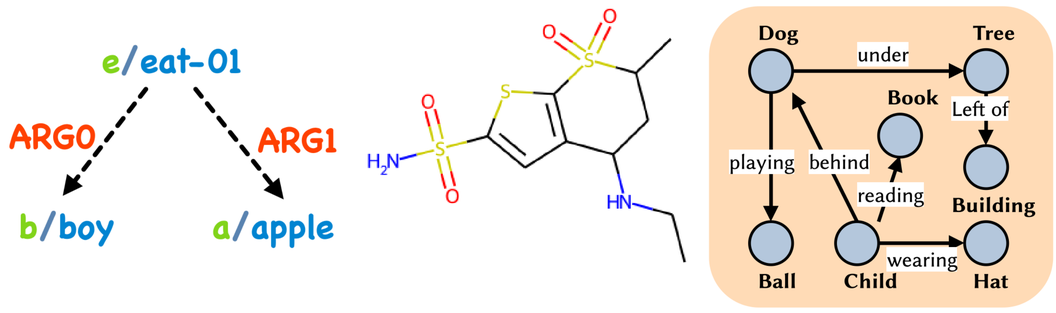
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Motivation

Problem: Graph-based Learning with Optimal Transport

Existing Works: Graph Classification with FGW [5], Graph Clustering with GW [4], FGW [5] or NGW [2], Graph Dictionary Learning with (F)GW [6], Supervised Graph Prediction with FGW [1]

Goal: We want to unlock OT-based learning for edge featured graphs. We target especially Supervised Graph Prediction task.

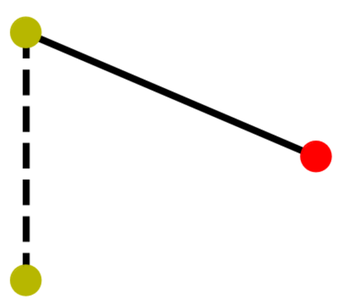


Node and Edge Featured Graph

A node and edge featured graph of size m is a quadruple of the form $(F, A, E, \boldsymbol{p})$ where

- $F \in \Psi^m$ is a tuple of points valued in a metric space (Ψ, d_Ψ)
- $A \in \mathbb{R}^{m \times m}$ is a real-valued matrix
- $E \in \Omega^{m \times m}$ is a tuple of points valued in a metric space (Ω, d_Ω)
- $\boldsymbol{p} \in \Sigma_m$ is a simplex histogram

We denote $\mathcal{G}$ as a set of such quadruples.



- $\Psi = \{\text{red}, \text{yellow}\}$: the node-color space
- $\Omega = \{\text{solid}, \text{dashed}, \text{non-edge}\}$: the edge-type space

$$F = \begin{bmatrix}[1,0]\\[1,0]\\[0,1]\end{bmatrix}, A = \begin{bmatrix}0&1&1\\1&0&0\\1&0&0\end{bmatrix}$$

$$E = \begin{bmatrix}[0,0,1] & [0,1,0] & [1,0,0]\\ [0,1,0] & [0,0,1] & [0,0,1]\\ [1,0,0] & [0,0,1] & [0,0,1]\end{bmatrix}, \boldsymbol{p} = \begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix}$$

Fused Network Gromov-Wasserstein Distance

Given $g = (F, A, E, \boldsymbol{p})$ of size m, $\tilde{g} = (\tilde{F}, \tilde{A}, \tilde{E}, \tilde{\boldsymbol{p}})$ of size $\tilde{m}$ corresponding to two tuples of $\mathcal{G}$, and trade-off parameters $(\alpha, \beta) \in [0,1]^2$, the Fused Network Gromov-Wasserstein distance between them for $(p, q) \in [1, \infty]$ is written as :

$$\mathrm{FNGW}_{\alpha,\beta,q,p}(g, \tilde{g}) = \min_{\pi \in \Pi(\boldsymbol{p}, \tilde{\boldsymbol{p}})} \mathcal{E}_{\alpha,\beta,q,p}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi)$$

with

$$\mathcal{E}_{\alpha,\beta,q,p}((F, A, E), (\tilde{F}, \tilde{A}, \tilde{E}), \pi) = \Bigg(\sum_{i,j,k,l} \Big[\alpha d_\Omega \left(E(i,k), \tilde{E}(j,l)\right)^q$$

$$+ \beta |A(i,k) - \tilde{A}(j,l)|^q + (1 - \alpha - \beta) d_\Psi \left(F(i), \tilde{F}(j)\right)^q \Big]^p \pi_{k,l} \pi_{i,j} \Bigg)^{\frac{1}{p}}$$

