Stanford

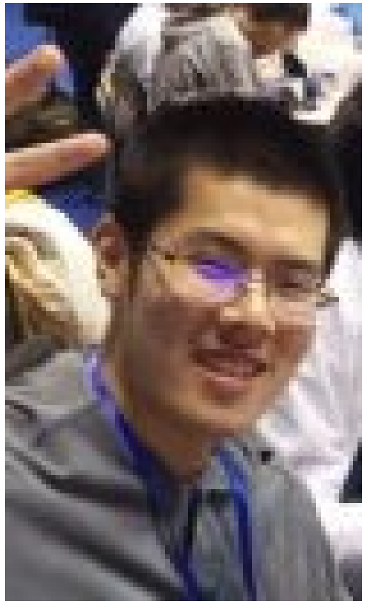
# Gromov-Wasserstein Distance Computation

## Infeasibility, Efficiency and Accuracy

**Jiajin Li**

SIAM Conference on Optimization (OP23)

# Collaborators

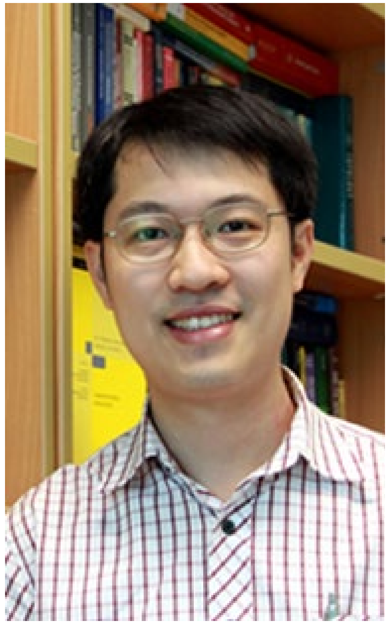


Jianheng Tang
[HKUST(GZ)]

Lemin Kong
[CUHK]

Huikang Liu
[SUFE]

Jia Li
[HKUST(GZ)]



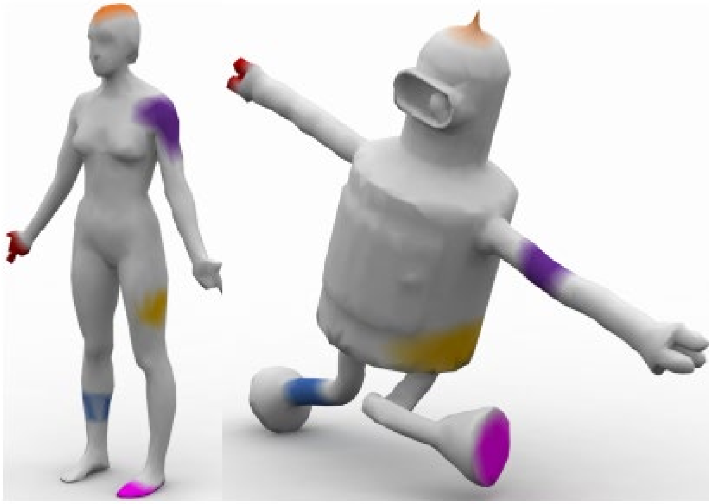
Man-Cho So
[CUHK]

Jose Blanchet
[Stanford]

