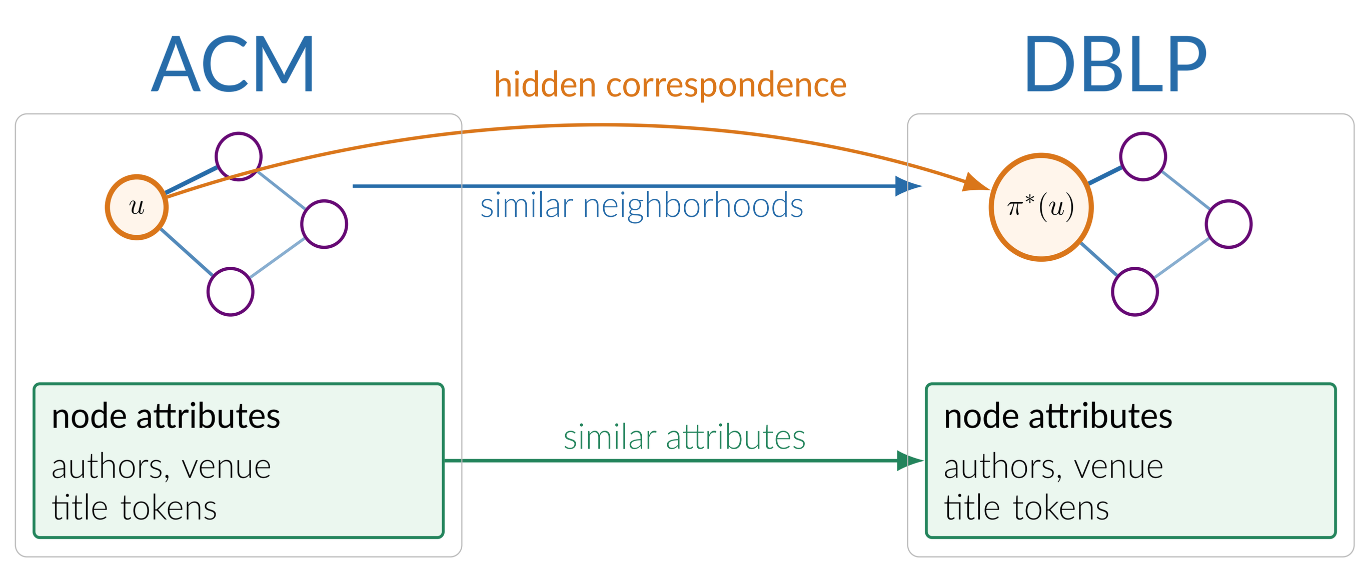
Attributed Network Alignment: Statistical Limits and Efficient Algorithm

Dong Huang   Chenyang Tian   Pengkun Yang

Tsinghua University

Attributed Alignment

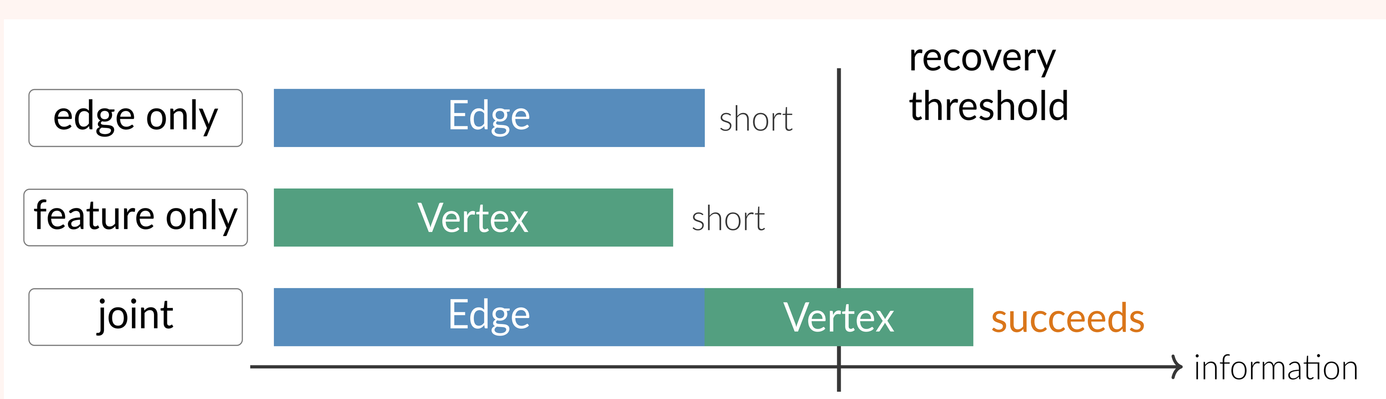
Many alignment tasks provide two noisy networks and node-side information. The same hidden correspondence must explain both channels.



