Fused Gromov-Wasserstein barycenter with application on graphs

Zhichen Zeng

10/13

ILLINOIS

Outline

- Motivation
- Preliminaries
- Fused Gromov-Wasserstein barycenter
- Experiments
- Takeaways

Outline

- Motivation
- Preliminaries
- Fused Gromov-Wasserstein barycenter
- Experiments
- Takeaways

Motivation

- Networks are everywhere
  - Alignment (node-level)
  - Clustering (subgraph-level)
  - Comparison (graph-level)
- Distance measures at different levels are important

Motivation

- Common distance measures
  - Invalid in certain cases
  - Ignore underlying structure

KL divergence: $KL(P||Q) = \int p(x) log \frac{p(x)}{q(x)} dx$

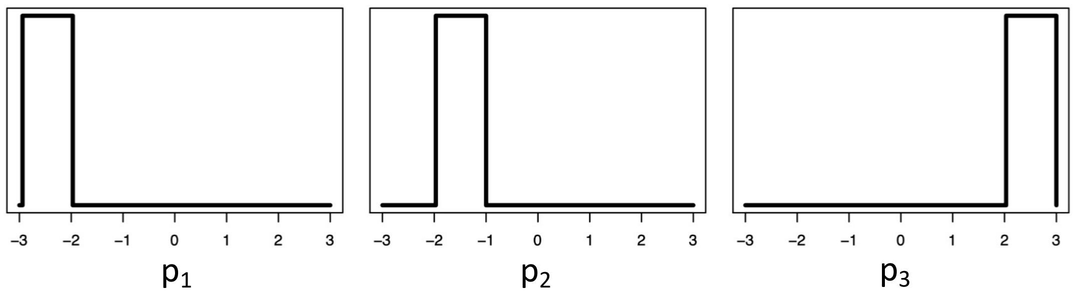
JS divergence: $JS(P||Q) = \frac{1}{2} KL(P|| \frac{P+Q}{2}) + \frac{1}{2} KL(Q|| \frac{P+Q}{2})$

Total variation: $\frac{1}{2} \int |p(x) - q(x)| dx$

Hellinger distance: $\sqrt{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx}$

L_2 distance: $\int (p(x) - q(x))^2 dx$

χ_2 distance: $\int \frac{(p(x) - q(x))}{q(x)} dx$



KL(p1,p2), KL(p1,p3): invalid
JS(p1,p2), JS(p1,p3): invalid

$$\frac{1}{2} \int |p_1(x) - p_2(x)| dx = \frac{1}{2} \int |p_1(x) - p_3(x)| dx = 1$$

- Optimal transport and Wasserstein distance

[1] Peyré, Gabriel, and Marco Cuturi. "Computational Optimal Transport." *arXiv preprint arXiv:1803.00567* (2018).

Outline

- Motivation
- Preliminaries
  - Optimal transport and Wasserstein distance
  - Gromov-Wasserstein distance
  - Fused Gromov-Wasserstein distance
- Fused Gromov-Wasserstein barycenter
- Experiments
- Takeaways

