The Planted Matching Problem: Sharp Threshold and Infinite-order Phase Transition

Jiaming Xu

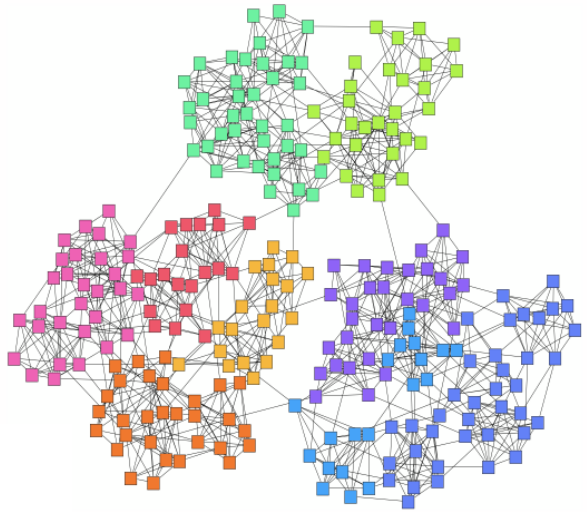
The Fuqua School of Business
Duke University

Joint work with
Jian Ding (UPenn), Yihong Wu (Yale) and Dana Yang (Duke)

April 12, 2021

Statistical model with planted structure

Question: How to recover latent structure from noisy data?



Classical examples

- Recovery of planted clique in Erdős-Rényi graphs
- Community detection under Stochastic Block Model
- Clustering in mixture models

Common theme: low-rank structure

- Underpinning of many phase transitions and algorithms, e.g. spectral method, SDP relaxation, etc

A new zoo of planted problems...

- Planted bipartite matching [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10]
- Graph matching (network alignment) [Pedarsani-Grossglauser '11]
- Planted Hamiltonian cycle problem (TSP) [Bagaria-Ding-Tse-W-Xu '18]
- Planted trees [Massoulié-Stephan-Towsley '18]
- Planted k-factors [Sicuro-Zdeborová '20]
- Planted k-nearest-neighbor graph [Ding-Wu-Xu-Yang '19]

A new zoo of planted problems...

- Planted bipartite matching [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10]
- Graph matching (network alignment) [Pedarsani-Grossglauser '11]
- Planted Hamiltonian cycle problem (TSP) [Bagaria-Ding-Tse-W-Xu '18]
- Planted trees [Massoulié-Stephan-Towsley '18]
- Planted k-factors [Sicuro-Zdeborová '20]
- Planted k-nearest-neighbor graph [Ding-Wu-Xu-Yang '19]

Common theme: Lack of low-rank structure $\Rightarrow$ new challenges in both statistical analysis and algorithm design

A new zoo of planted problems...

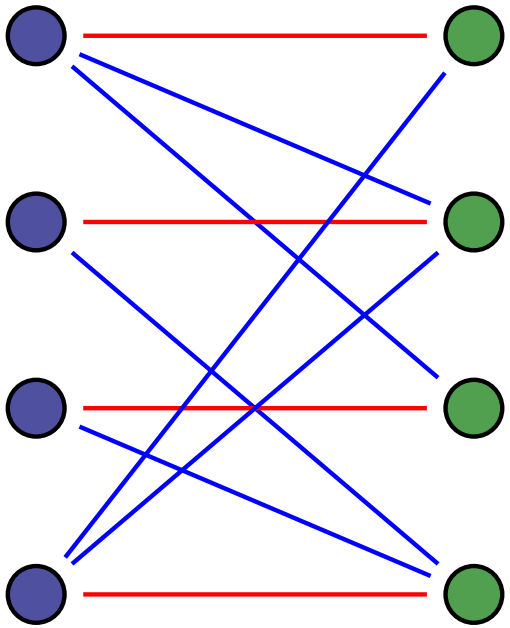
- Planted bipartite matching [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10]
- Graph matching (network alignment) [Pedarsani-Grossglauser '11]
- Planted Hamiltonian cycle problem (TSP) [Bagaria-Ding-Tse-W-Xu '18]
- Planted trees [Massoulié-Stephan-Towsley '18]
- Planted k-factors [Sicuro-Zdeborová '20]
- Planted k-nearest-neighbor graph [Ding-Wu-Xu-Yang '19]

Common theme: Lack of low-rank structure $\Rightarrow$ new challenges in both statistical analysis and algorithm design

Today:

- Planted bipartite matching

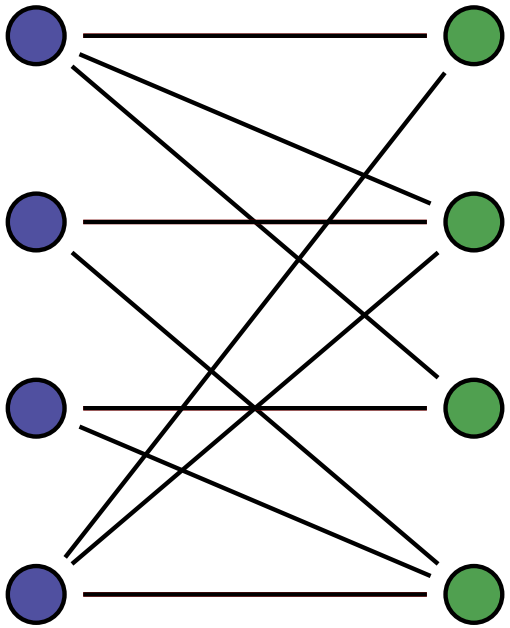
The planted matching model



- A weighted bipartite graph G
- A hidden perfect matching M^*
- All $n(n-1)$ pairs not in M^* are connected w.p. $\frac{d}{n}$
- Edge weight

$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in M^* \\ Q & e \notin M^* \end{cases}$$

The planted matching model

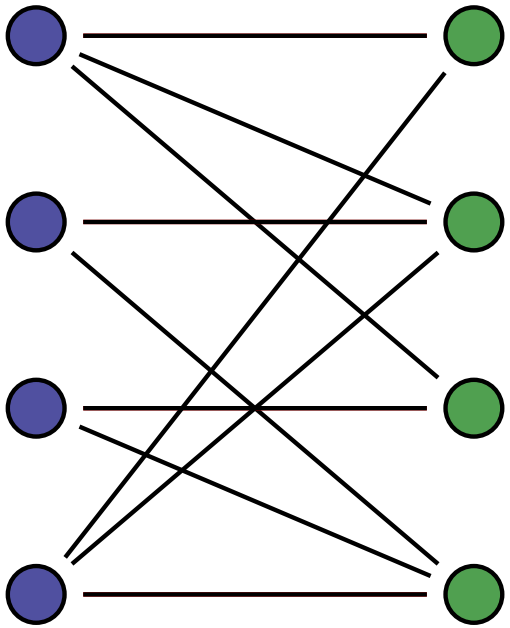


- A weighted bipartite graph G
- A hidden perfect matching M^*
- All $n(n-1)$ pairs not in M^* are connected w.p. $\frac{d}{n}$
- Edge weight

$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in M^* \\ Q & e \notin M^* \end{cases}$$

- Goal: recover M^* from G

The planted matching model



- A weighted bipartite graph G
- A hidden perfect matching M^*
- All $n(n-1)$ pairs not in M^* are connected w.p. $\frac{d}{n}$
- Edge weight

$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in M^* \\ Q & e \notin M^* \end{cases}$$