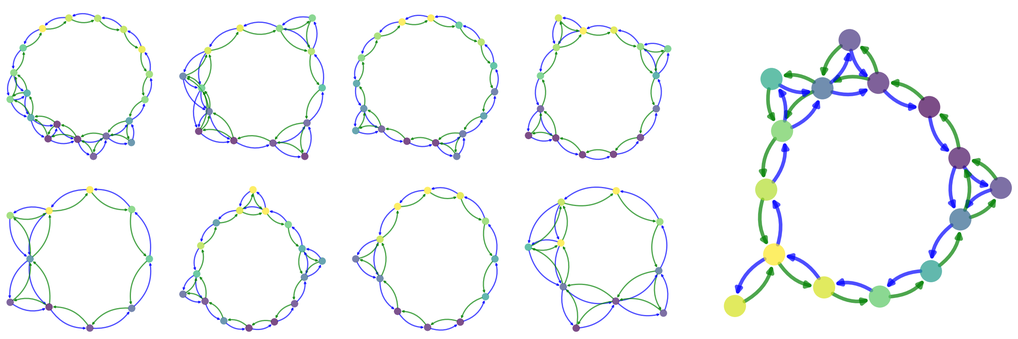
The FNGW distance satisfies the following **metric** properties: positivity, symmetry, equality with a corresponding notion of weak isomorphism, relaxed triangle inequality with a factor of 2^{q-1}.

FNGW Barycenter

Given a set $\{g_k\}_{k=1}^K$ and a set of weights $\{\lambda_k\}_{k=1}^K$ such that $\sum_k \lambda_k = 1$, the FNGW Barycenter for a pre-defined histogram $\boldsymbol{p} \in \Sigma_n$ is defined as follows:

$$\mathrm{Bary}(\{\lambda_k\}_k, \{g_k\}_k, \boldsymbol{p}) = \underset{F \in \mathbb{R}^{n \times S}, A \in \mathbb{R}^{n \times n}, E \in \mathbb{R}^{n \times n \times T}}{\arg\min} \sum_k \lambda_k \mathrm{FNGW}_{\alpha,\beta}((F, A, E, \boldsymbol{p}), g_k)$$

We employ the Block Coordinate Descent (BCD) algorithm to obtain the FNGW barycenter where the tensor E can be updated by $E = \frac{1}{\mathcal{I}_{n \times T \times_2 \boldsymbol{p}\boldsymbol{p}^\top}} \sum_k \lambda_k (E_k \times_2 \pi_k) \times_1 \pi_k$.



(a) Random circle samples

(b) Circle barycenter

Supervised Graph Prediction with FNGW

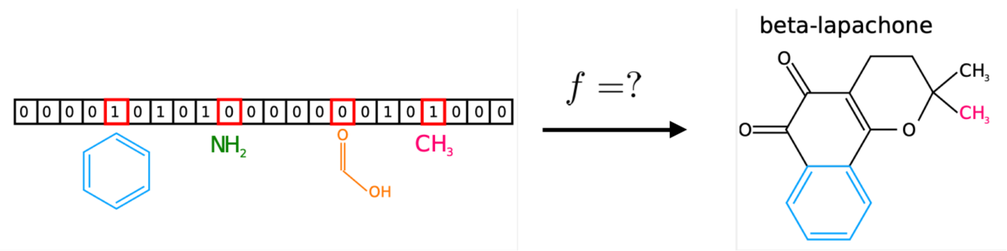
Given input space $\mathcal{X}$, output graph space $\mathcal{G}$, relaxed graph space $\mathcal{G}_m = \{(F, A, E, \boldsymbol{p}) \mid C \in [0,1]^{m \times m}, F = (F_i)_{i=1}^m \in \mathrm{Conv}(\mathcal{F})^m, E = (E_{ij}) \in \mathrm{Conv}(\mathcal{T})^{m \times m}, \boldsymbol{p} = m^{-1}\mathbb{1}_m\}$ where $\mathcal{F} \subset \mathbb{R}^S$ and $\mathcal{T} \subset \mathbb{R}^T$ are finite node and edge feature spaces, and training samples $\{(x_i, g_i)\}_{i=1}^n$, Supervised Graph Prediction requires finding an estimator $f : \mathcal{X} \to \mathcal{G}_m$ of the minimizer f^* of the expected risk $\mathcal{R}(f) = \mathbb{E}_\rho[\mathrm{FNGW}_{\alpha,\beta}(f(X), G)]$. Based on the work of [3, 1], we propose an estimator of the form