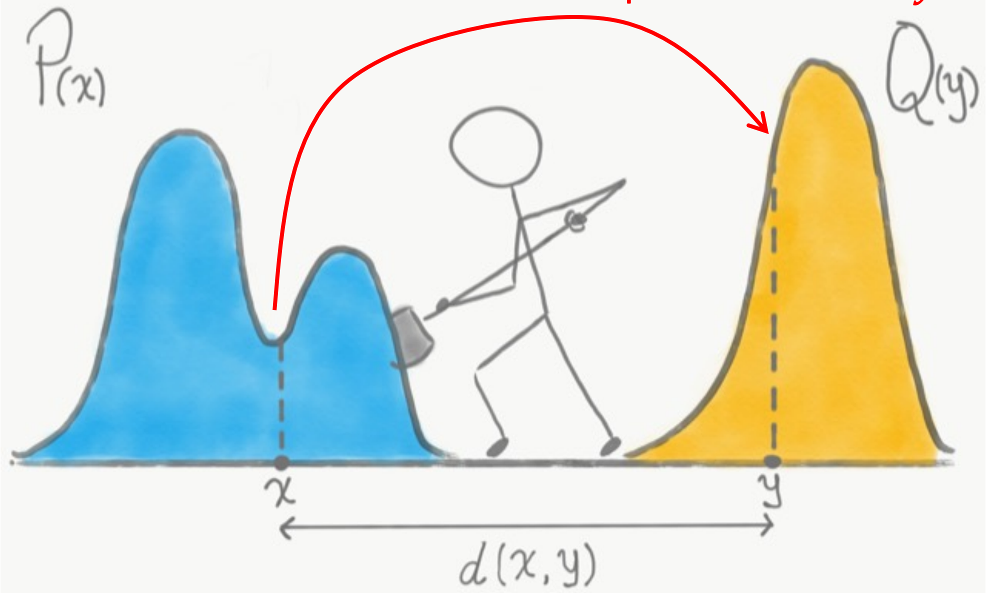
Preliminaries

- Optimal Transport
  - How to transport P(x) to Q(x) with minimum effort?

$$\min_{\boldsymbol{T}} \sum_{x \in \boldsymbol{P}, y \in \boldsymbol{Q}} \boldsymbol{T}(x,y) d(x,y)$$

$\boldsymbol{T}(x,y)$

how much mass to transport from x to y



Preliminaries

- Wasserstein distance
  - Given:
    - Two distribution density $\boldsymbol{P_1} \in \mathbb{R}^{n_1}, \boldsymbol{P_2} \in \mathbb{R}^{n_2}$
    - Attribute matrices $\boldsymbol{X_1} \in \mathbb{R}^{n_1 \times d}, \boldsymbol{X_2} \in \mathbb{R}^{n_2 \times d}$
    - A cross-cost matrix $\boldsymbol{C} \in \mathbb{R}^{n_1 \times n_2}$ based on $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$
  - Output:
    - The p-Wasserstein distance between $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$:

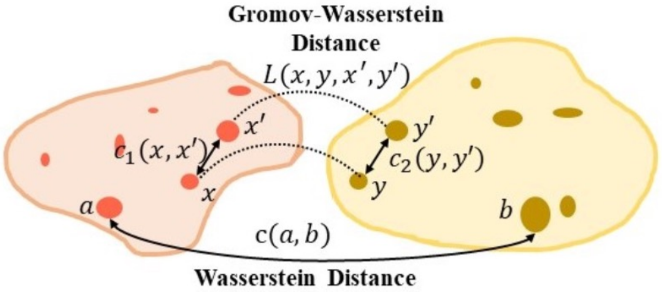
$$\mathrm{W}_p(\boldsymbol{P_1}, \boldsymbol{P_2}) = \left(\min_{\underline{\boldsymbol{T} \in \Pi(\boldsymbol{P_1}, \boldsymbol{P_2})}} \sum_{\substack{x \in \boldsymbol{P_1} \\ y \in \boldsymbol{P_2}}} \boldsymbol{C}^p(x, y) \boldsymbol{T}(x, y)\right)^{1/p}$$

$$\boxed{\begin{cases} \sum_{x \in \boldsymbol{P_1}} \boldsymbol{T}(x, y) = \boldsymbol{P_2} \\ \sum_{y \in \boldsymbol{P_2}} \boldsymbol{T}(x, y) = \boldsymbol{P_1} \end{cases}}$$

Minimum effort of transporting $\boldsymbol{P_1}$ to $\boldsymbol{P_2}$ in terms of distance between samples

[1] Peyré, Gabriel, and Marco Cuturi. "Computational Optimal Transport." *arXiv preprint arXiv:1803.00567* (2018).