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

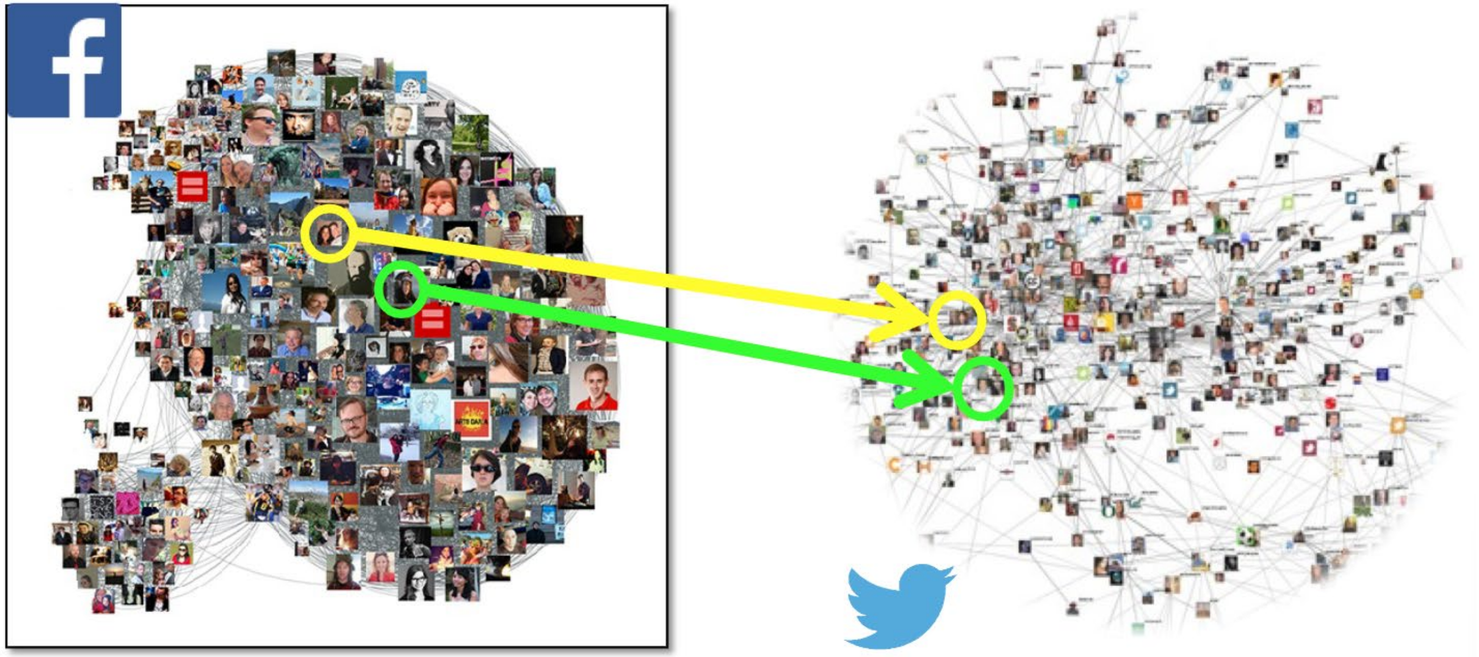
# Heterogeneous & Structured Data

**Dataset Matching:** Various applications require matching structured datasets.



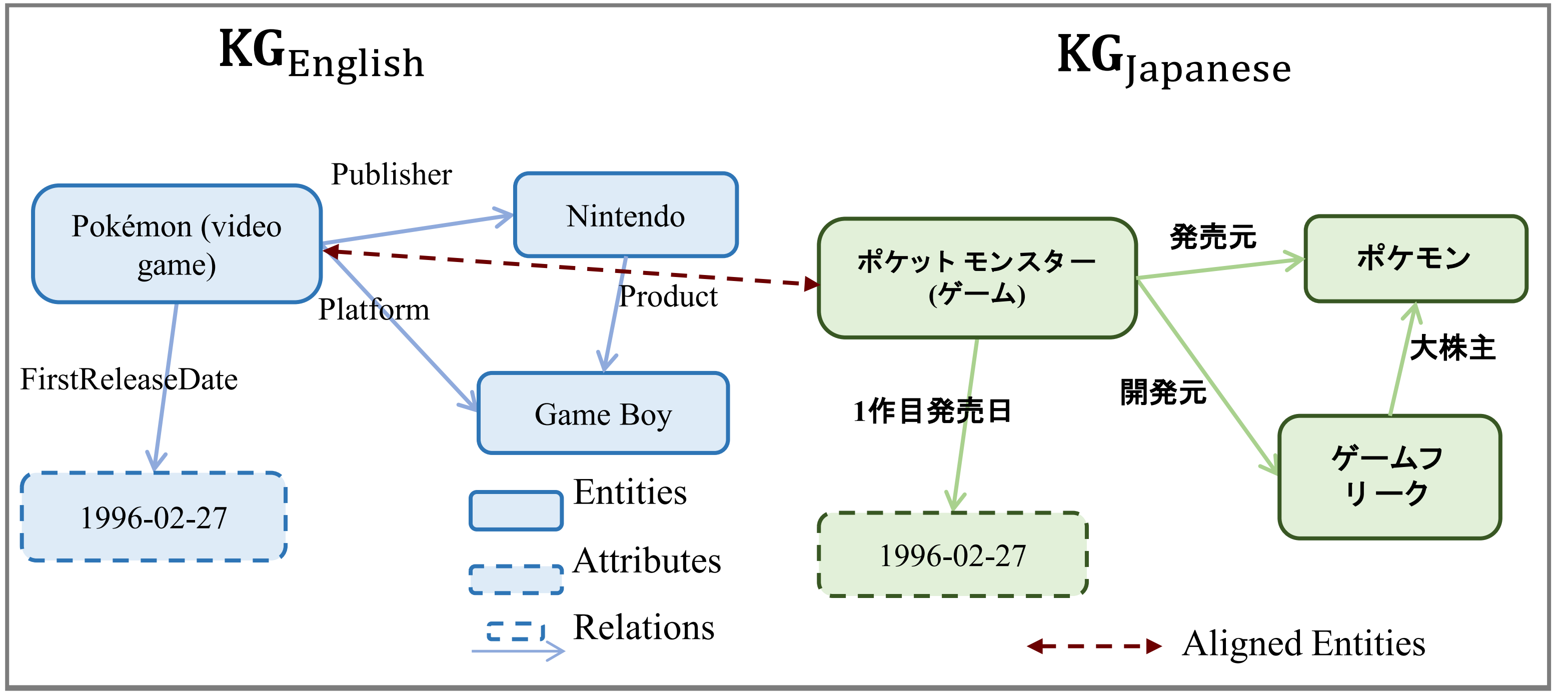
Shape Correspondence

Social Network Alignment



# Heterogeneous & Structured Data

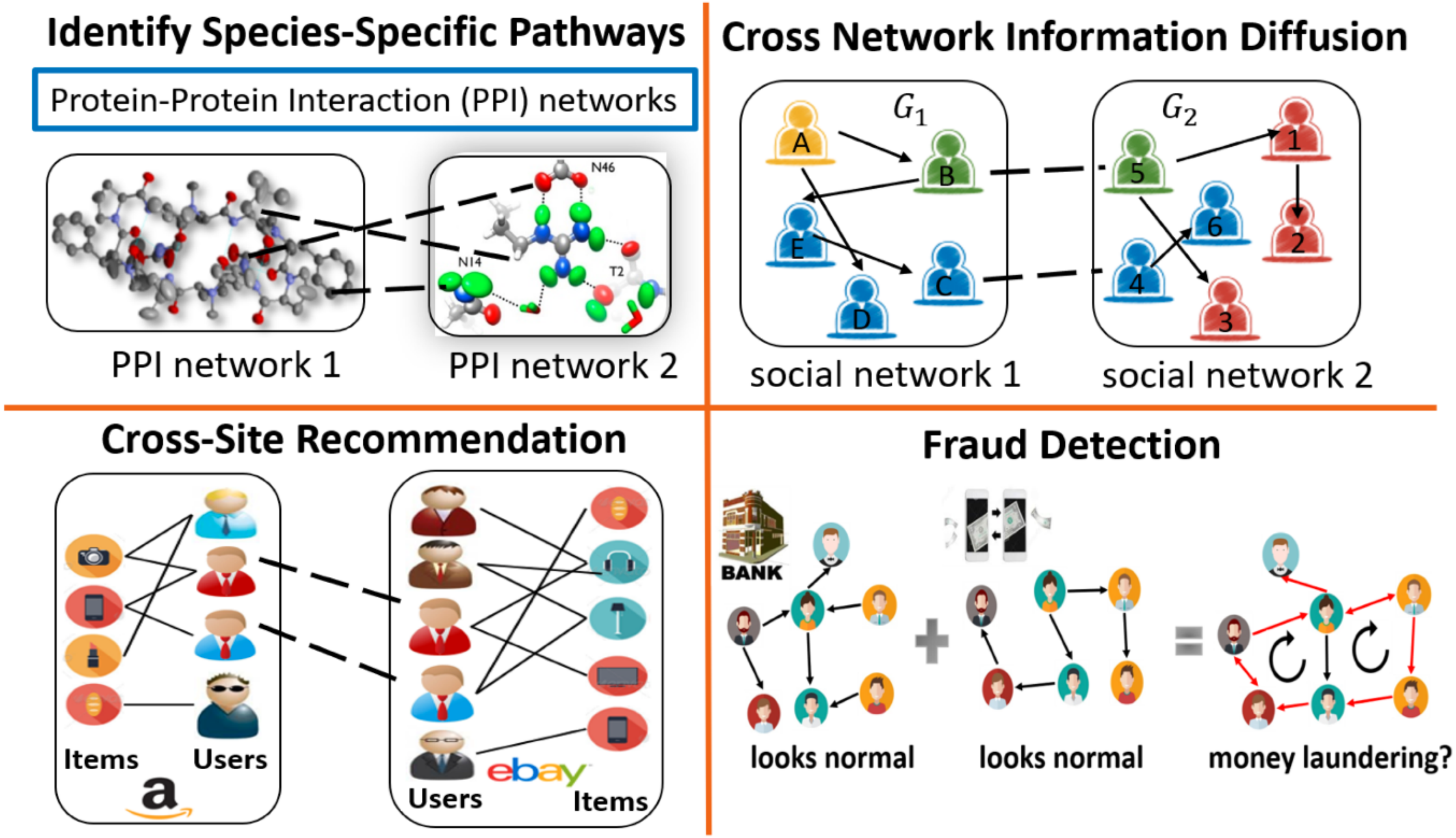
**Dataset Matching:** Various applications require matching structured datasets.