$$\hat{f}(x) = \underset{g \in \mathcal{G}_m}{\arg\min} \sum_{i=1}^n \xi(x)_i \mathrm{FNGW}_{\alpha,\beta}(g, g_i)$$

with $\xi(x) = \mathbf{K}S^\top(S\mathbf{K}^2S^\top + n\lambda S\mathbf{K}S^\top)^\dagger S\boldsymbol{\kappa}_x$ where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the input kernel Gram matrix, $\boldsymbol{\kappa}_x = (k(x, x_1), \ldots, k(x, x_n))^\top \in \mathbb{R}^n$, and $S \in \mathbb{R}^{s \times n}$ with $s \ll n$ is a sketching matrix.

- The FNGW loss admits an Implicit Loss Embedding (ILE) $\to \hat{f}$ is universally consistent and its learning rate is of order $n^{-1/4}$ with additional assumptions.
- Sketched ILE enables the supervised graph prediction with more than 100,000 training points.

Experiment: Fingerprint to Molecule



	GED w/o edge feature ↓	GED w/ edge feature ↓
NNBary-FGW	5.000 ± 0.140	-
NNBary-FNGW	5.311 ± 0.090	5.756 ± 0.073
Sketched ILE-FGW	3.037 ± 0.111	-
Sketched ILE-FNGW	$\mathbf{1.449 \pm 0.034}$	$\mathbf{1.534 \pm 0.029}$

Table. Graph edit distances of different methods on the Fin2Mol test set.

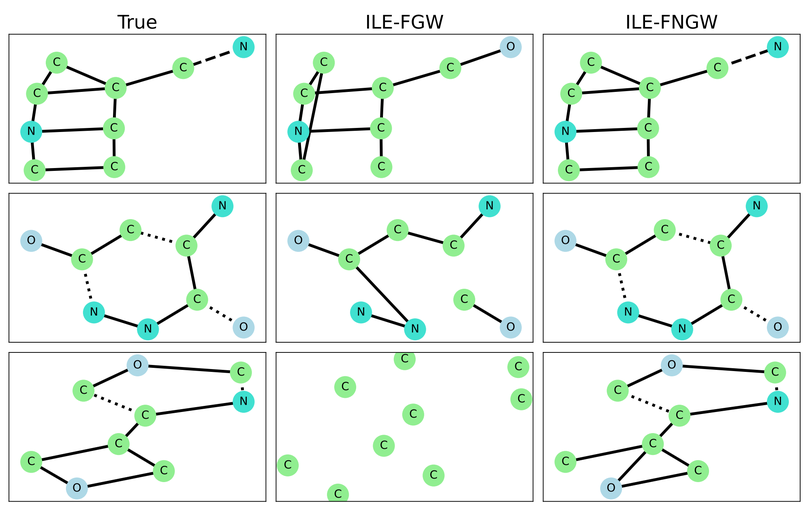


Figure. Qualitative comparison of the predicted QM9 molecules.

Experiment: Metabolite Identification

To solve Metabolite Identification problem, the learning algorithm is expected to predict the metabolite (small molecules) given a tandem mass spectra. For each input spectra, a known set of metabolite candidates is provided.

	Top-1	Top-10	Top-20
WL kernel	9.8%	29.1%	37.4%
Fingerprint with linear kernel	28.6%	54.5%	59.9%
Fingerprint with gaussian kernel	**41.0%**	**62.0%**	**67.8%**
FGW diffuse	28.1%	53.6%	59.9%
FNGW diffuse + Bond stereo	27.7%	55.2%	60.9%
FNGW diffuse + Bond type	34.6%	55.1%	60.0%
FNGW diffuse + Mix	36.2%	58.2%	61.9%

Table. Top-k accuracies on the metabolite identification test set.

References

[1] L. Brogat-Motte, R. Flamary, C. Brouard, J. Rousu, and F. D'Alché-Buc. Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2321–2335, July 2022.

[2] S. Chowdhury and F. Mémoli. The Gromov–Wasserstein Distance Between Networks and Stable Network Invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

[3] C. Ciliberto, L. Rosasco, and A. Rudi. A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings. *JMLR*, 21(98):1–67, 2020.

[4] G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672, June 2016.

[5] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Optimal Transport for structured data with application on graphs. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284, June 2019.

[6] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli, and N. Courty. Online Graph Dictionary Learning. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 10564–10574, July 2021.

Acknowledgements

This research work is supported by the Hi! PARIS Center and the Institut Polytechnique de Paris.

junjie.yang@telecom-paris.fr https://chunchiehy.github.io/