- **Bibliographic databases:** align ACM and DBLP papers from co-author links plus author, venue, or title features.
- **Social platforms:** align online/offline users from interaction strengths plus profile or location attributes.
- **Biological networks:** align genes/proteins from weighted interaction graphs plus functional annotations.

Core Challenge

Topology alone can be ambiguous; attributes alone can be noisy or weak. The useful signal is often distributed across both.

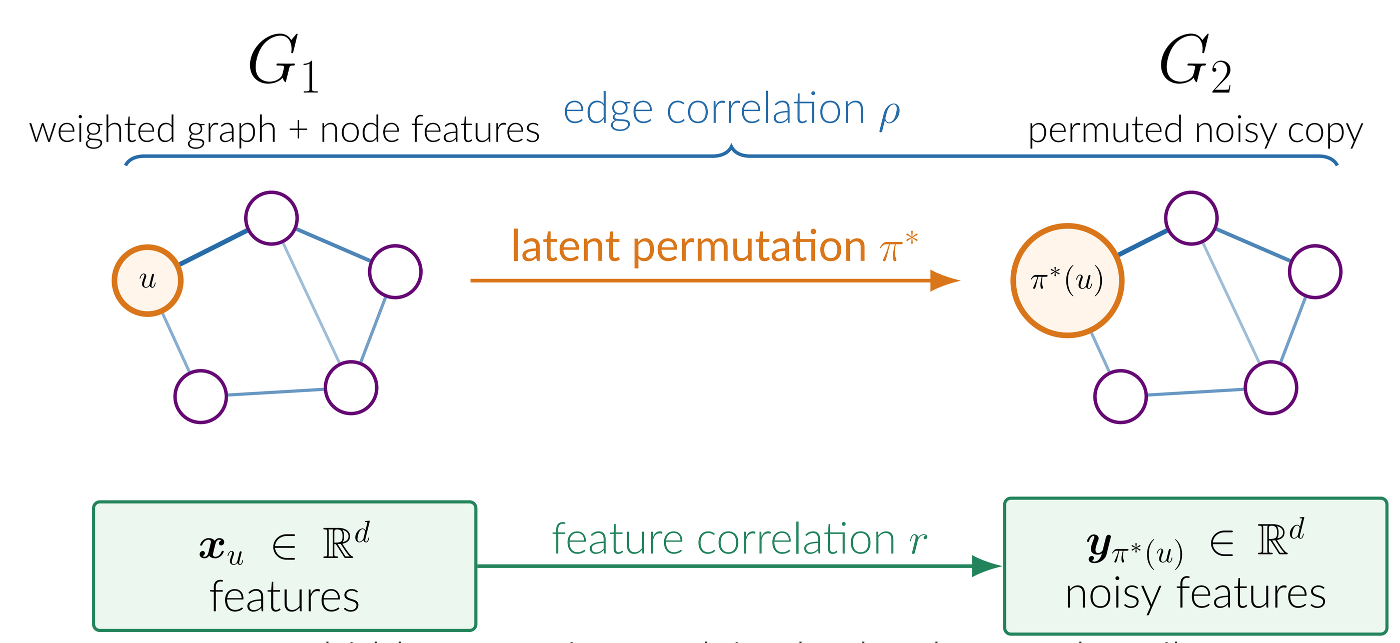


- Edge information and feature information each contribute mutual information.
- The recovery boundary is reached when their combined signal crosses the threshold.

Main Messages

- We introduce a weighted attributed random graph model: **featured correlated Gaussian Wigner model**.
- We characterize sharp information-theoretic thresholds for partial and exact recovery.
- We propose **QPAlign**, a quadratic-programming relaxation that combines weighted edges and node features efficiently.

Problem Setting



same hidden mapping explains both edges and attributes

Let $\pi^* \in \mathcal{S}_n$ be the unknown vertex mapping. In $\mathcal{G}(n, d, \rho, r)$:

- Edges: $(\beta_{uv}(G_1), \beta_{\pi^*(u)\pi^*(v)}(G_2))$ are correlated standard normals with correlation ρ.
- Features: $(\boldsymbol{x}_u, \boldsymbol{y}_{\pi^*(u)})$ are jointly Gaussian with covariance

$$\Sigma_d = \begin{bmatrix} I_d & rI_d \\ rI_d & I_d \end{bmatrix}.$$

- Edges and features are independent conditional on π^*.