Preliminaries



- Gromov-Wasserstein distance
  - Given:
    - Two distribution density $\boldsymbol{P_1} \in \mathbb{R}^{n_1}, \boldsymbol{P_2} \in \mathbb{R}^{n_2}$
    - Two intra-cost matrices $\boldsymbol{C_{P_1}} \in \mathbb{R}^{n_1 \times n_1}, \boldsymbol{C_{P_2}} \in \mathbb{R}^{n_2 \times n_2}$
  - Output:
    - The p-GW distance between $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$:

$$\mathrm{GW}_p(\boldsymbol{P_1}, \boldsymbol{P_2}) = \min_{\boldsymbol{T} \in \Pi(\boldsymbol{P_1}, \boldsymbol{P_2})} \left(\sum_{\substack{x_i, x_j \in \boldsymbol{P_1} \\ y_l, y_m \in \boldsymbol{P_2}}} \boxed{\left(\boldsymbol{C_{P_1}}(x_i, x_j) - \boldsymbol{C_{P_2}}(y_l, y_m)\right)^p} \boldsymbol{T}(x_i, y_l) \boldsymbol{T}(x_j, y_m) \right)^{1/p}$$

distance between sample pairs;
Similar pairwise relation across distributions

Minimum effort of transporting $\boldsymbol{P_1}$ to $\boldsymbol{P_2}$ in terms of distance between sample pairs

[2] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.
[3] Chen, Liqun, et al. "Graph optimal transport for cross-domain alignment." International Conference on Machine Learning. PMLR, 2020.

Preliminaries

- Fused Gromov-Wasserstein (FGW) distance
  - Linear combination of W_p and GW_p

$$\mathrm{FGW}_p(\boldsymbol{P_1}, \boldsymbol{P_2}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{P_1},\boldsymbol{P_2})} \left[\sum_{\substack{x_i,x_j\in\boldsymbol{P_1} \\ y_l,y_m\in\boldsymbol{P_2}}} \left((1-\alpha)\boldsymbol{C}^p(x_i,y_l) + \alpha\left(\boldsymbol{C}_{\boldsymbol{P_1}}(x_i,x_j) - \boldsymbol{C}_{\boldsymbol{P_2}}(y_l,y_m)\right)^p \right) \boldsymbol{T}(x_i,y_l)\boldsymbol{T}(x_j,y_m) \right]^{1/p}$$

  - Special case: L_2 norm as cross-cost, p=2
    - $\mathrm{W}_2^2(\boldsymbol{P_1}, \boldsymbol{P_2}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{P_1},\boldsymbol{P_2})} \langle \boldsymbol{C}, \boldsymbol{T} \rangle; \boldsymbol{C}(x,y) = \left|\left|\boldsymbol{X_1}(x) - \boldsymbol{X_2}(y)\right|\right|_2^2$
    - $\mathrm{GW}_2^2(\boldsymbol{P_1}, \boldsymbol{P_2}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{P_1},\boldsymbol{P_2})} \langle \boldsymbol{L}, \boldsymbol{T} \rangle; \boldsymbol{L} = \boldsymbol{C}_{\boldsymbol{P_1}}^2 \boldsymbol{P_1} \boldsymbol{1}_{n_2}^T + \boldsymbol{1}_{n_1} \boldsymbol{P}_{\boldsymbol{2}}^T \boldsymbol{C}_{\boldsymbol{P_2}}^{2^T} - 2\boldsymbol{C}_{\boldsymbol{P_1}} \boldsymbol{T} \boldsymbol{C}_{\boldsymbol{P_2}}^T$
    - $\mathrm{FGW}_2^2(\boldsymbol{P_1}, \boldsymbol{P_2}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{P_1},\boldsymbol{P_2})} \langle (1-\alpha)\boldsymbol{C} + \alpha\boldsymbol{L}, \boldsymbol{T} \rangle$