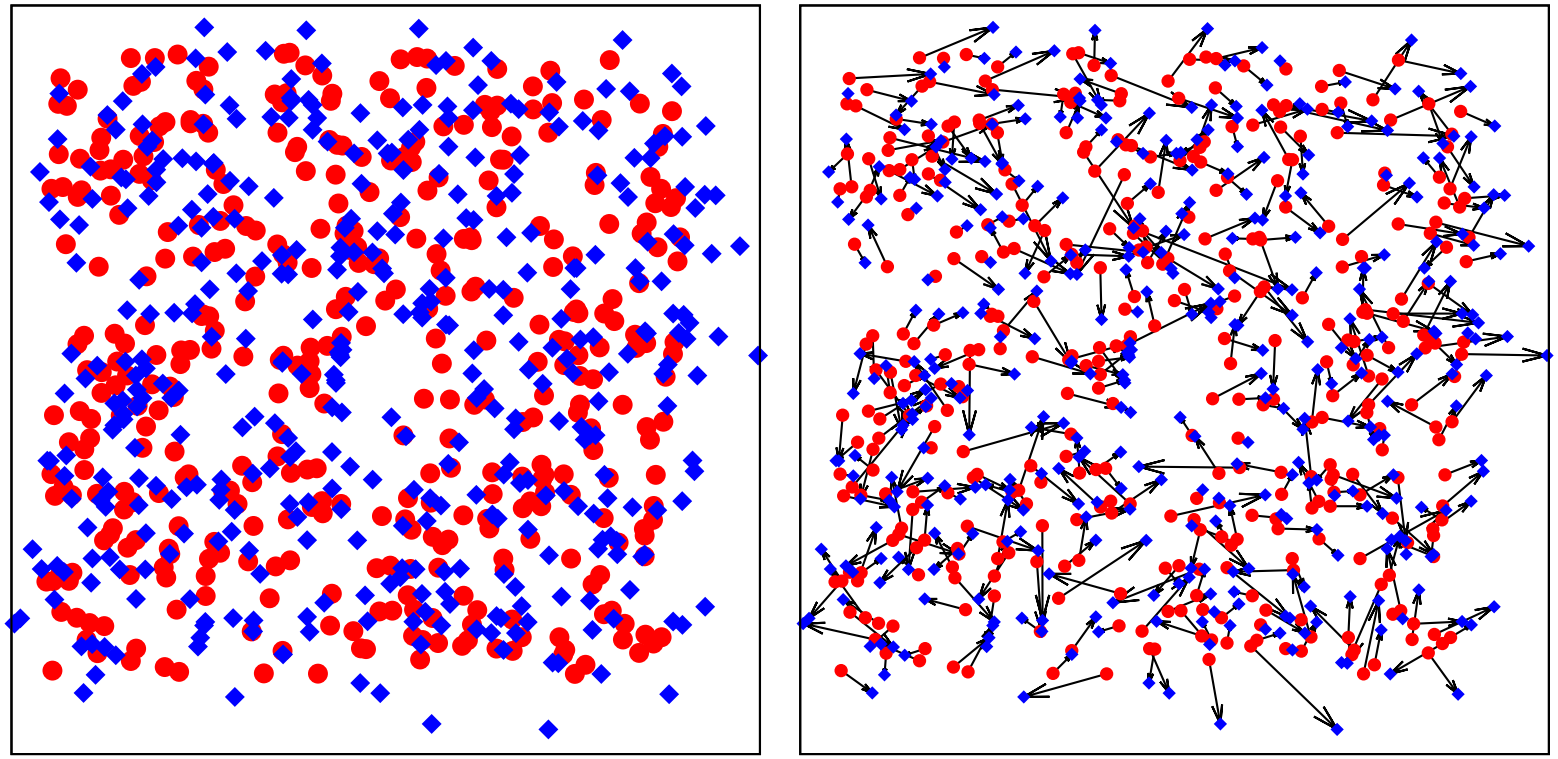
- Goal: recover M^* from G

$d = n$ and $P = Q = \exp(1)$: celebrated random assignment problem

[Walkup'79, Mézard-Parisi'87, Steele'97, Aldous'01, Nair-Prabhakar-Sharma'05, Wästlund'09]

$$\mathbb{E}\left[\min_{M \in \mathcal{M}} \sum_{e \in M} W_e\right] = 1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{n^2} \to \frac{\pi^2}{6}$$

Motivating application: particle tracking



[Chertkov-Kroc-Krzakala-Vergassola-Zdeborová PNAS'10]

- Tracking particles advected by turbulent fluid flow
- Goal: recover the latent correspondence between particles
- $d = n$, $\mathcal{P} = |\mathcal{N}(0, \sigma^2)|$ and $\mathcal{Q} = \text{Uniform}[0, n]$

Maximum likelihood estimation as linear assignment

Maximum likelihood estimation reduces to max-weighted matching:

$$\widehat{M}_{\mathrm{ML}} = \arg\max_{M \in \mathcal{M}} \sum_{e \in M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$$

- Linear assignment: computable in polynomial time
- For certain distributions e.g. exponentials, further reduce to min-weighted matching in terms of W_e

Maximum likelihood estimation as linear assignment

Maximum likelihood estimation reduces to max-weighted matching:

$$\widehat{M}_{\mathrm{ML}} = \arg\max_{M \in \mathcal{M}} \sum_{e \in M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$$

- Linear assignment: computable in polynomial time
- For certain distributions e.g. exponentials, further reduce to min-weighted matching in terms of W_e
- How much does $\widehat{M}_{\mathrm{ML}}$ have in common with M^*?

$$\mathrm{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \triangleq \frac{1}{n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \cap M^*\right| = 1 - \frac{1}{2n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \triangle M^*\right|$$

Maximum likelihood estimation as linear assignment

Maximum likelihood estimation reduces to max-weighted matching:

$$\widehat{M}_{\mathrm{ML}} = \arg\max_{M \in \mathcal{M}} \sum_{e \in M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$$

- Linear assignment: computable in polynomial time
- For certain distributions e.g. exponentials, further reduce to min-weighted matching in terms of W_e
- How much does $\widehat{M}_{\mathrm{ML}}$ have in common with M^*?

$$\mathrm{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \triangleq \frac{1}{n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \cap M^*\right| = 1 - \frac{1}{2n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \triangle M^*\right|$$

- Information-theoretic limit for reconstruction, in terms of
  1. Average degree d
  2. Similarity between $\mathcal{P}$ and $\mathcal{Q}$

Maximum likelihood estimation as linear assignment

Maximum likelihood estimation reduces to max-weighted matching:

$$\widehat{M}_{\mathrm{ML}} = \arg\max_{M \in \mathcal{M}} \sum_{e \in M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$$

- Linear assignment: computable in polynomial time
- For certain distributions e.g. exponentials, further reduce to min-weighted matching in terms of W_e
- How much does $\widehat{M}_{\mathrm{ML}}$ have in common with M^*?

$$\mathrm{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \triangleq \frac{1}{n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \cap M^*\right| = 1 - \frac{1}{2n}\mathbb{E}\left|\widehat{M}_{\mathrm{ML}} \triangle M^*\right|$$

- Information-theoretic limit for reconstruction, in terms of
  1. Average degree d
  2. Similarity between $\mathcal{P}$ and $\mathcal{Q}$
     - Bhattacharyya coefficient (Hellinger affinity) $B(\mathcal{P}, \mathcal{Q}) \triangleq \int \sqrt{\mathrm{d}\mathcal{P}\mathrm{d}\mathcal{Q}}$

Main result: phase transition threshold

Theorem (Ding-Wu-X.-Yang '21)

- *If* $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) \le 1$, *then* $\mathsf{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \to 1$;

Main result: phase transition threshold

Theorem (Ding-Wu-X.-Yang '21)

- *If* $\sqrt{d}\ B(\mathcal{P}, \mathcal{Q}) \le 1$, *then* $\mathrm{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \to 1$;
- *Conversely, assuming* d, P, Q *independent of* n, *if* $\sqrt{d}\ B(\mathcal{P}, \mathcal{Q}) \ge 1 + \epsilon$, *then for all* $\widehat{M}$ *and some* $c = c(\epsilon)$

$$\mathrm{overlap}(\widehat{M}, M^*) \le 1 - c.$$

- Resolve the conjecture in [Semerjian-Sicuro-Zdeborová '20]

Main result: phase transition threshold

Theorem (Ding-Wu-X.-Yang '21)

- *If* $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) \le 1$, *then* $\mathrm{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) \to 1$;
- *Conversely, assuming* d, P, Q *independent of* n, *if* $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) \ge 1 + \epsilon$, *then for all* $\widehat{M}$ *and some* $c = c(\epsilon)$

$$\mathrm{overlap}(\widehat{M}, M^*) \le 1 - c.$$

- Resolve the conjecture in [Semerjian-Sicuro-Zdeborová '20]
- Generalize to dense model: $d \equiv d(n) \to \infty$ and Q is under proper scaling: e.g. $d = n$,
  - $P = |\mathcal{N}(0, \sigma^2)|$, $Q = \mathrm{Uniform}[0, n]$:

$$\text{Sharp threshold } \sigma^2 = \frac{1}{2\pi}$$

  - $P = \exp(\lambda)$, $Q = \exp(1/n)$ (mean $1/\lambda$ vs. n):

$$\text{Sharp threshold } \lambda = 4$$

Infinite-order phase transition under exponential model

Theorem (Ding-Wu-X.-Yang '21)

Assume $\lambda = 4 - \epsilon$. There exist absolute constants c_1, c_2:

$$\text{overlap}(\widehat{M}_{\text{ML}}, M^*) \geq 1 - e^{-\frac{c_1}{\sqrt{\epsilon}}};$$

Conversely, for all $\widehat{M}$,

$$\text{overlap}(\widehat{M}, M^*) \leq 1 - e^{-\frac{c_2}{\sqrt{\epsilon}}}.$$

Infinite-order phase transition under exponential model

Theorem (Ding-Wu-X.-Yang '21)

Assume $\lambda = 4 - \epsilon$. There exist absolute constants c_1, c_2:

$$\text{overlap}(\widehat{M}_{\text{ML}}, M^*) \geq 1 - e^{-\frac{c_1}{\sqrt{\epsilon}}};$$

Conversely, for all $\widehat{M}$,

$$\text{overlap}(\widehat{M}, M^*) \leq 1 - e^{-\frac{c_2}{\sqrt{\epsilon}}}.$$

- Optimal reconstruction error is $\exp\left(-\Theta(1/\sqrt{\epsilon})\right)$
- Resolve the ∞-order phase transition conjecture [Semerjian-Sicuro-Zdeborová '20]

Infinite-order phase transition diagram

Using local weak convergence, [Maharrami-Moore-Xu '19] determines the exact overlap of MLE

$$\lim_{n\to\infty} \text{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) = \alpha(\lambda)$$

Infinite-order phase transition diagram

Using local weak convergence, [Maharrami-Moore-Xu '19] determines the exact overlap of MLE

$$\boxed{\lim_{n\to\infty} \text{overlap}(\widehat{M}_{\mathrm{ML}}, M^*) = \alpha(\lambda)}$$

$\alpha(\lambda) = 1 - 2\int_0^\infty (1 - F(x))\,(1 - G(x))\,V(x)W(x)\,dx$, and F, G, V, W is the unique solution to the ODE system