Cross-lingual Knowledge Graph Alignment

# Heterogeneous & Structured Data

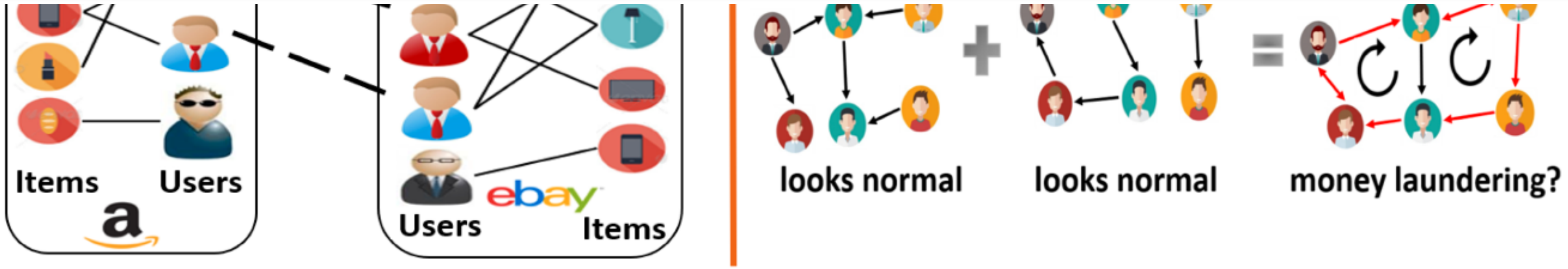
Other Applications



# Heterogeneous & Structured Data


**Goal:**

1. Compare how similar/different two datasets are?
2. Obtain alignment that preserves the geometric (graph) structure



# Gromov-Wasserstein Distance

**Definition(Memoli' 11):** The GW distance between two metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ is

$$\mathrm{GW}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \iint |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 d\pi(x, y) d\pi(x', y')$$

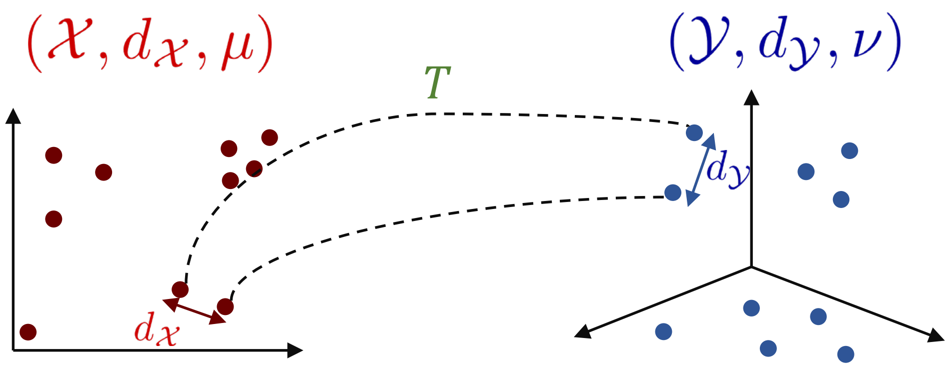
where $\Pi(\mu, \nu)$ is the set of probabilities whose marginals are $\mu \text{ and } \nu$.

# Gromov-Wasserstein Distance

Definition(Memoli' 11): The GW distance between two metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ is

$$\mathrm{GW}(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \iint |d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')|^2 d\pi(x,y) d\pi(x',y')$$

where $\Pi(\mu,\nu)$ is the set of probabilities whose marginals are $\mu$ and $\nu$.




- Finding the transport map T: $\mathcal{X} \to \mathcal{Y}$
- Preserve the **isometric** structure

$$d_{\mathcal{X}}(x,x') \approx d_{\mathcal{Y}}(T(y),T(y'))$$

# Computational Hardness

GW distance computation is a nonconvex quadratic problem.

$$\min_{\pi \in \mathbb{R}^{n \times m}} \; -\mathrm{Tr}(D_X \pi D_Y \pi^T)$$

$$\text{s.t.} \quad \boxed{\pi \mathbf{1}_m = \hat{\mu}_n, \; \pi^T \mathbf{1}_n = \hat{\nu}_m, \; \pi \geq 0.}$$

Birkhoff Polytope (Doubly Stochastic Matrix).

# Computational Hardness

GW distance computation is a nonconvex quadratic problem.

$$\min_{\pi \in \mathbb{R}^{n\times m}} -\mathrm{Tr}(D_X \pi D_Y \pi^T)$$

$$\text{s.t.} \quad \boxed{\pi \mathbf{1}_m = \hat{\mu}_n,\ \pi^T \mathbf{1}_n = \hat{\nu}_m,\ \pi \geq 0.}$$

Replace by permutation matrix -> Quadratic Assignment Problem (QAP)

**NP Complete [Commander '05]**

# Computational Hardness

GW distance computation is a nonconvex quadratic problem.

$$\min_{\pi \in \mathbb{R}^{n \times m}} \; -\mathrm{Tr}(D_X \pi D_Y \pi^T)$$

$$\text{s.t.} \quad \boxed{\pi \mathbf{1}_m = \hat{\mu}_n,\ \pi^T \mathbf{1}_n = \hat{\nu}_m,\ \pi \geq 0.}$$

Converge to local minima?