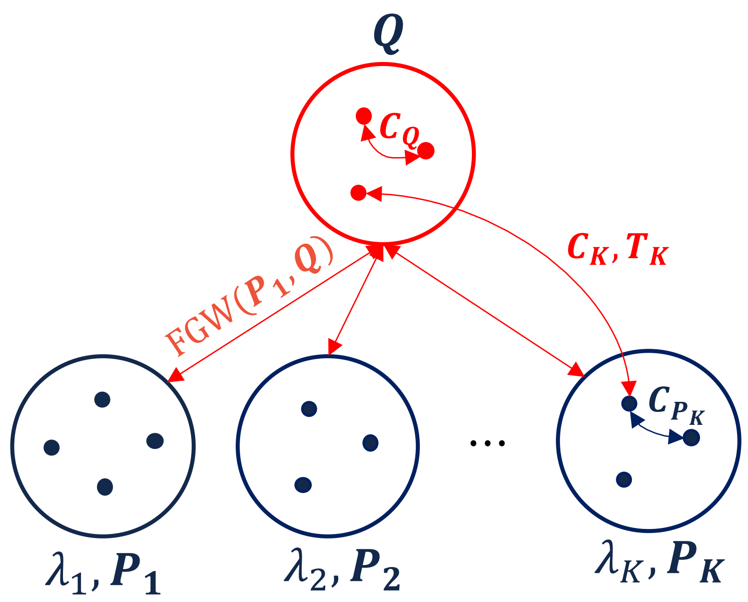
[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

Outline

- Motivation
- Preliminaries
- Fused Gromov-Wasserstein barycenter
  - Problem definition
  - Optimization
- Experiments
- Takeaways

FGW barycenter

- A distribution close to given distributions in terms of FGW distance
- Given:
  - distribution density $\boldsymbol{P}_1, \dots, \boldsymbol{P}_K$
  - weight for each distribution $\lambda_1, \dots, \lambda_K$
  - intra-cost matrices $\boldsymbol{C}_{\boldsymbol{P}_1}, \dots, \boldsymbol{C}_{\boldsymbol{P}_K}$
  - Barycenter density $\boldsymbol{Q} \in \mathbb{R}^m$
- Output:
  - Intra-cost matrix $\boldsymbol{C}_{\boldsymbol{Q}} \in \mathbb{R}^{m \times m}$
  - Attribute matrix $\boldsymbol{X}_{\boldsymbol{Q}} \in \mathbb{R}^{m \times d}$



$$\operatorname{argmin}_{\boldsymbol{C}_{\boldsymbol{Q}}, \boldsymbol{X}_{\boldsymbol{Q}}} \sum_{k=1}^{K} \lambda_k \text{FGW}(\boldsymbol{P}_k, \boldsymbol{Q})$$

$$= \operatorname{argmin}_{\boldsymbol{C}_{\boldsymbol{Q}}, \boldsymbol{X}_{\boldsymbol{Q}}} \sum_{k=1}^{K} \lambda_k \min_{\boldsymbol{T}_k \in \Pi(\boldsymbol{P}_k, \boldsymbol{Q})} \langle (1-\alpha) \boldsymbol{C}_k + \alpha \boldsymbol{L}_k, \boldsymbol{T}_k \rangle$$

$$\text{where} \begin{cases} \boldsymbol{C}_k(x, y) = \left\| \boldsymbol{X}_{\boldsymbol{P}_k}(x, :) - \boldsymbol{X}_{\boldsymbol{Q}}(y, :) \right\|_2 \\ \boldsymbol{L}_k = \boldsymbol{C}_{\boldsymbol{P}_k}^2 \boldsymbol{P}_k \mathbf{1}_m^T + \mathbf{1}_{n_k} \boldsymbol{Q}^T \boldsymbol{C}_{\boldsymbol{Q}}^{2^T} - 2 \boldsymbol{C}_{\boldsymbol{P}_k} \boldsymbol{T}_k \boldsymbol{C}_{\boldsymbol{Q}}^T \end{cases}$$

[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

FGW barycenter

$$\text{argmin}_{C_Q, X_Q} \sum_{k=1}^{K} \lambda_k \min_{T_k \in \Pi(P_K, Q)} \langle (1-\alpha) C_k + \alpha L_k, T_k \rangle$$

$$\text{where} \begin{cases} C_k(x,y) = \|X_{P_k}(x,:) - X_Q(y,:)\|_2 \\ L_k = C_{P_k}^2 P_k 1_m^T + 1_{n_k} Q^T C_Q^{2^T} - 2 C_{P_k} T_k C_Q^T \end{cases}$$

- Block Coordinate Descent
  - Iteratively minimize w.r.t. T_k, X_Q, C_Q
- Fix $X_Q^{(t-1)}, C_Q^{(t-1)}$; Optimize $T_k^{(t)}$
  - *Observation: $T_1^{(t)}, \ldots, T_K^{(t)}$ are decoupled*
  - Solve K problems respectively

$$T_k^{(t)} = \underset{T \in \Pi(P_k, Q)}{\text{argmin}} \left\langle (1-\alpha) C_k^{(t-1)} + \alpha L_k^{(t-1)}, T \right\rangle$$

  - Sinkhorn algorithm for solution
    - Entropy regularization → strict convexity
    - Iterative matrix scaling → coupling constraint

[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.
[5] Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." Advances in neural information processing systems 26 (2013).