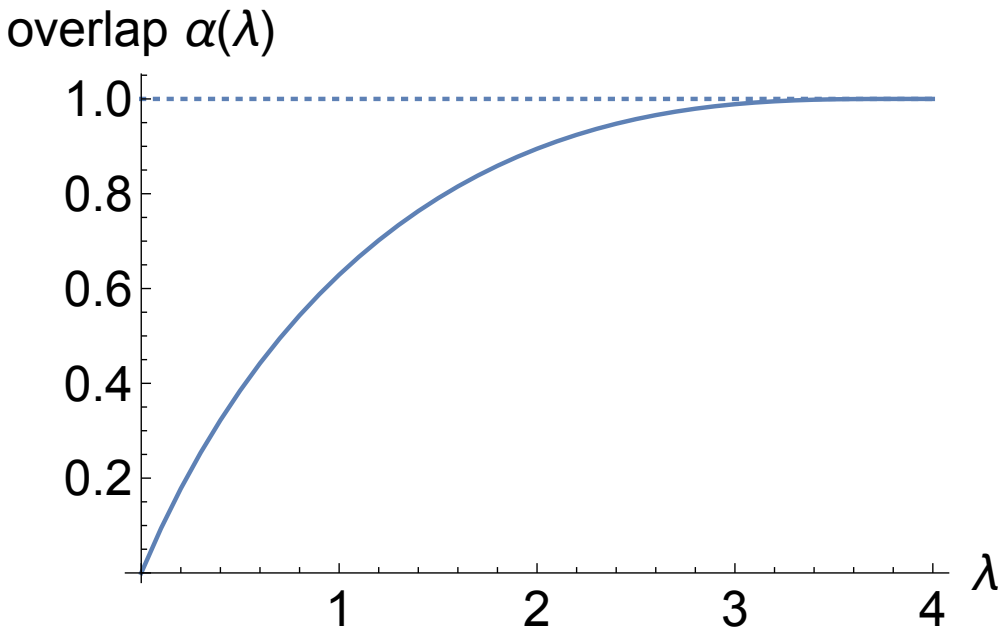
$$\begin{aligned}
\dot{F} &= (1-F)(1-G)V \\
\dot{G} &= -(1-F)(1-G)W \\
\dot{V} &= \lambda(V-F) \\
\dot{W} &= -\lambda(W-G)
\end{aligned}$$

Boundary conditions: $F(x), V(x), G(-x), W(-x) \to \begin{cases} 1 & x \to +\infty \\ 0 & x \to -\infty \end{cases}$

Infinite-order phase transition diagram

Using local weak convergence, [Maharrami-Moore-Xu '19] determines the exact overlap of MLE

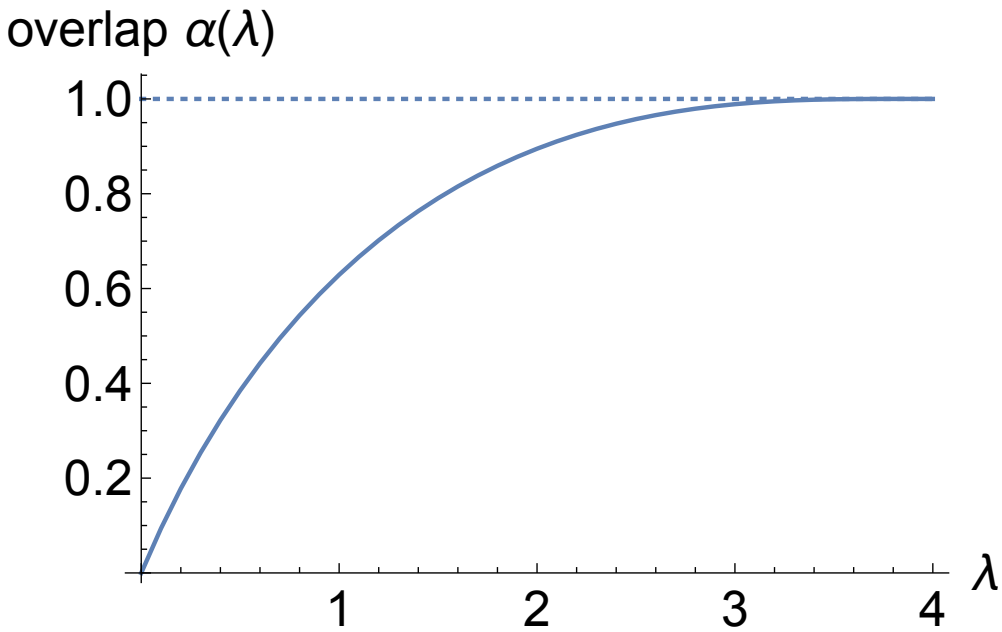
$$\lim_{n\to\infty} \text{overlap}(\widehat{M}_{\text{ML}}, M^*) = \alpha(\lambda)$$



Infinite-order phase transition diagram

Using local weak convergence, [Maharrami-Moore-Xu '19] determines the exact overlap of MLE

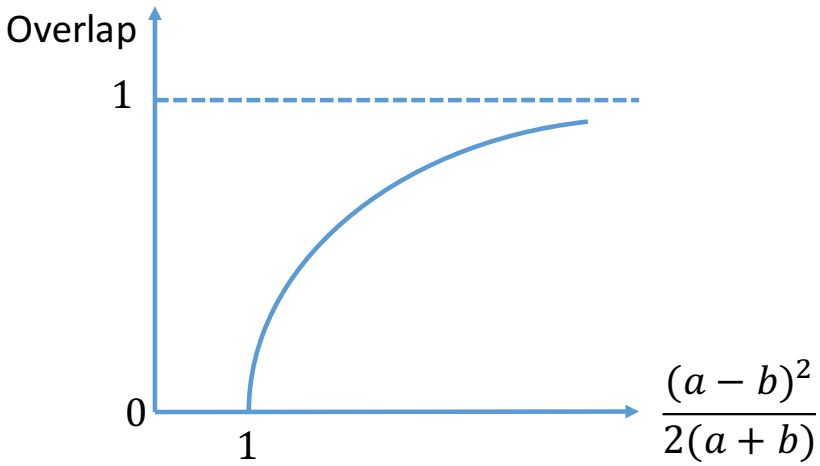
$$\lim_{n\to\infty} \text{overlap}(\widehat{M}_{\text{ML}}, M^*) = \alpha(\lambda)$$



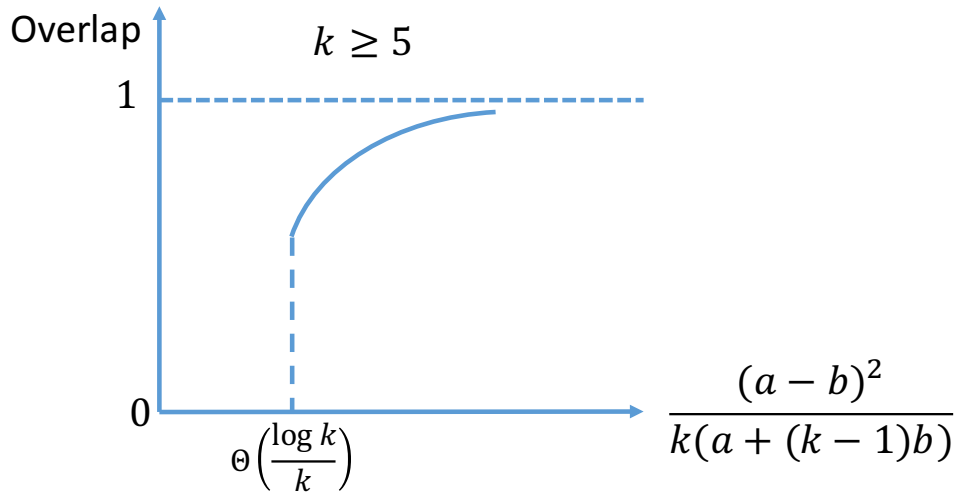
- $\alpha(\lambda)$ is infinitely differentiable at threshold $\lambda = 4$
- Our overlap lower bound follows from analyzing the ODE system

Comparison of phase transition orders

Drastically different from the other well-known planted models such as stochastic block model (conjecture, not fully proven yet)



Second-order phase transition
with two groups
[DKMZ'11, MNS'12 13, Massoulié'13]



First-order phase transition
with five or more groups
[DKMZ'11, BMNN '16, AS'16]

Analysis

- Proof of positive result via maximum likelihood
- Proof of negative result via analyzing posterior distribution
- Proof of tight error lower bound under exponential model

Proof of positive result via maximum likelihood

- At most $\binom{n}{t} t!$ matchings M with $|M \triangle M^*| = 2t$
- Probability that M has higher likelihood than M^* is

$$\mathbb{P}\left\{ \sum_{e \in M \setminus M^*} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \geq \sum_{e \in M^* \setminus M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \right\} \leq \left(\frac{d}{n} B^2(\mathcal{P}, \mathcal{Q}) \right)^t$$

Proof of positive result via maximum likelihood

- At most $\binom{n}{t} t!$ matchings M with $|M \triangle M^*| = 2t$
- Probability that M has higher likelihood than M^* is

$$\mathbb{P}\left\{ \sum_{e \in M \setminus M^*} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \geq \sum_{e \in M^* \setminus M} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \right\} \leq \left(\frac{d}{n} B^2(\mathcal{P}, \mathcal{Q}) \right)^t$$

- Taking union bound $\Rightarrow$

$$\mathbb{P}\left[\exists\ M \text{ with } |M \triangle M^*| \geq \beta n \text{ has higher likelihood than } M^*\right]$$
$$\leq \sum_{t \geq \beta n} \binom{n}{t} t! \left(\frac{d}{n} B^2(\mathcal{P}, \mathcal{Q}) \right)^t$$
$$\to 0 \text{ for some } \beta = o(1), \text{ if } \sqrt{d}\ B(\mathcal{P}, \mathcal{Q}) \leq 1$$

Analysis

- Proof of positive result via maximum likelihood
- Proof of negative result via analyzing posterior distribution
- Proof of tight error lower bound under exponential model

Negative result via analyzing posterior distribution

- ...But MLE does not maximize overlap
- Instead, analyze posterior distribution: Gibbs distribution over perfect matchings

$$\mu_W(m) \propto \exp\left(\sum_{e \in m} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)\right)$$

Negative result via analyzing posterior distribution

- ...But MLE does not maximize overlap
- Instead, analyze posterior distribution: Gibbs distribution over perfect matchings

$$\mu_W(m) \propto \exp\left(\sum_{e \in m} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)\right)$$

Crucial observation

Sampling from posterior distribution is optimal within a factor of two.