# Limitations

The main bottleneck: projecting onto the Birkhoff polytope.

# Limitations

The main bottleneck: projecting onto the Birkhoff polytope.

1. Double-Loop Iterative Scheme
2. Approximation (Relaxation)
3. Without any Theoretical Guarantee
4. Overlook Task Information

# Limitations

The main bottleneck: projecting onto the Birkhoff polytope.

☹

1. Double-Loop Iterative Scheme
2. Approximation (Relaxation)
3. Without any Theoretical Guarantee
4. Overlook Task Information

Entropic Regularization (**eBPG**)

$$\min_{\pi \in \mathbb{R}^{n \times m}} -\mathrm{Tr}(D_X \pi D_Y \pi^T) + \epsilon D_{\mathrm{KL}}(\pi \mid \hat{\mu}_n \otimes \hat{\nu}_m)$$

$$\text{s.t.} \quad \pi \mathbf{1}_m = \hat{\mu}_n, \; \pi^T \mathbf{1}_n = \hat{\nu}_m, \; \pi \geq 0.$$

- Sinkhorn -> subroutine
- Sensitive to hyperparameters
- Result in suboptimal performance on shape correspondence.

Solomon J, Peyré G, Kim V G, et al. Entropic metric alignment for correspondence problems. TOG, 2016.

# Limitations

The main bottleneck: projecting onto the Birkhoff polytope.

☹

1. Double-Loop Iterative Scheme
2. Approximation (Relaxation)
3. Without any Theoretical Guarantee
4. Overlook Task Information

Bregman Projected Gradient Descent (**BPG**)

$$\min_{\pi \in \mathbb{R}^{n \times m}} -\mathrm{Tr}(D_X \pi_k D_Y \pi^T) + \eta D_{\mathrm{KL}}(\pi \mid \pi_k)$$

$$\text{s.t.} \quad \pi \mathbf{1}_m = \hat{\mu}_n, \, \pi^T \mathbf{1}_n = \hat{\nu}_m, \, \pi \geq 0.$$

- Sinkhorn -> subroutine
- Result in suboptimal performance on graph alignment/partition.

Xu H, Luo D, Zha H, et al. Gromov-wasserstein learning for graph matching and node embedding, ICML 2019.

# Limitations

The main bottleneck: projecting onto the Birkhoff polytope.

☹

1. Double-Loop Iterative Scheme
2. Approximation (Relaxation)
3. Without any Theoretical Guarantee
4. Overlook Task Information

**(Heuristic) BPG-S**

$$\min_{\pi \in \mathbb{R}^{n \times m}} -\mathrm{Tr}(D_X \pi_k D_Y \pi^T) + \eta D_{\mathrm{KL}}(\pi \mid \pi_k)$$
$$\text{s.t.} \quad \pi \mathbf{1}_m = \hat{\mu}_n,\ \pi^T \mathbf{1}_n = \hat{\nu}_m,\ \pi \geq 0.$$

- Sinkhorn -> subroutine **(only one step)**
- Result in suboptimal performance on graph alignment/partition.

Xu H, Luo D, Carin L. Scalable gromov-wasserstein learning for graph partitioning and matching. NeurIPS, 2019.

# Limitations

**The main bottleneck: projecting onto the Birkhoff polytope.**

☹

1. Double-Loop Iterative Scheme
2. Approximation (Relaxation)
3. Without any Theoretical Guarantee
4. Overlook Task Information

Frank-Wolfe (**FW**)

$$\min_{\pi \in \mathbb{R}^{n\times m}} -\mathrm{Tr}(D_X \pi_k D_Y \pi^T)$$
$$\text{s.t.} \quad \pi \mathbf{1}_m = \hat{\mu}_n,\ \pi^T \mathbf{1}_n = \hat{\nu}_m,\ \pi \geq 0.$$

- Linear programming-> subroutine
- Line search
- Medium size datasets?

Vayer, Titouan, et al. Optimal Transport for structured data with application on graphs. ICML, 2019.

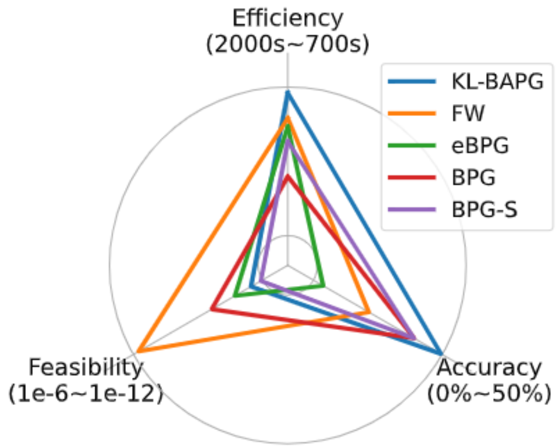
# Our Main Contribution

Develop **two** theoretically solid algorithms **tailored** to different graph learning tasks.

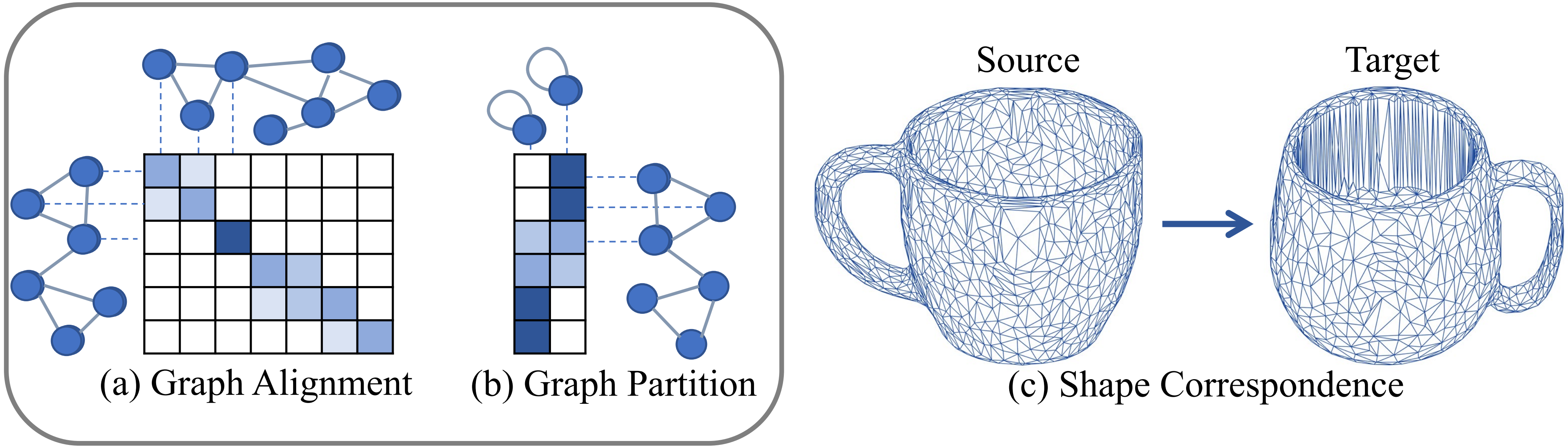
1. Bregman Alternating Projected Gradient (**BAPG**)