FGW barycenter

$$\text{argmin}_{C_Q,X_Q}\sum_{k=1}^{K}\lambda_k \min_{T_k\in\Pi(P_K,Q)}\langle(1-\alpha)C_k+\alpha L_k, T_k\rangle$$

$$\text{where}\begin{cases} C_k(x,y)=\|X_{P_k}(x,:)-X_Q(y,:)\|_2 \\ L_k = C_{P_k}^2 P_k 1_m^T + 1_{n_k} Q^T C_Q^{2^T} - 2C_{P_k} T_k C_Q^T \end{cases}$$

- Fix $T_k^{(t)}, X_Q^{(t-1)}$; Optimize $C_Q^{(t)}$

$$C_Q^{(t)} = \underset{C_Q}{\text{argmin}} \sum_{k=1}^{K}\lambda_k \left\langle 1_{n_k} Q^T C_Q^{2^T} - 2C_{P_k} T_k^{(t)} C_Q^T, T_k^{(t)}\right\rangle$$

– First-order optimality

$$C_Q^{(t)} = \frac{\sum_{k=1}^{K}\lambda_k \boxed{T_k^{(t)^T} C_{P_k} T_k^{(t)}}}{QQ^T}$$

Average of C_{P_k} based on λ_k and $T_k^{(t)}$

$$\underline{\left(T_k^{(t)^T} \otimes T_k^{(t)^T}\right)} \text{vec}(C_{P_k})$$

row (i,j), column (l,m): $T_k^{(t)}(l,i)\, T_k^{(t)}(m,j)$
Similarity between intra-relation $(l,m) \in P_k$ and $(i,j) \in Q$

[2] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.
[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

FGW barycenter

$$\mathrm{argmin}_{C_Q, X_Q} \sum_{k=1}^{K} \lambda_k \min_{T_k \in \Pi(P_K, Q)} \langle (1-\alpha) C_k + \alpha L_k, T_k \rangle$$

$$\text{where} \begin{cases} C_k(x,y) = \left\| X_{P_k}(x,:) - X_Q(y,:) \right\|_2 \\ L_k = C_{P_k}^2 P_k 1_m^T + 1_{n_k} Q^T C_Q^{2^T} - 2 C_{P_k} T_k C_Q^T \end{cases}$$

- Fix $T_k^{(t)}, C_Q^{(t)}$; Optimize $X_Q^{(t)}$

$$X_Q^{(t)} = \mathrm{argmin}_{X_Q} \sum_{k=1}^{K} \lambda_k \left\langle C_k^{(t)}, T_k^{(t)} \right\rangle$$

$$\text{where } C_k^{(t)} = \underbrace{\boxed{\mathrm{diag}(X_k X_k^T) 1_m^T}}_{\text{constant}} + \underbrace{\boxed{1_{n_k}}}_{1_{n_k} T_k = Q} \mathrm{diag}\left(X_Q^{(t)} X_Q^{(t)^T}\right) - 2 X_k X_Q^{(t)^T}$$

$$X_Q^{(t)} = \mathrm{argmin}_{X_Q} \sum_{k=1}^{K} \lambda_k \left\| \mathrm{diag}\left(Q^{\frac{1}{2}}\right) X_Q - \mathrm{diag}\left(Q^{-\frac{1}{2}}\right) T_k^{(t)^T} X_k \right\|^2$$

$$X_Q^{(t)} = \mathrm{diag}(Q^{-1}) \boxed{\sum_{k=1}^{K} \lambda_k T_k^{(t)^T} X_k}$$