Negative result via analyzing posterior distribution

- ...But MLE does not maximize overlap
- Instead, analyze posterior distribution: Gibbs distribution over perfect matchings

$$\mu_W(m) \propto \exp\left(\sum_{e\in m} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)\right)$$

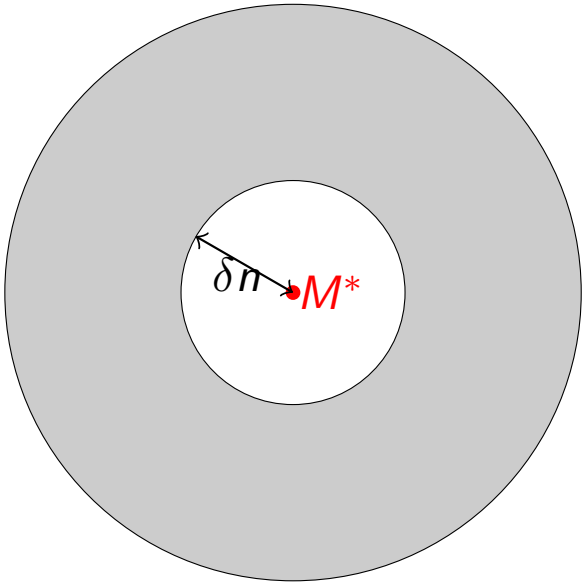
Crucial observation

Sampling from posterior distribution is optimal within a factor of two.

Proof: Let $\widetilde{M}$ be sampled from posterior distribution. Then for any estimator $\widehat{M}$, $(M^*, \widehat{M}) \overset{\text{law}}{=} (\widetilde{M}, \widehat{M})$ and

$$\mathbb{E}|\widetilde{M} \triangle M^*| \le \mathbb{E}|\widetilde{M} \triangle \widehat{M}| + \mathbb{E}|\widehat{M} \triangle M^*| = 2\mathbb{E}|\widehat{M} \triangle M^*|.$$

Analysis of posterior distribution



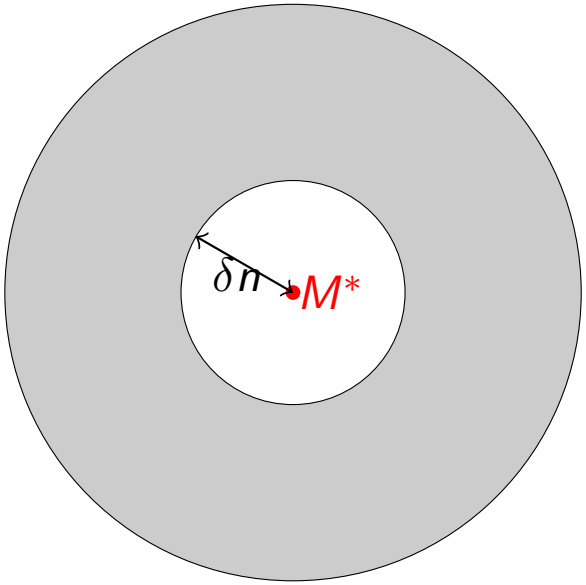
- Upper bound the posterior mass of matchings near M^*:

$$\frac{\mu_W(\mathcal{M}_{\text{near}})}{\mu_W(M^*)} \leq e^{7\epsilon\delta n} \tag{1}$$

- Lower bound the posterior mass of matchings far away from M^*:

$$\frac{\mu_W(\mathcal{M}_{\text{far}})}{\mu_W(M^*)} \geq e^{14\epsilon\delta n} \tag{2}$$

Analysis of posterior distribution



- Upper bound the posterior mass of matchings near M^*:

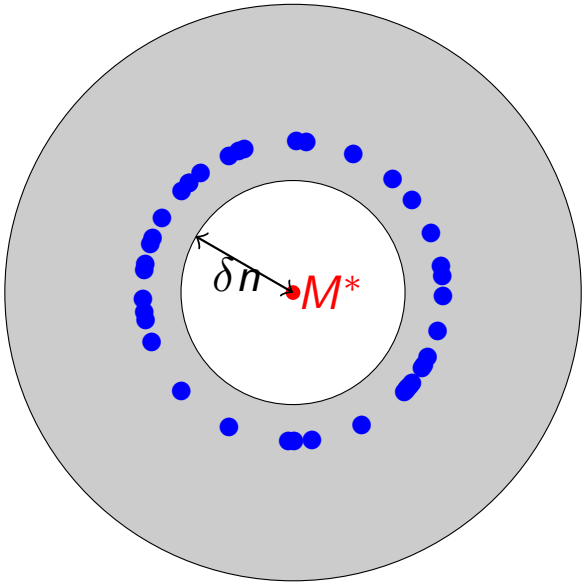
$$\frac{\mu_W(\mathcal{M}_{\text{near}})}{\mu_W(M^*)} \leq e^{7\epsilon\delta n} \tag{1}$$

- Lower bound the posterior mass of matchings far away from M^*:

$$\frac{\mu_W(\mathcal{M}_{\text{far}})}{\mu_W(M^*)} \geq e^{14\epsilon\delta n} \tag{2}$$

- Proof of (1) is straightforward: truncated first moment

Analysis of posterior distribution



- Upper bound the posterior mass of matchings near M^*:

$$\frac{\mu_W(\mathcal{M}_{\text{near}})}{\mu_W(M^*)} \leq e^{7\epsilon\delta n} \tag{1}$$

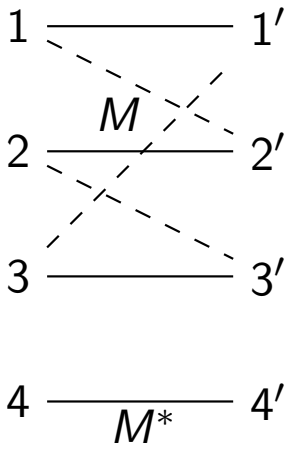
- Lower bound the posterior mass of matchings far away from M^*:

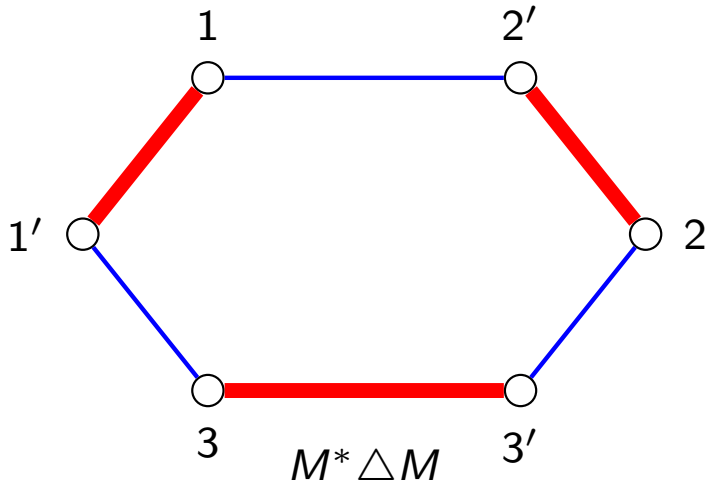
$$\frac{\mu_W(\mathcal{M}_{\text{far}})}{\mu_W(M^*)} \geq e^{14\epsilon\delta n} \tag{2}$$

- Proof of (1) is straightforward: truncated first moment
- Proof of (2) is constructive: find exponentially many matchings $M \in \mathcal{M}_{\text{far}}$ whose likelihood exceeds that of M^*

Lower bound: Augmenting alternating cycles

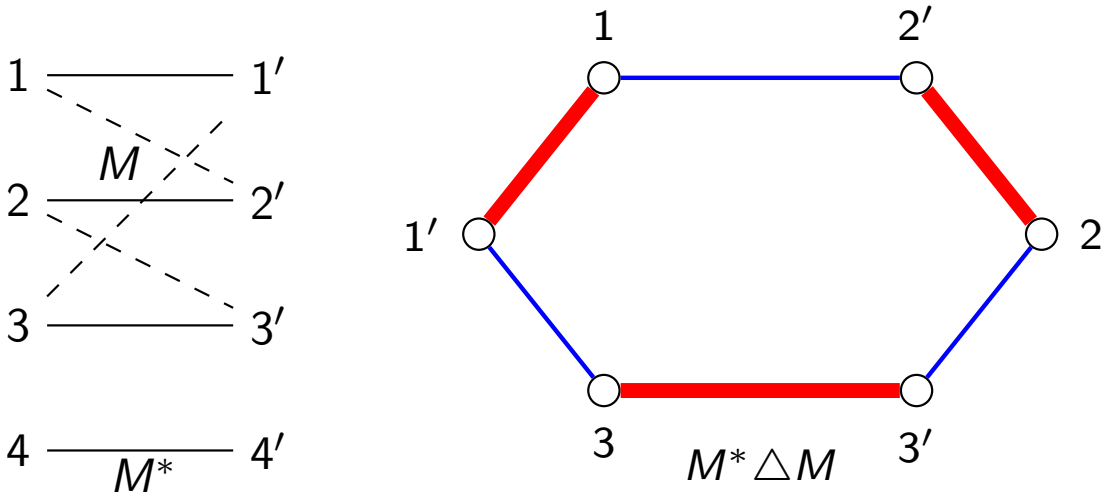
For perfect matching M, $M \triangle M^* =$ disjoint union of alternating cycles