Novel Relaxation

- ✓ **Single-loop** algorithm
- ✓ Compatibility with **GPU** implementation,
- ✓ **Robustness** to the step size (the only hyperparameter)
- ✓ Low **memory** cost
- ✗ **Infeasible** method

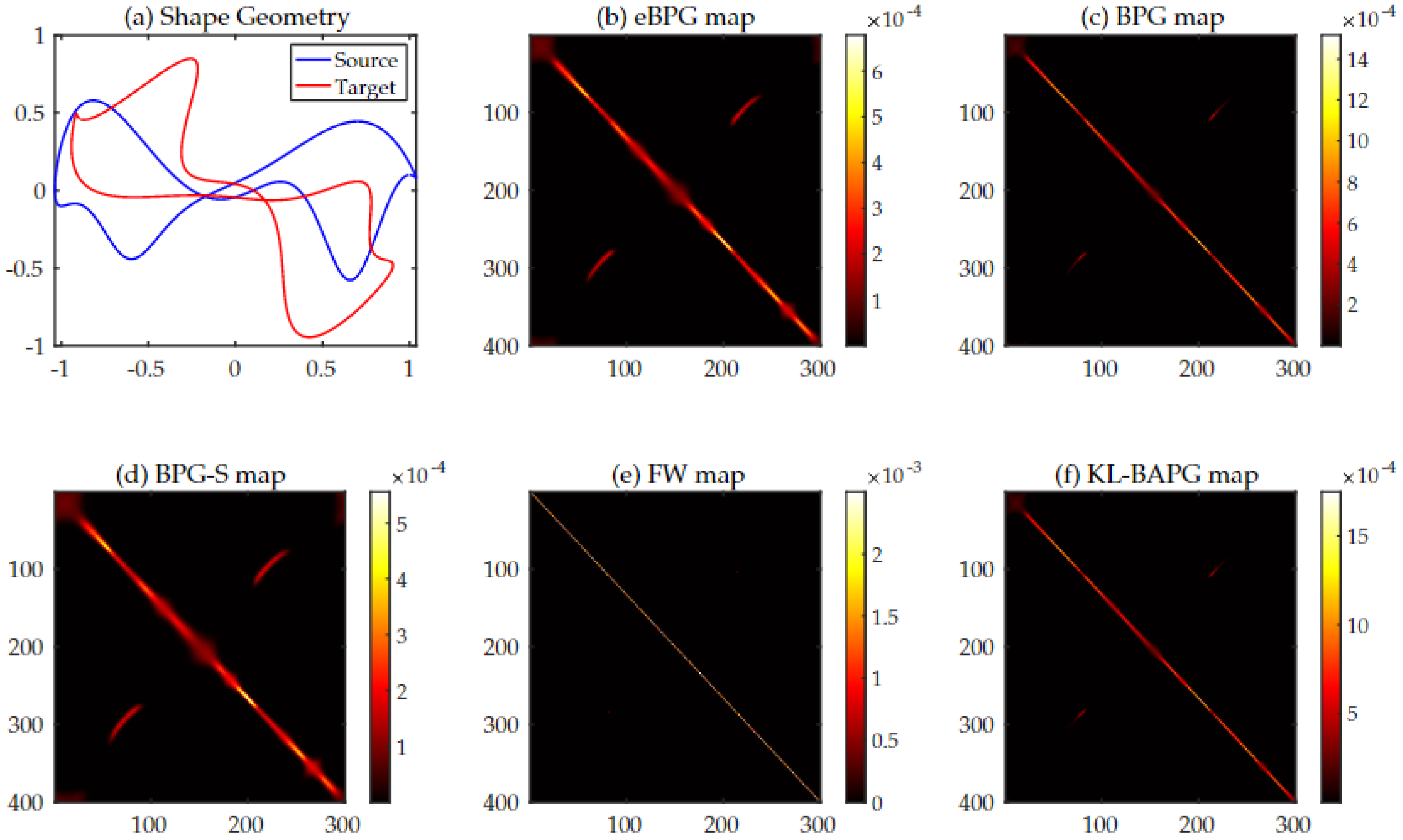


# BAPG




Sacrifice some **feasibility** to gain both **efficiency** and **accuracy** !

# 2D Toy Example



# Our Main Contribution

Develop **two** theoretically solid algorithms **tailored** to different graph learning tasks.