Average of node attribute X_k based on λ_k and $T_k^{(t)}$

[2] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.
[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

FGW barycenter

Algorithm 1 FGW barycenter

Input distributions $\boldsymbol{P}_1, \dots, \boldsymbol{P}_K$; weight $\lambda_1, \dots, \lambda_K$; intra-cost matrices $\mathbf{C}_{\boldsymbol{P}_1}, \dots, \mathbf{C}_{\boldsymbol{P}_K}$; attribute matrices $\mathbf{X}_1, \dots, \mathbf{X}_K$; barycenter distribution $\boldsymbol{Q}$

Output intra-cost matrix $\mathbf{C}_{\boldsymbol{Q}}$; attribute matrix $\mathbf{X}_{\boldsymbol{Q}}$

1: Initialize $\mathbf{T}_k^{(0)}, \mathbf{C}_{\boldsymbol{Q}}^{(0)}, \mathbf{X}_{\boldsymbol{Q}}^{(0)}$;

2: **for** $t = 1, 2, \dots$ **do**

3: Update $\mathbf{T}_k^{(t)}$ based on $\mathbf{C}_{\boldsymbol{Q}}^{(t-1)}, \mathbf{X}_{\boldsymbol{Q}}^{(t-1)}$ via Sinkhorn algorithm.

4: Update intra-cost $\mathbf{C}_{\boldsymbol{Q}}^{(t)} = \frac{\sum_{k=1}^{K} \lambda_k {\mathbf{T}_k^{(t)}}^T \mathbf{C}_{\boldsymbol{P}_k} \mathbf{T}_k^{(t)}}{\mathbf{Q}\mathbf{Q}^T}$.

5: Update attribute matrix $\mathbf{X}_{\boldsymbol{Q}}^{(t)} = \mathrm{diag}(\boldsymbol{Q}^{-1}) \sum_{k=1}^{K} \lambda_k {\mathbf{T}_k^{(t)}}^T \mathbf{X}_k$.

6: **end for**

7: **return** $\mathbf{T}_k, \mathbf{C}_{\boldsymbol{Q}}, \mathbf{X}_{\boldsymbol{Q}}$.

---

[2] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.

[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

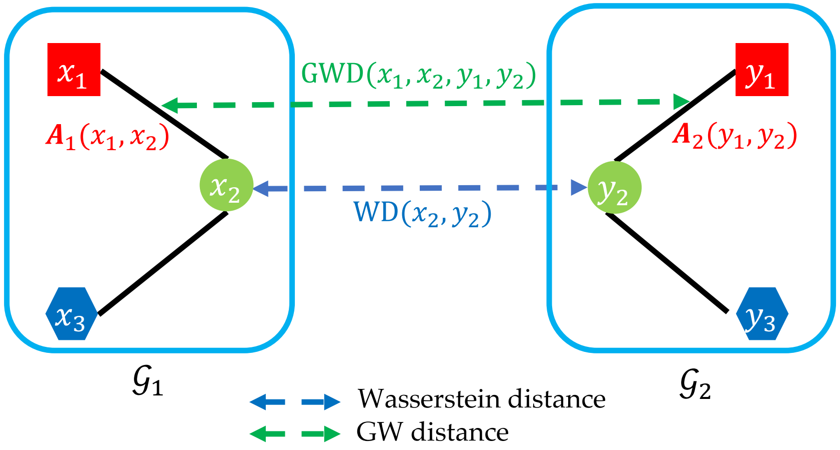
[5] Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." Advances in neural information processing systems 26 (2013).

Outline

- Motivation
- Preliminaries
- Fused Gromov-Wasserstein barycenter
- Experiments
- Takeaways