Lower bound: Augmenting alternating cycles

For perfect matching M, $M \triangle M^* =$ disjoint union of alternating cycles

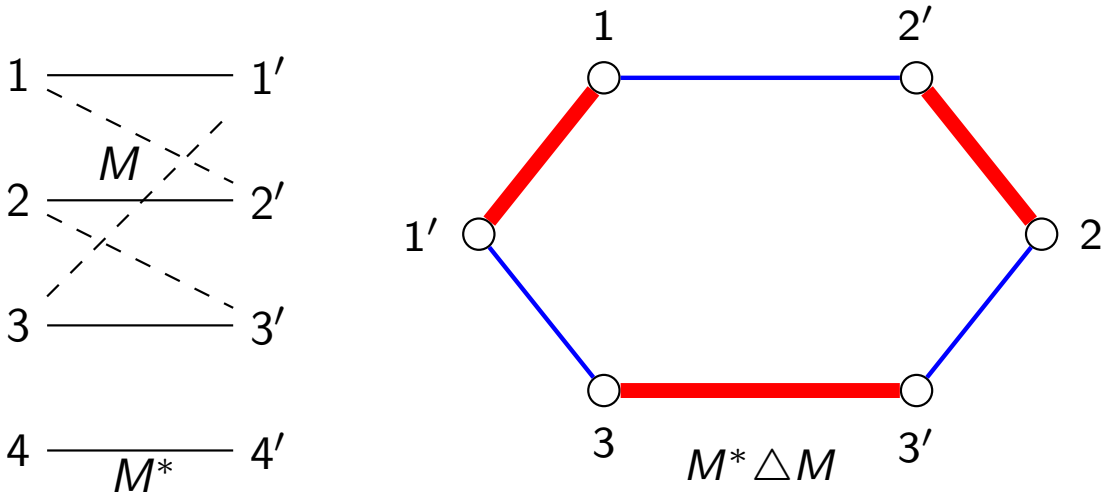


Goal: Find exponentially many long alternating cycles C that are augmenting:

$$\sum_{e \in E_{\text{blue}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \geq \sum_{e \in E_{\text{red}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e).$$

Lower bound: Augmenting alternating cycles

For perfect matching M, $M \triangle M^* =$ disjoint union of alternating cycles



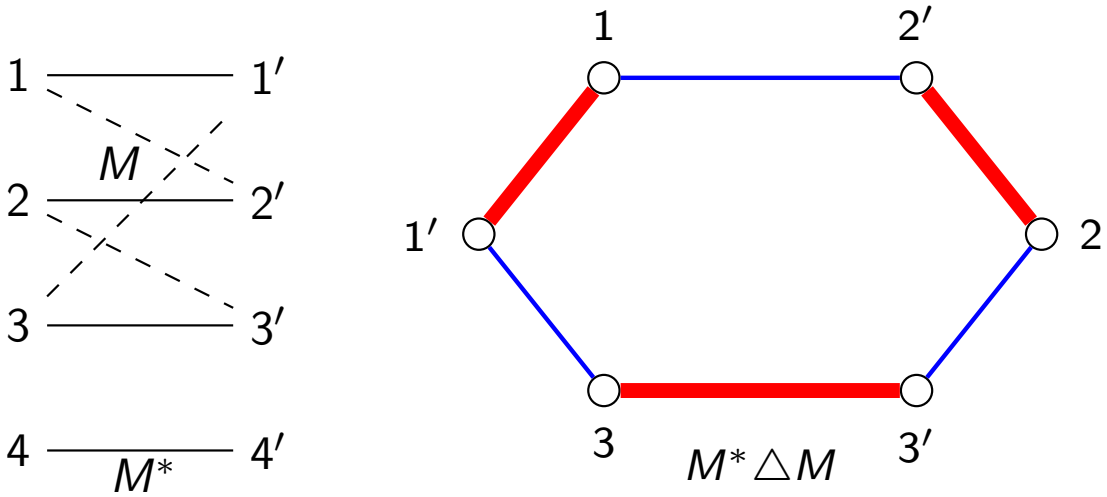
Goal: Find exponentially many long alternating cycles C that are augmenting:

$$\sum_{e \in E_{\text{blue}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \geq \sum_{e \in E_{\text{red}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e).$$

- Augmenting alternating cycles are rare;

Lower bound: Augmenting alternating cycles

For perfect matching M, $M \triangle M^* =$ disjoint union of alternating cycles



Goal: Find exponentially many long alternating cycles C that are augmenting:

$$\sum_{e \in E_{\text{blue}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e) \geq \sum_{e \in E_{\text{red}}(C)} \log \frac{\mathcal{P}}{\mathcal{Q}}(W_e).$$

- Augmenting alternating cycles are rare; but there are many alternating cycles

Failure of second-moment in counting alternating cycles

Natural attempt: first and second moment method.

- Let S be the set of augmenting alternating cycles in G of length at least cn. Then $\mathbb{E}|S| = e^{\Omega(n)}$.

Failure of second-moment in counting alternating cycles

Natural attempt: first and second moment method.

- Let S be the set of augmenting alternating cycles in G of length at least cn. Then $\mathbb{E}|S| = e^{\Omega(n)}$.
- If $\mathbb{E}(|S|^2) \lesssim (\mathbb{E}|S|)^2$, then $|S| = e^{\Omega(n)}$ with constant probability.

Failure of second-moment in counting alternating cycles

Natural attempt: first and second moment method.

- Let S be the set of augmenting alternating cycles in G of length at least cn. Then $\mathbb{E}|S| = e^{\Omega(n)}$.
- If $\mathbb{E}(|S|^2) \lesssim (\mathbb{E}|S|)^2$, then $|S| = e^{\Omega(n)}$ with constant probability.
- However, $\mathbb{E}(|S|^2) \gg (\mathbb{E}|S|)^2$ due to the excessive correlation between long cycles.

Failure of second-moment in counting alternating cycles

Natural attempt: first and second moment method.

- Let S be the set of augmenting alternating cycles in G of length at least cn. Then $\mathbb{E}|S| = e^{\Omega(n)}$.
- If $\mathbb{E}(|S|^2) \lesssim (\mathbb{E}|S|)^2$, then $|S| = e^{\Omega(n)}$ with constant probability.
- However, $\mathbb{E}(|S|^2) \gg (\mathbb{E}|S|)^2$ due to the excessive correlation between long cycles.

Key idea

First find many disjoint short paths, then connect the paths into long cycles [Aldous '98, Ding '13, ...]

Existence of many long augmenting alternating cycles

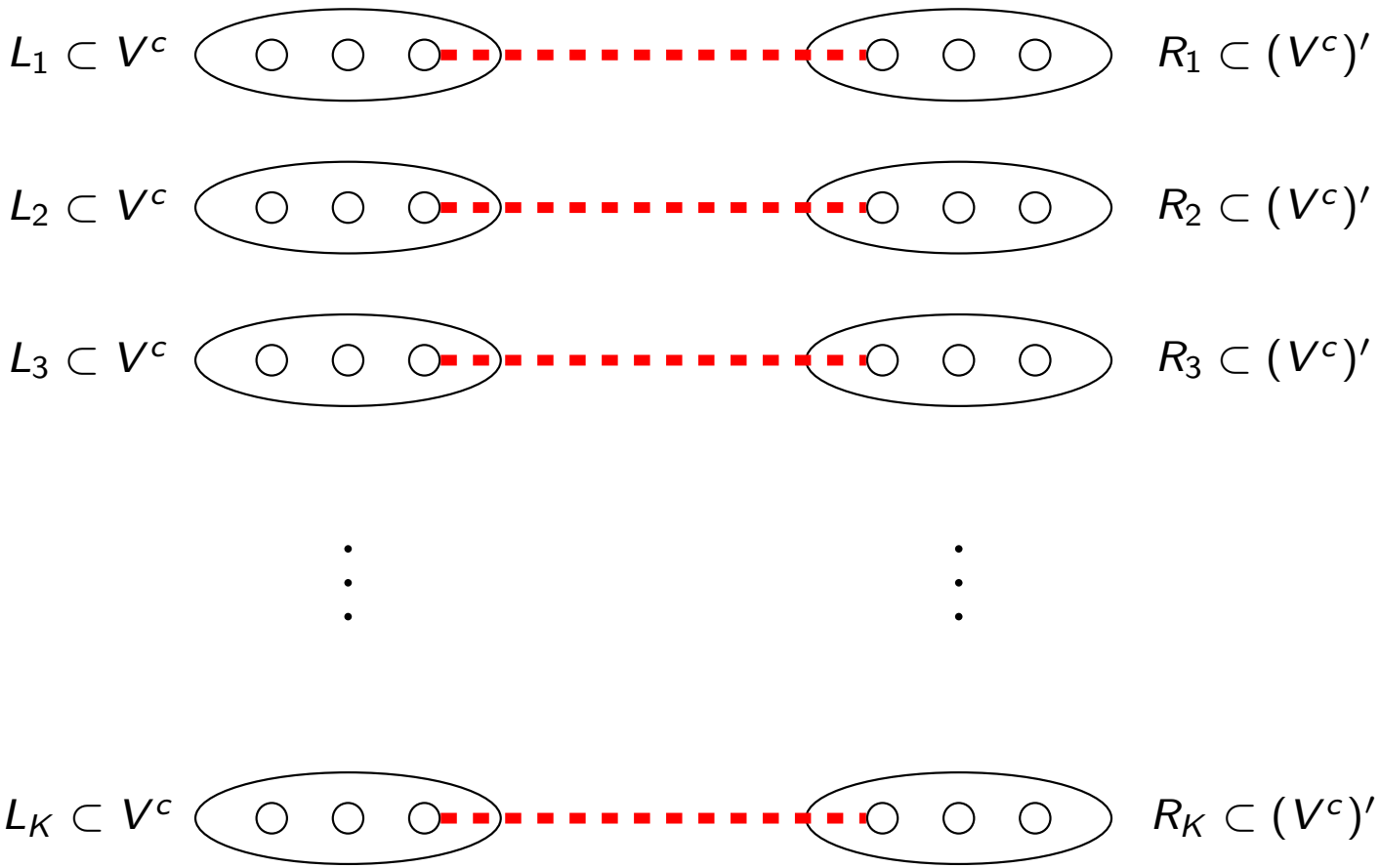
Two-stage cycle-finding scheme

Reserve a set V of γn vertices for some small $\gamma > 0$.