2. Hybrid Bregman Gradient Descent (**hBPG**)

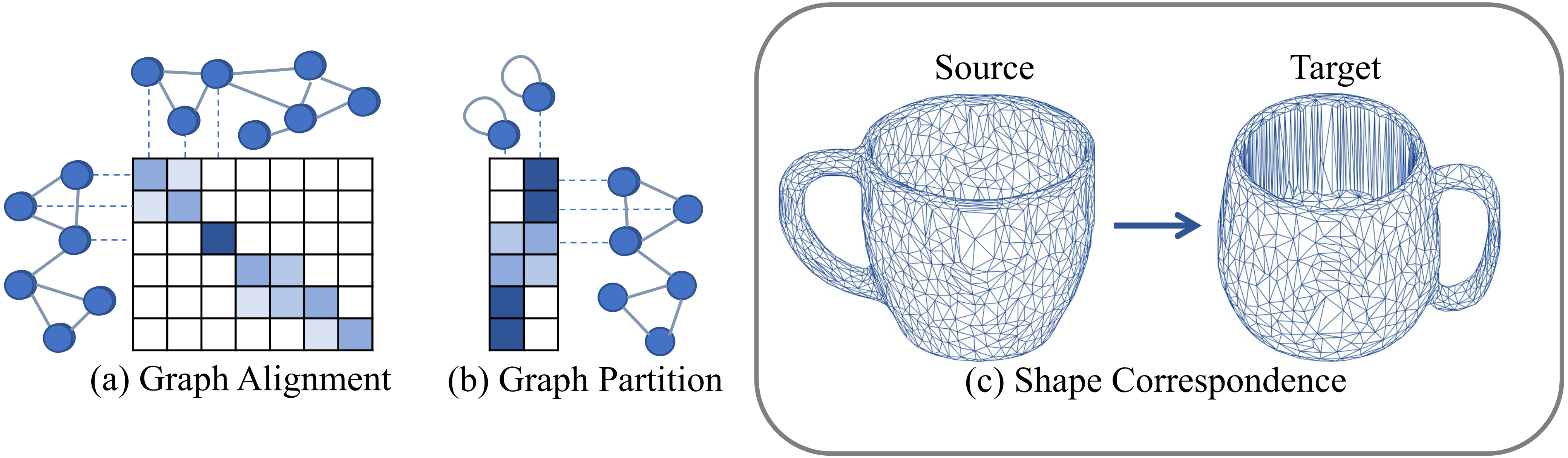
✓ **Feasible** method

✓ Take benefits from both BPG and eBPG

✓ Faster local convergence rate

X **Double-loop** algorithm

# Hybrid BPG




**Feasibility** is directly connected with task performance.

# Algorithm Design

# Bregman Alternating Gradient Descent

1. "**Operator Splitting**" strategy to decouple the Birkhoff polytope:

$$\begin{aligned} \min_{\pi, w \in \mathbb{R}^{n \times m}} \quad & -\mathrm{Tr}(D_X \pi D_Y w^T) \\ \text{s.t.} \quad & \pi \mathbf{1}_m = \hat{\mu}_n, \pi \geq 0 \\ & w^T \mathbf{1}_n = \hat{\nu}_m, w \geq 0 \\ & \pi = w. \end{aligned}$$

2. Alternating projected gradient descent on a new **penalized function**

$$F_\rho(\pi, w) := -\mathrm{Tr}(D_X \pi D_Y w^T) + \boxed{\rho} D_h(\pi, w)$$

Infeasibility

# KL-BAPG

Choosing h as the relative entropy, we have closed-form update:

$$\pi \leftarrow \pi \odot \exp(D_X \pi D_Y / \rho)$$
$$\pi \leftarrow \mathrm{diag}(\mu ./ \pi \mathbf{1}_m)\pi$$
$$\pi \leftarrow \pi \odot \exp(D_X \pi D_Y / \rho)$$
$$\pi \leftarrow \pi \mathrm{diag}(\nu ./ \pi^T \mathbf{1}_n).$$

✓ **Single-loop** algorithm

✓ Compatibility with **GPU** implementation,

✓ **Robustness** to the step size (the only hyperparameter)

✓ Low **memory** cost

*X* **Infeasible** method

# Any Theoretical Guarantee?

# Main Technical Tool

$$\min_{\pi \in \mathbb{R}^{n \times m}} \quad -\mathrm{Tr}(D_X \pi D_Y \pi^T)$$

$$\text{s.t.} \quad \boxed{\pi \mathbf{1}_m = \hat{\mu}_n, \pi \geq 0} \; C_1$$

$$\boxed{\pi^T \mathbf{1}_n = \hat{\nu}_m, \pi \geq 0.} \; C_2$$

Proposition (**Luo-Tseng Error Bound Condition**):

$$\mathrm{dist}(\pi, \mathcal{X}) \leq \|\pi - \mathrm{proj}_{C_1 \cap C_2}(\pi + D_X \pi D_Y)\|$$

where $\mathcal{X}$ is the critical point set.

Luo Z Q, Tseng P. Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM Journal on Optimization, 1992.

# Approximation Bound of BAPG

**Theorem**: If the point $(\pi^\star, w^\star)$ belongs to the fixed-point set of BAPG, then the infeasibility error satisfies

$$\|\pi^\star - w^\star\| \leq \frac{\tau_1}{\rho}$$

and

$$\mathrm{dist}\left(\frac{\pi^\star + w^\star}{2}, \mathcal{X}\right) \leq \frac{\tau_1}{\rho}.$$

Perturbation Analysis!

# Convergence Results for BAPG

**Theorem (Diminish Step Size)** If $\rho_k = \mathcal{O}(\sqrt{k})$ then the infeasibility error decays

$$\mathrm{dist}(\pi_k, C_1 \cap C_2) \le \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$

and

$$\min_{1 \le i \le k} \|\boxed{\nabla f_\lambda(\pi_i)}\| \le \mathcal{O}\left(\frac{1}{\sqrt{k}}\right).$$

Moreau envelope of $f(\pi) := -\mathrm{tr}(D_X \pi D_Y \pi^T) + \mathbf{I}_{C_1 \cap C_2}(\pi)$

# Convergence Results for hBPG

**Theorem (Local Linear Convergence Rate**) Suppose that the sequence $\{\pi_k\}_{k\geq 0}$ has **an element-wise lower bound $\epsilon$**, i.e., the sequence of solutions $\{\pi_k\}_{k\geq 0}$ generated by BPG converges R-linearly to an element in the critical point set.

I. Sufficient decrease property

$$F\left(\pi^{k+1}\right) - F\left(\pi^{k}\right) \leq -\kappa_1 \left\|\pi^{k+1} - \pi^{k}\right\|^2 .$$

II. Cost-to-Go estimate

$$F\left(\pi^{k+1}\right) - F\left(\pi^{\star}\right) \leq -\kappa_2 \left(\mathrm{dist}^2(\pi^k, \mathcal{X}) + \left\|\pi^{k+1} - \pi^{k}\right\|^2\right) .$$

III. Safyguard property

$$\|\pi^k - \mathrm{Proj}_{C_1 \cap C_2}(\pi^k - \nabla f(\pi^k))\| \leq -\kappa_3 \|\pi^{k+1} - \pi^k\|^2.$$

# Graph Alignment

Graph alignment aims to identify the node correspondence between two graphs possibly with different topology structures.