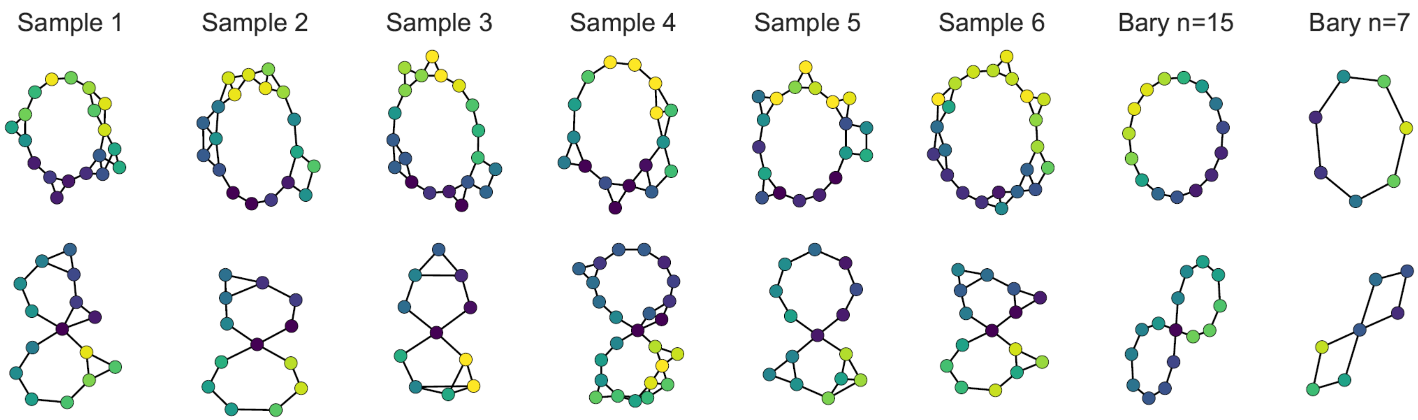
Experiments

- FGW barycenter for graphs
  - $G_1 = \{V_1, \boldsymbol{A}_1, \boldsymbol{X}_1\}, G_2 = \{V_2, \boldsymbol{A}_2, X_2\}$
  - $\boldsymbol{P}_1 = \frac{\mathbf{1}_{|V_1|}}{|V_1|}, \boldsymbol{P}_2 = \frac{\mathbf{1}_{|V_2|}}{|V_2|}$
  - Cross-cost $\boldsymbol{C}$: L_2 norm between $\boldsymbol{X}_1, \boldsymbol{X}_2$
  - Intra-cost $\boldsymbol{C}_{\boldsymbol{P}_1}, \boldsymbol{C}_{\boldsymbol{P}_2}$: adjacency matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$
  - WD: node relation; attribute
  - GWD: edge relation; structure



Experiments

- Graph barycenter



- Graph classification (FGW distance as SVM kernel)

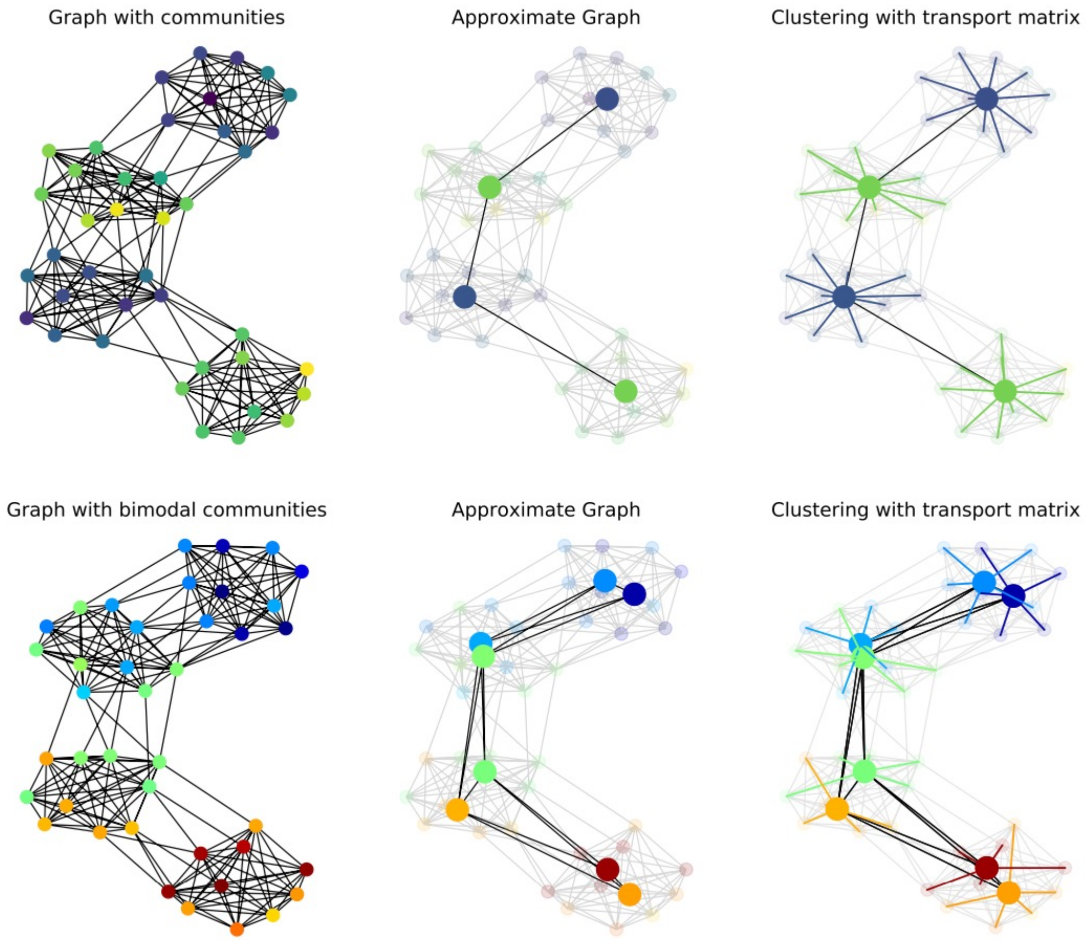
Table 1. Average classification accuracy on the graph datasets with vector attributes.