1. Stage 1 (path construction): Find many disjoint short (constant length) augmenting alternating paths, using vertices in V^c.
2. Stage 2 (sprinkling): Connect the paths into long cycles, using vertices in V.

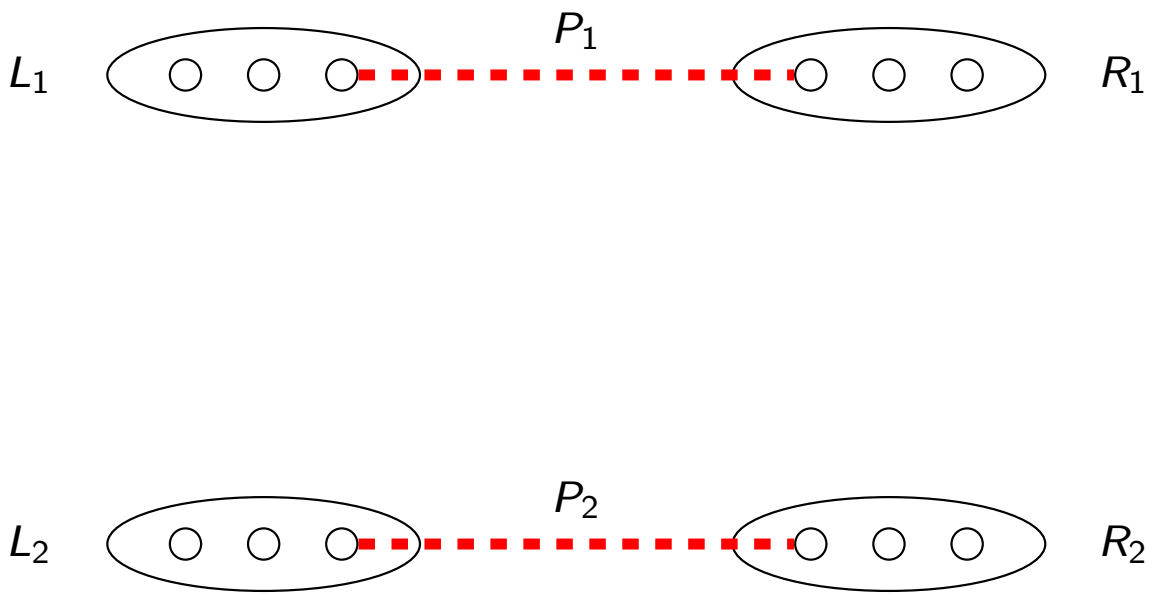
Two-stage cycle-finding scheme

Stage 1 (path construction): Find $\{L_k, R_k\}_{k=1}^{K}$ for $K = \Omega(n)$ such that every vertex in L_k is connected to every vertex in R_k via an augmenting alternating path (length = large constant)



Two-stage cycle-finding scheme

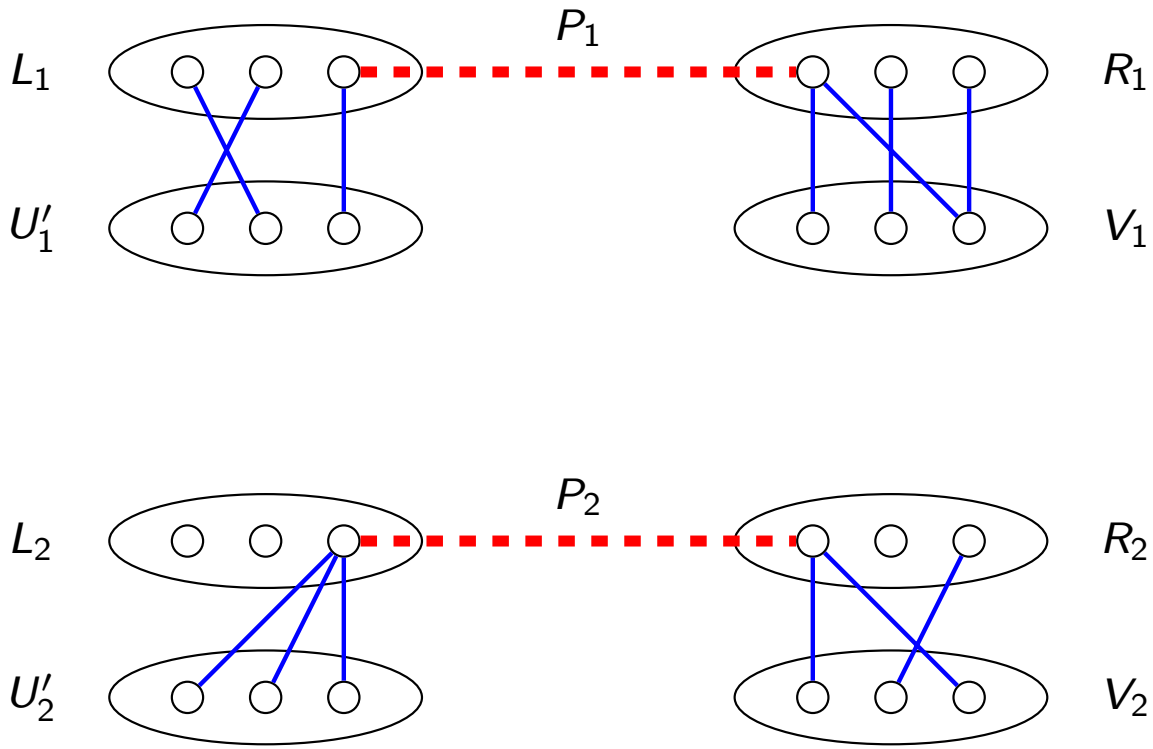
Stage 2 (sprinkling):



Two-stage cycle-finding scheme

Stage 2 (sprinkling):

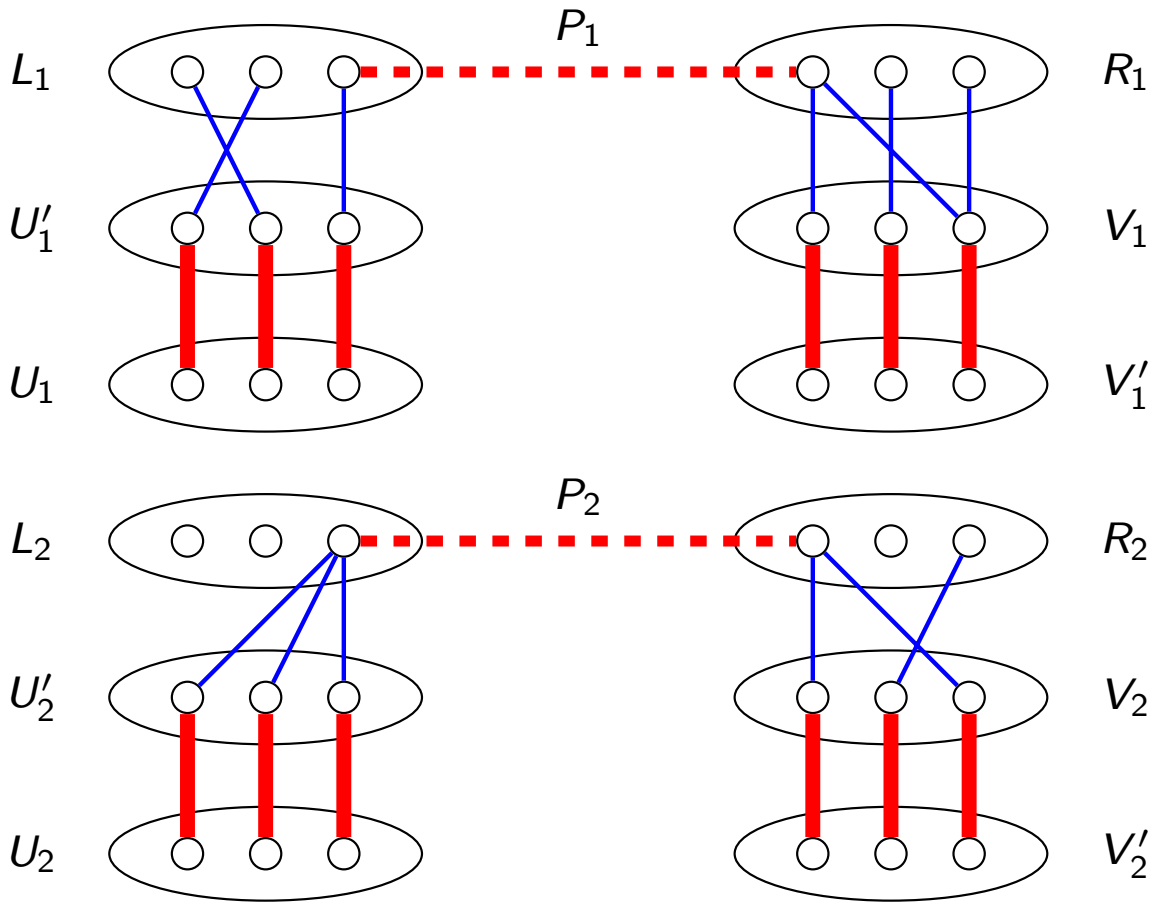
1. Let U'_k be set of reserved vertices connecting to L_k
   Let V_k be set of reserved vertices connecting to R_k