Table 2: Statistics of databases for graph alignment.

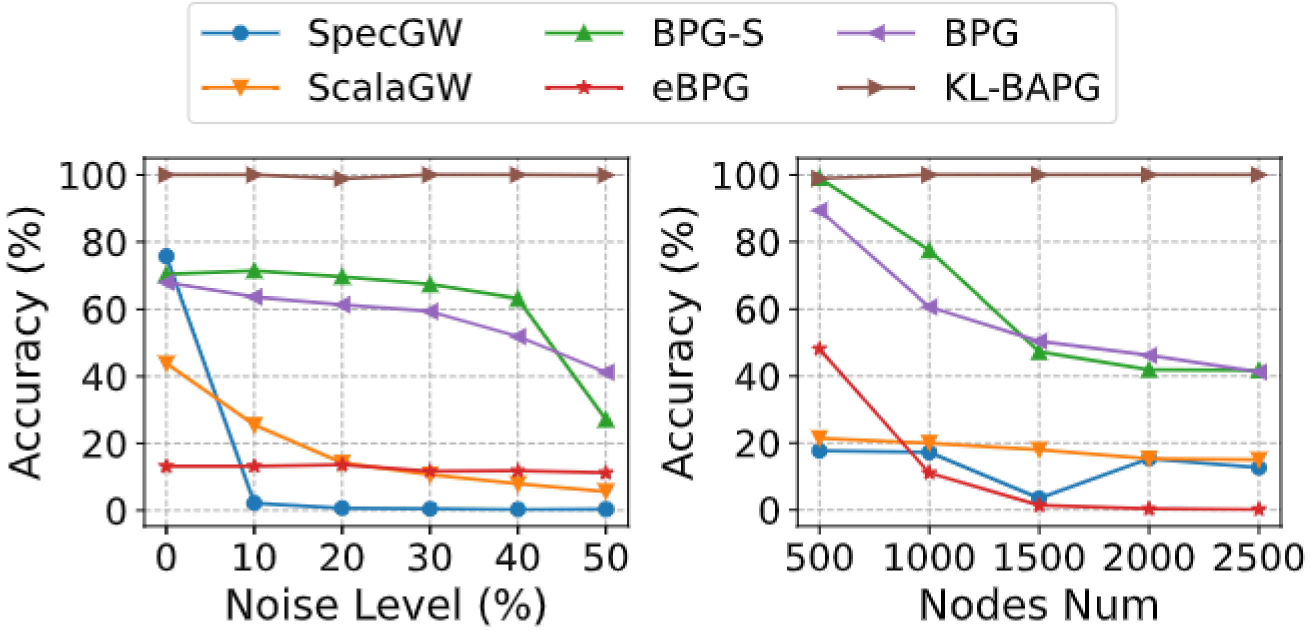
| Dataset | # Samples | Ave. Nodes | Ave. Edges |
|---|---|---|---|
| Synthetic | 300 | 1500 | 56579 |
| Proteins | 1113 | 39.06 | 72.82 |
| Enzymes | 600 | 32.63 | 62.14 |
| Reddit | 500 | 375.9 | 449.3 |

Table 3: Comparison of the matching accuracy (%) and wall-clock time (seconds) on graph alignment. For BAPG, we also report the time of GPU implementation.

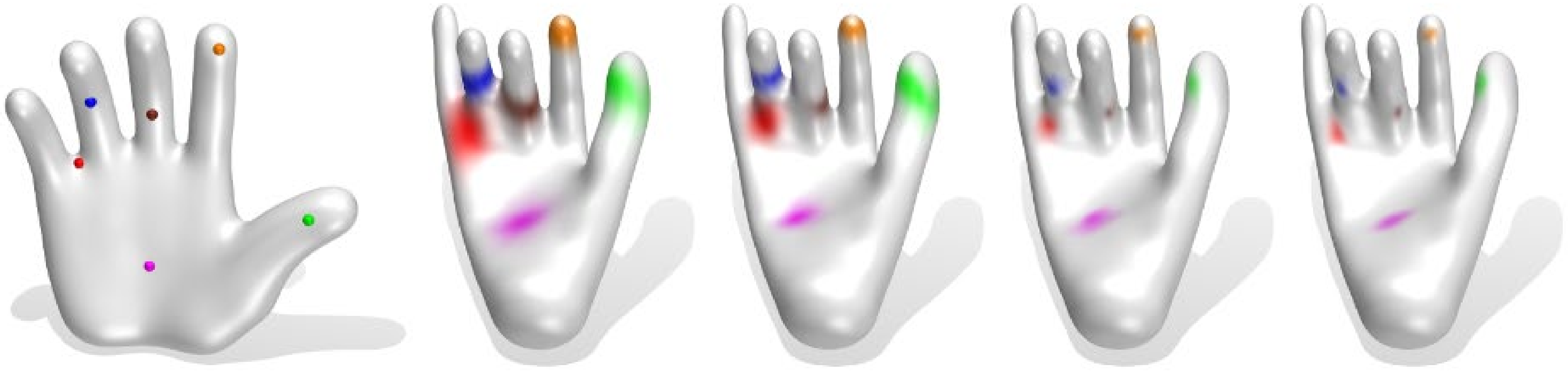
| Method | Synthetic | | Proteins | | | Enzymes | | | Reddit | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Time | Raw | Noisy | Time | Raw | Noisy | Time | Raw | Noisy | Time |
| IPFP | - | - | 43.84 | 29.89 | 87.0 | 40.37 | 27.39 | 23.7 | - | - | - |
| RRWM | - | - | 71.79 | 33.92 | 239.3 | 60.56 | 30.51 | 114.1 | - | - | - |
| SpecMethod | - | - | 72.40 | 22.92 | 40.5 | 71.43 | 21.39 | 9.6 | - | - | - |
| FW | 24.50 | 8182 | 29.96 | 20.24 | 54.2 | 32.17 | 22.80 | 10.8 | 21.51 | 17.17 | 1121 |
| ScalaGW | 17.93 | 12002 | 16.37 | 16.05 | 372.2 | 12.72 | 11.46 | 213.0 | 0.54 | 0.70 | 1109 |
| SpecGW | 13.27 | 1462 | 78.11 | 19.31 | **30.7** | 79.07 | 21.14 | **6.7** | 50.71 | 19.66 | 1074 |
| eBPG | 34.33 | 9502 | 67.48 | 45.85 | 208.2 | 78.25 | 60.46 | 499.7 | 3.76 | 3.34 | 1234 |
| BPG | 57.56 | 22600 | 71.99 | 52.46 | 130.4 | 79.19 | 62.32 | 73.1 | 39.04 | 36.68 | 1907 |
| BPG-S | 61.48 | 18587 | 71.74 | 52.74 | 40.4 | 79.25 | 62.21 | 13.4 | 39.04 | 36.68 | 1431 |
| hBPG | 51.57 | 13279 | 70.07 | 49.01 | 245.9 | 78.57 | 62.26 | 560.0 | 47.15 | 45.58 | 1447 |
| BAPG | **99.79** | 9024 | **78.18** | **57.16** | 59.1 | **79.66** | **62.85** | 14.8 | **50.93** | **49.45** | 780 |
| BAPG-GPU | - | **1253** | - | - | 75.4 | - | - | 21.8 | - | - | **115** |