Vector Attributes	BZR	COX2	CUNEIFORM	ENZYMES	PROTEIN	SYNTHETIC
FGW sp	**85.12 ± 4.15** *	77.23 ± 4.86	**76.67 ± 7.04**	71.00 ± 6.76	**74.55 ± 2.74**	**100.00 ± 0.00**
HOPPERK	84.15 ± 5.26	**79.57 ± 3.46**	32.59 ± 8.73	45.33 ± 4.00	71.96 ± 3.22	90.67 ± 4.67
PROPAK	79.51 ± 5.02	77.66 ± 3.95	12.59 ± 6.67	**71.67 ± 5.63** *	61.34 ± 4.38	64.67 ± 6.70
PSCN k = 10	80.00 ± 4.47	71.70 ± 3.57	25.19 ± 7.73	26.67 ± 4.77	67.95 ± 11.28	**100.00 ± 0.00**
PSCN k = 5	82.20 ± 4.23	71.91 ± 3.40	24.81 ± 7.23	27.33 ± 4.16	71.79 ± 3.39	**100.00 ± 0.00**

[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

Experiments

- Clustering



[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

Outline

- Motivation
- Preliminaries
- Fused Gromov-Wasserstein barycenter
- Experiments
- Takeaways

Takeaways

- Motivation
- Preliminaries
  - OT, WD, GWD, FGWD
- Fused Gromov-Wasserstein barycenter
  - Block coordinate descent
  - Optimize transport plan $\boldsymbol{T_k}$: solve K optimal transport problems
  - Optimize intra-cost $\boldsymbol{C_Q}$: Average of $\boldsymbol{C_k}$ based on λ_k and $\boldsymbol{T_k}$
  - Optimize attribute $\boldsymbol{X_Q}$: Average of $\boldsymbol{X_k}$ based on λ_k and $\boldsymbol{T_k}$
- Experiments
  - Graph barycenter
  - Graph classification
  - Graph clustering

References

[1] Peyré, Gabriel, and Marco Cuturi. "Computational Optimal Transport." arXiv preprint arXiv:1803.00567 (2018).

[2] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.

[3] Chen, Liqun, et al. "Graph optimal transport for cross-domain alignment." International Conference on Machine Learning. PMLR, 2020.

[4] Vayer, Titouan, et al. "Fused gromov-wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

[5] Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." Advances in neural information processing systems 26 (2013).

Thanks for listening!
Q&A

Zhichen Zeng

10/13