Two-stage cycle-finding scheme

Stage 2 (sprinkling):

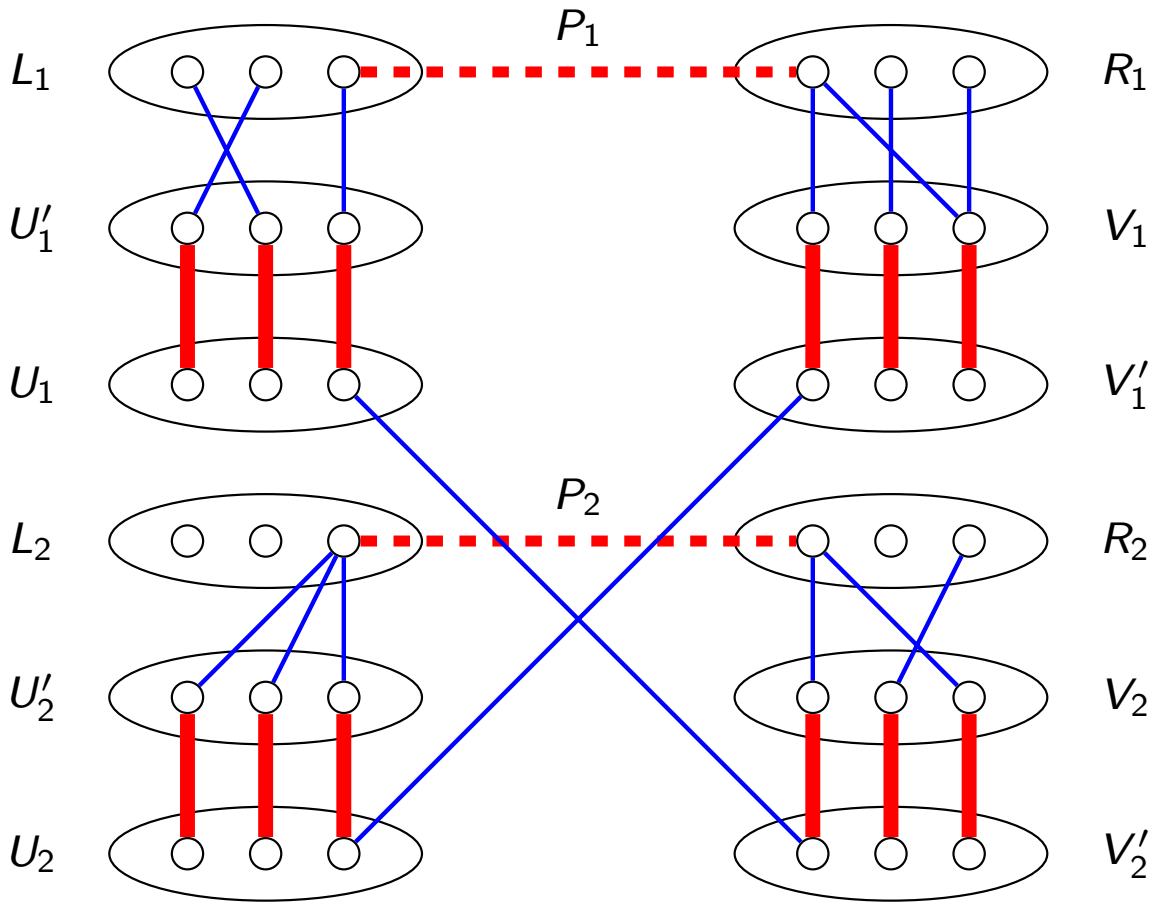
1. Let U'_k be set of reserved vertices connecting to L_k
   Let V_k be set of reserved vertices connecting to R_k



Two-stage cycle-finding scheme

Stage 2 (sprinkling):

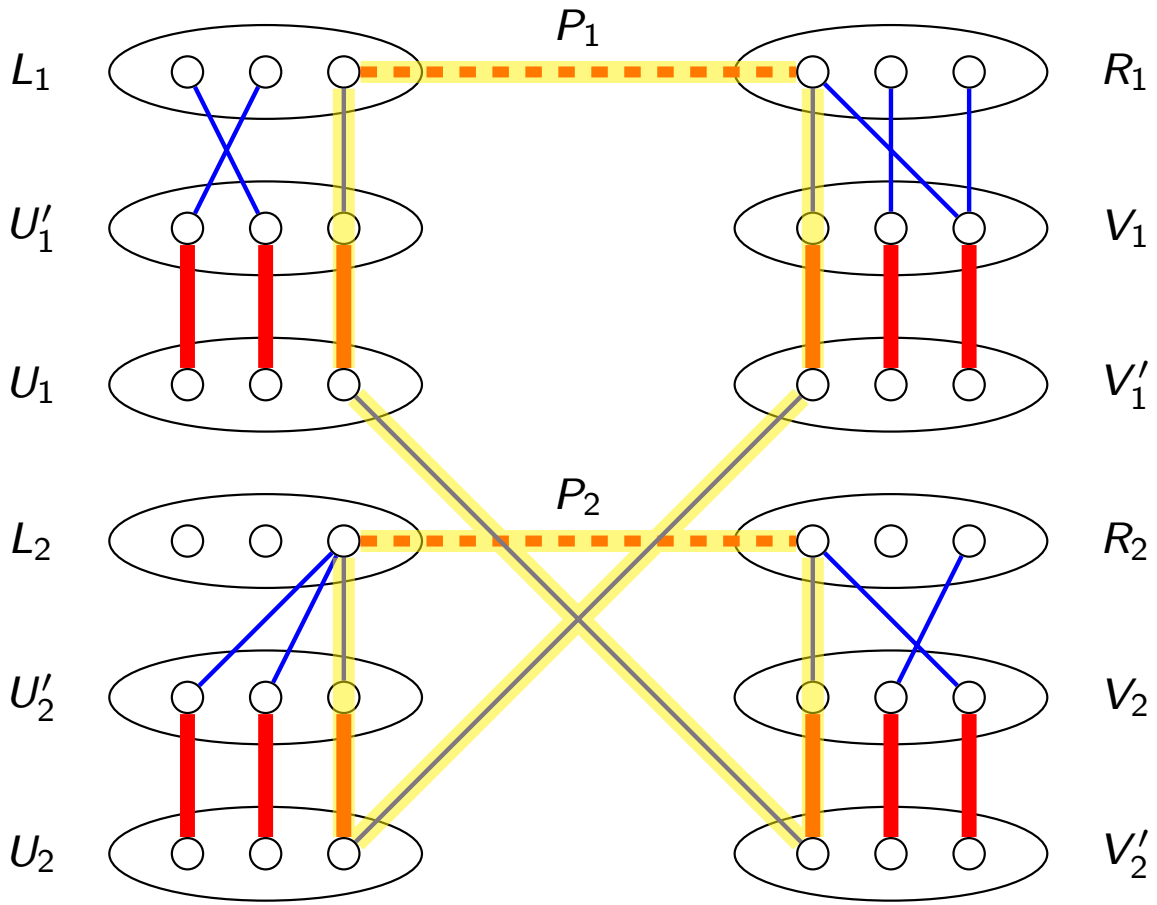
1. Let U'_k be set of reserved vertices connecting to L_k
   Let V_k be set of reserved vertices connecting to R_k
2. Find blue edges connecting $\{U_k\}, \{V'_k\}$.



Two-stage cycle-finding scheme

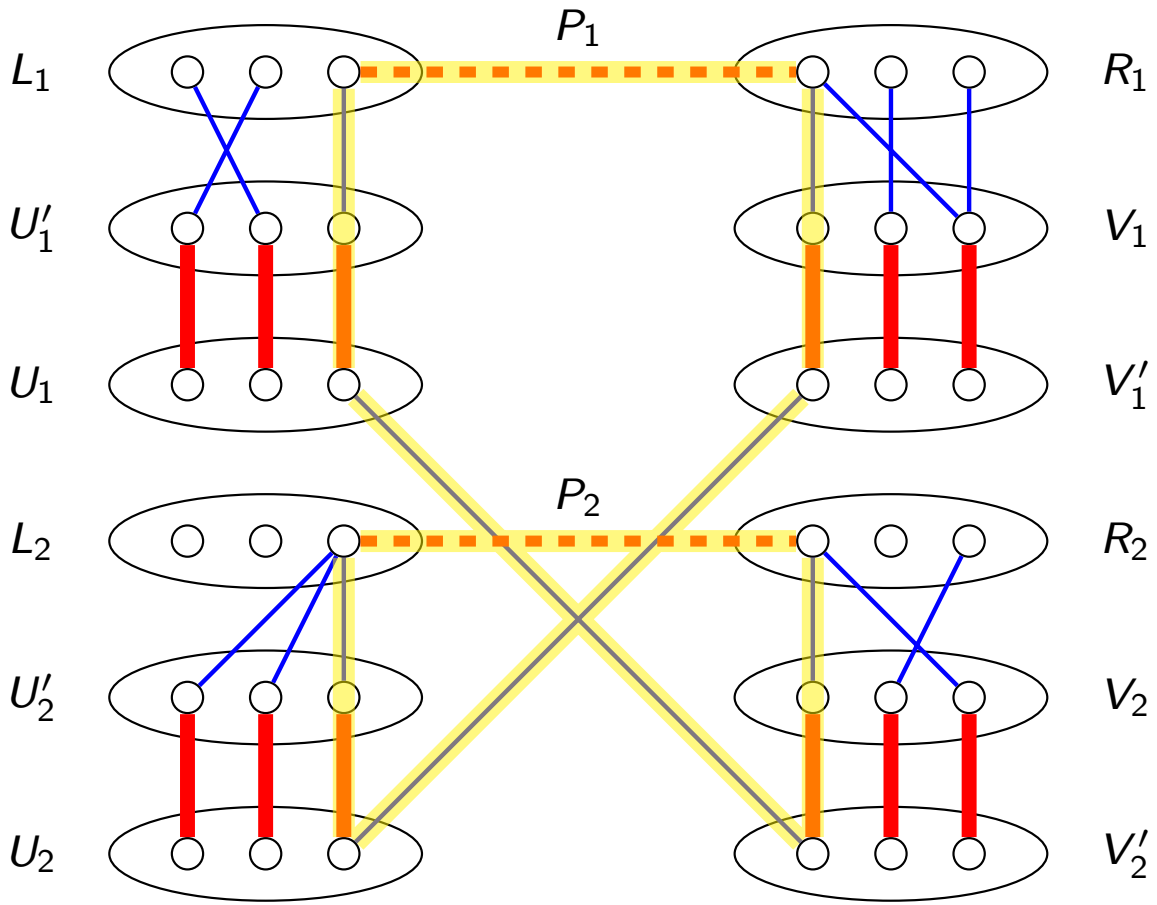
Stage 2 (sprinkling):