# Graph Alignment

Graph alignment aims to identify the node correspondence between two graphs possibly with different topology structures.

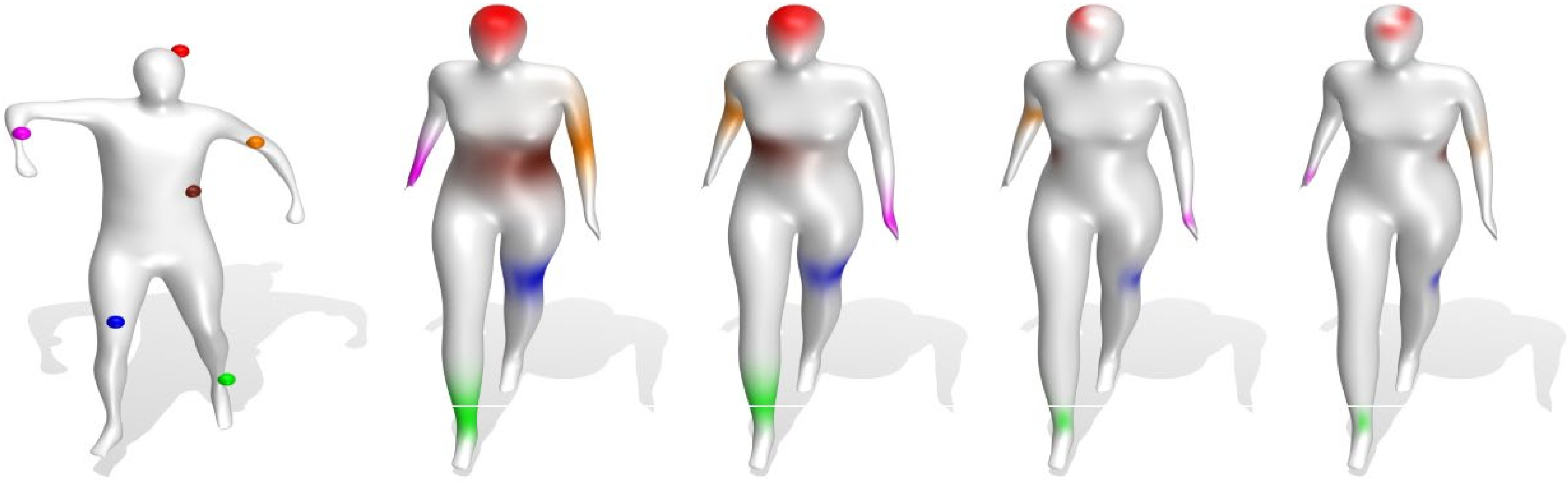


# Shape Correspondence



(a) Source Surface (b) eBPG (c) BAPG (d) BPG (e) hBPG

# Shape Correspondence



(u) Source Surface (v) eBPG (w) BAPG (x) BPG (y) hBPG

# Shape Correspondence

Table 5: Comparison of the infeasibility error (i.e., $\frac{\|\pi^T 1_n - \nu\|}{m} + \frac{\|\pi 1_m - \mu\|}{n}$ ) and CPU wall-clock time.

| Method | Hand | | Octopus | | Mug | | Chair | | Human | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Error | Time | Error | Time | Error | Time | Error | Time | Error | Time |
| eBPG | 2.95e-10 | 9.37 | 7.21e-10 | 12.08 | 6.41e-10 | 27.86 | 1.77e-10 | 12.87 | 5.52e-10 | 11.46 |
| BPG | 3.00e-07 | 389.72 | 2.00e-07 | 32.55 | 3.50e-07 | 196.93 | 4.00e-07 | 304.98 | 2.53e-07 | 93.27 |
| hBPG | 3.00e-07 | 193.85 | 2.00e-07 | 23.14 | 3.50e-07 | 90.59 | 4.00e-07 | 189.41 | 2.53e-07 | 53.29 |
| BAPG | 4.54e-06 | 61.77 | 2.14e-05 | 6.19 | 6.62e-05 | 30.83 | 2.13e-05 | 127.78 | 5.62e-05 | 8.26 |
| BAPG-GPU | - | 3.10 | - | 1.28 | - | 1.39 | - | 3.22 | - | 0.78 |

# Take Home Message

- **GW distance**: A powerful tool for aligning distinct structured datasets.
- When the sharpness of coupling does not matter, we can sacrifice some feasibility to gain both efficiency and accuracy – choose **BAPG**.
- When **coupling feasibility** affects task performance directly, consider using hBPG as an alternative.
- **Local error bound**: A useful technical tool for perturbation analysis to facilitate analysis.

# Reference


1. **Jiajin Li**, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, Jose Blanchet. *A Convergent Single-Loop Algorithm for Relaxation of Gromov-Wasserstein in Graph Data,* International Conference on Learning Representation (**ICLR**), 2023.

2. **Jiajin Li**, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, Jose Blanchet. *Fast Provably Convergent Algorithms for Gromov-Wasserstein in Graph Data,* *arXiv:2205.08115*

3. Jianheng Tang, Weiqi Zhang, **Jiajin Li**, Kangfei Zhao, Fugee Tsung, Jia Li. *Robust Attributed Graph Alignment via Joint Structure Learning and Optimal Transport,* International Conference on Data Engineering (**ICDE**), 2023.

4. He Chen, **Jiajin Li**, Anthony Man-Cho So. Random Projected Descent Method for Weakly Convex Functions. Working Paper.