Goal: Recover the Hidden Permutation

For an estimator $\hat{\pi}$, define $\mathrm{overlap}(\hat{\pi}, \pi^*) = \frac{1}{n}|\{u : \hat{\pi}(u) = \pi^*(u)\}|$.

- **Partial recovery:** $\mathrm{overlap}(\hat{\pi}, \pi^*) \geq \delta$ for a fixed $\delta \in (0, 1)$.
- **Exact recovery:** $\mathrm{overlap}(\hat{\pi}, \pi^*) = 1$.

The maximum-likelihood score balances the two channels:

$$\hat{\pi}_\lambda = \arg\max_{\pi \in \mathcal{S}_n} \left\{ \lambda \sum_{i<j} \beta_{ij}(G_1)\beta_{\pi(i)\pi(j)}(G_2) + (1-\lambda) \sum_i \boldsymbol{x}_i^\top \boldsymbol{y}_{\pi(i)} \right\}.$$

Sharp Statistical Limits

Define the information contributions

$$\textsf{Edge} = n \log \frac{1}{1-\rho^2}, \qquad \textsf{Vertex} = d \log \frac{1}{1-r^2}.$$

For any fixed constant $\epsilon > 0$:

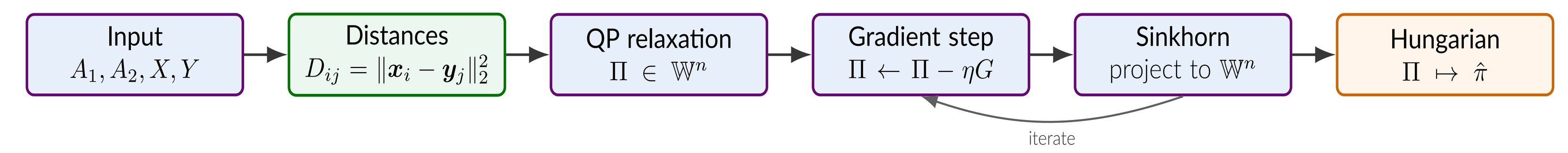
Task	Theorem condition	Guarantee
Partial recovery	$d = \omega(\log n)$, Edge $+ 2$ Vertex $\geq (4+\epsilon)\log n$	$\mathbb{P}(\mathrm{overlap}(\hat{\pi}, \pi^*) \geq \delta) = 1 - o(1)$
Exact recovery	$d = \omega(\log n)$, Edge $+$ Vertex $\geq (4+\epsilon)\log n$	$\mathbb{P}(\hat{\pi} = \pi^*) = 1 - o(1)$

- Partial recovery is possible when **Edge** $+ 2$ **Vertex** $\leq (4-\epsilon)\log n$.
- Exact recovery is possible when **Edge** $+$ **Vertex** $\leq (4-\epsilon)\log n$.
- Joint regime: if each channel is below $4\log n$ but their sum exceeds $(4+\epsilon)\log n$, using both channels enables recovery.

QPAlign: Efficient Relaxation

The exact MLE searches over $n!$ permutations. QPAlign rewrites it as a quadratic assignment problem, then relaxes permutation matrices to the Birkhoff polytope $\mathbb{W}^n$:

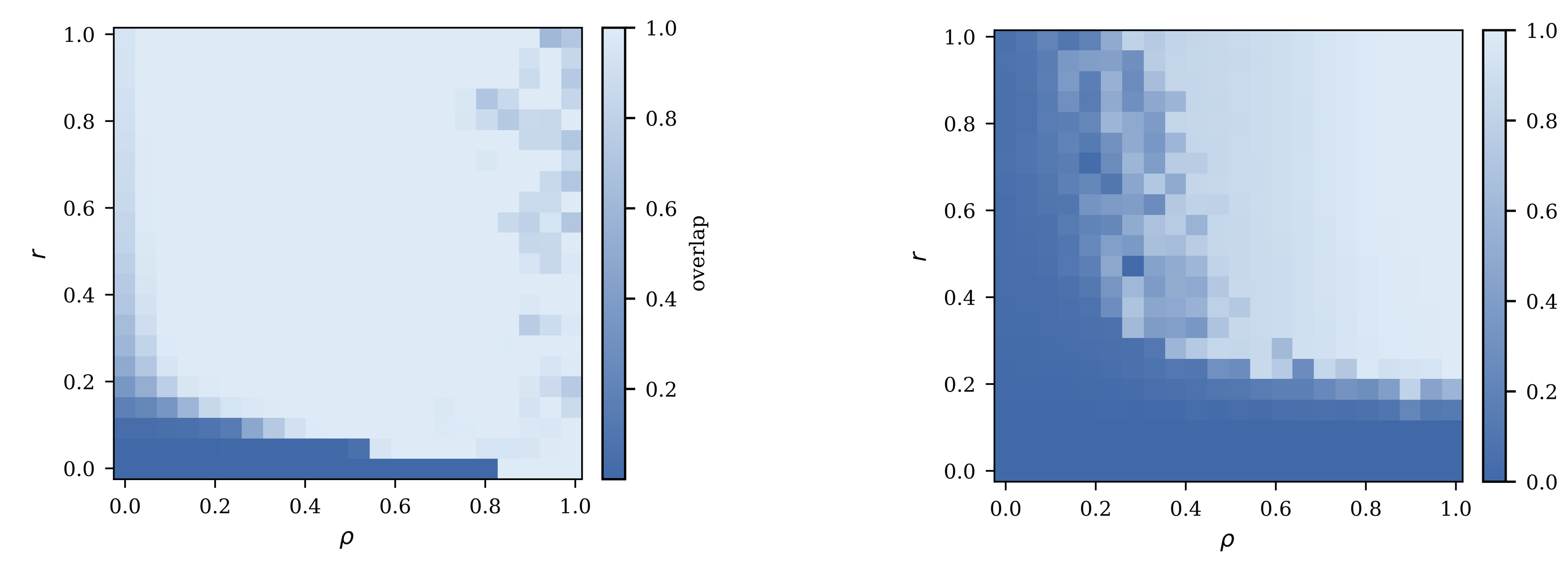
$$\min_{\Pi \in \mathbb{W}^n} \lambda \|A_1\Pi - \Pi A_2\|_F^2 + (1-\lambda)\sum_{i,j} D_{ij}\Pi_{ij}^2 + \mu \sum_{i,j} \Pi_{ij}(1 - \Pi_{ij}).$$



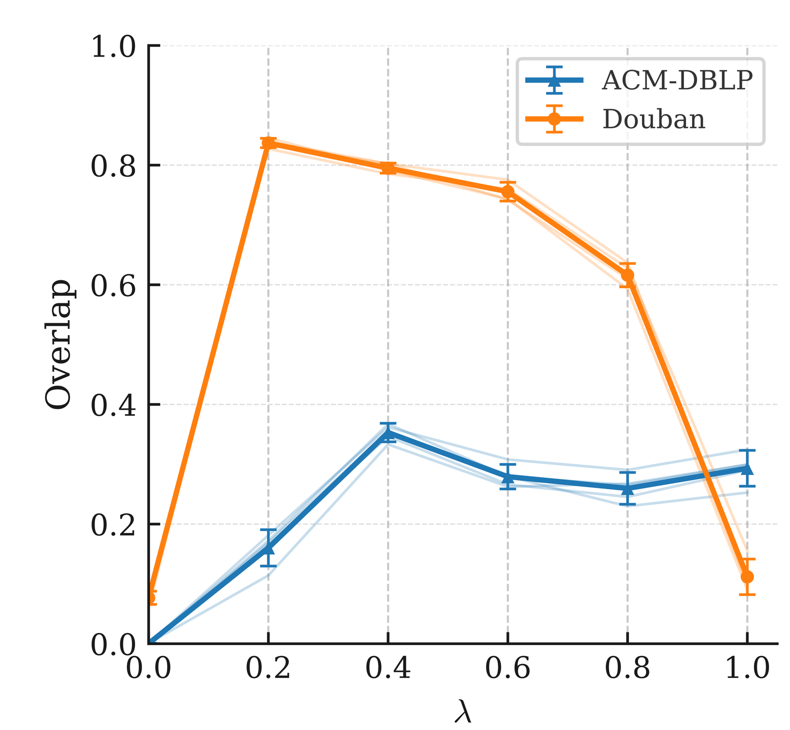
- $D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2$ injects node features, while $\|A_1\Pi - \Pi A_2\|_F^2$ aligns weighted neighborhoods.
- Each iteration takes a gradient step, truncates negative entries, and uses Sinkhorn as a fast projection onto $\mathbb{W}^n$.
- Hungarian rounding returns $\hat{\pi}$; total time is $O((d + T(K+n))n^2)$.

Experiments: Synthetic and Real Data

Synthetic. QPAlign moves from zero to full overlap as edge correlation ρ or feature correlation r strengthens.



Real networks. Combining sources improves ACM-DBLP alignment and remains competitive on Douban.



Method	ACM	Douban
QPAlign	**0.3445**	0.8370
PARROT	0.0441	**0.8462**
FGW	0.0018	0.2773
REGAL	0.0301	0.1118
Edge only	0.2896	0.1118
Feature only	0.0004	0.0767