1. Let U'_k be set of reserved vertices connecting to L_k
   Let V_k be set of reserved vertices connecting to R_k
2. Find blue edges connecting $\{U_k\}, \{V'_k\}$.



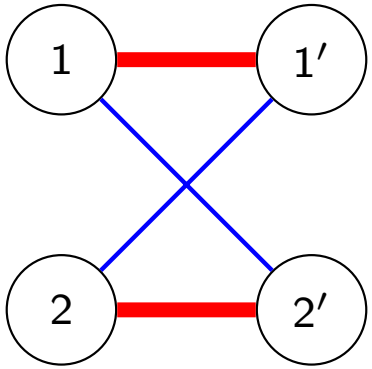
Two-stage cycle-finding scheme

Super graph: Define G_{super} on $[K] \times [K]'$, such that (k, k') is a red edge for all k, and (i, j') is a blue edge iff U_i and V'_j is connected by at least one blue edge.



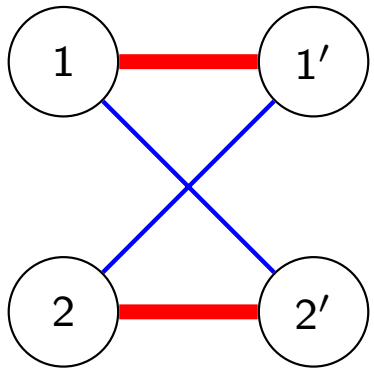
Two-stage cycle-finding scheme

Super graph: Define G_{super} on $[K] \times [K]'$, such that (k, k') is a red edge for all k, and (i, j') is a blue edge iff U_i and V'_j is connected by at least one blue edge.



Two-stage cycle-finding scheme

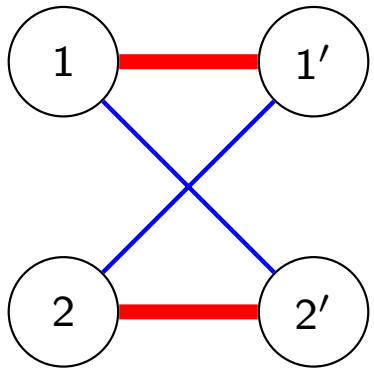
Super graph: Define G_{super} on $[K] \times [K]'$, such that (k, k') is a red edge for all k, and (i, j') is a blue edge iff U_i and V'_j is connected by at least one blue edge.



1. Each alternating cycle on G_{super} expands into an augmenting alternating cycle in G

Two-stage cycle-finding scheme

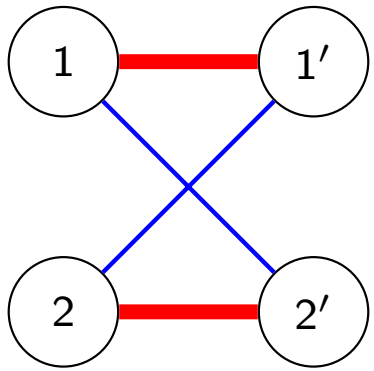
Super graph: Define G_{super} on $[K] \times [K]'$, such that (k, k') is a red edge for all k, and (i, j') is a blue edge iff U_i and V_j' is connected by at least one blue edge.



1. Each alternating cycle on G_{super} expands into an augmenting alternating cycle in G
2. G_{super} is a very supercritical Erdős-Rényi bipartite graph with a planted perfect matching

Two-stage cycle-finding scheme

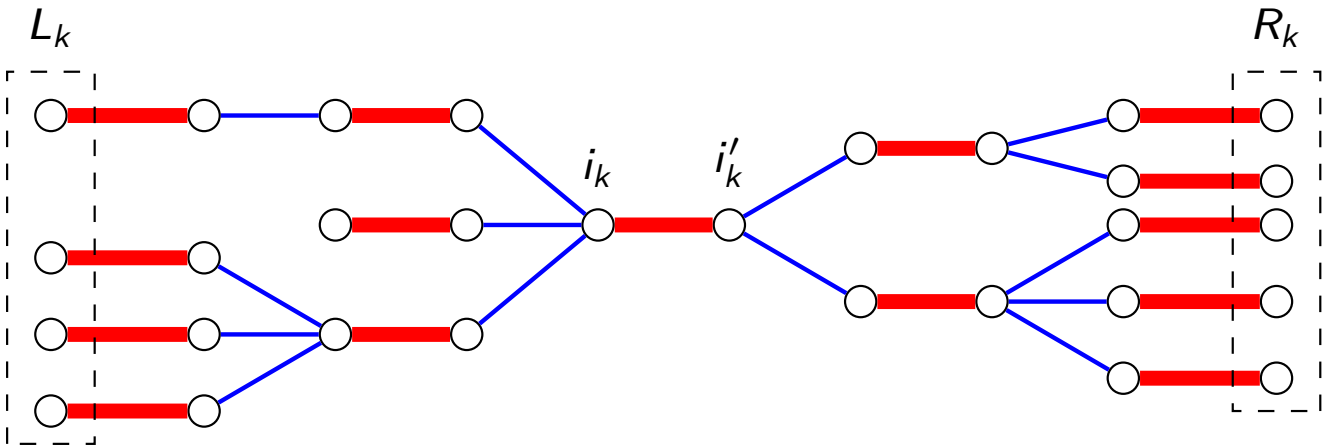
Super graph: Define G_{super} on $[K] \times [K]'$, such that (k, k') is a red edge for all k, and (i, j') is a blue edge iff U_i and V_j' is connected by at least one blue edge.



1. Each alternating cycle on G_{super} expands into an augmenting alternating cycle in G
2. G_{super} is a very supercritical Erdős-Rényi bipartite graph with a planted perfect matching
3. G_{super} contains $e^{\Omega(K)} = e^{\Omega(n)}$ alternating cycles of length $\Omega(K) = \Omega(n)$ (standard DFS argument [Krivelevich-Lee-Sudakov '13])

Path construction via neighborhood exploration process

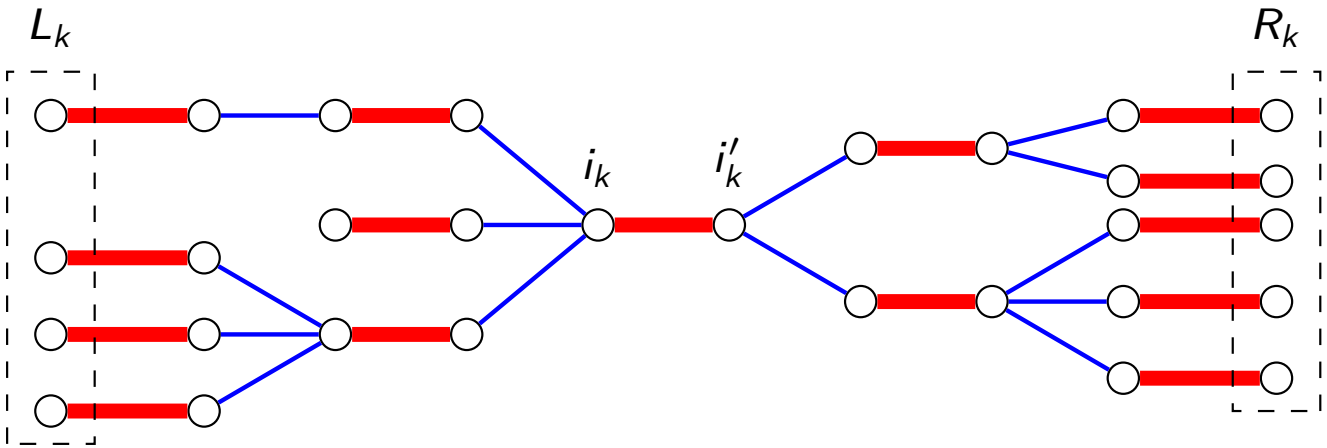
Two-sided tree:



Starting from i_k, grow a tree, remove the inspected vertices, and then grow another tree from i'_k

Path construction via neighborhood exploration process

Two-sided tree:

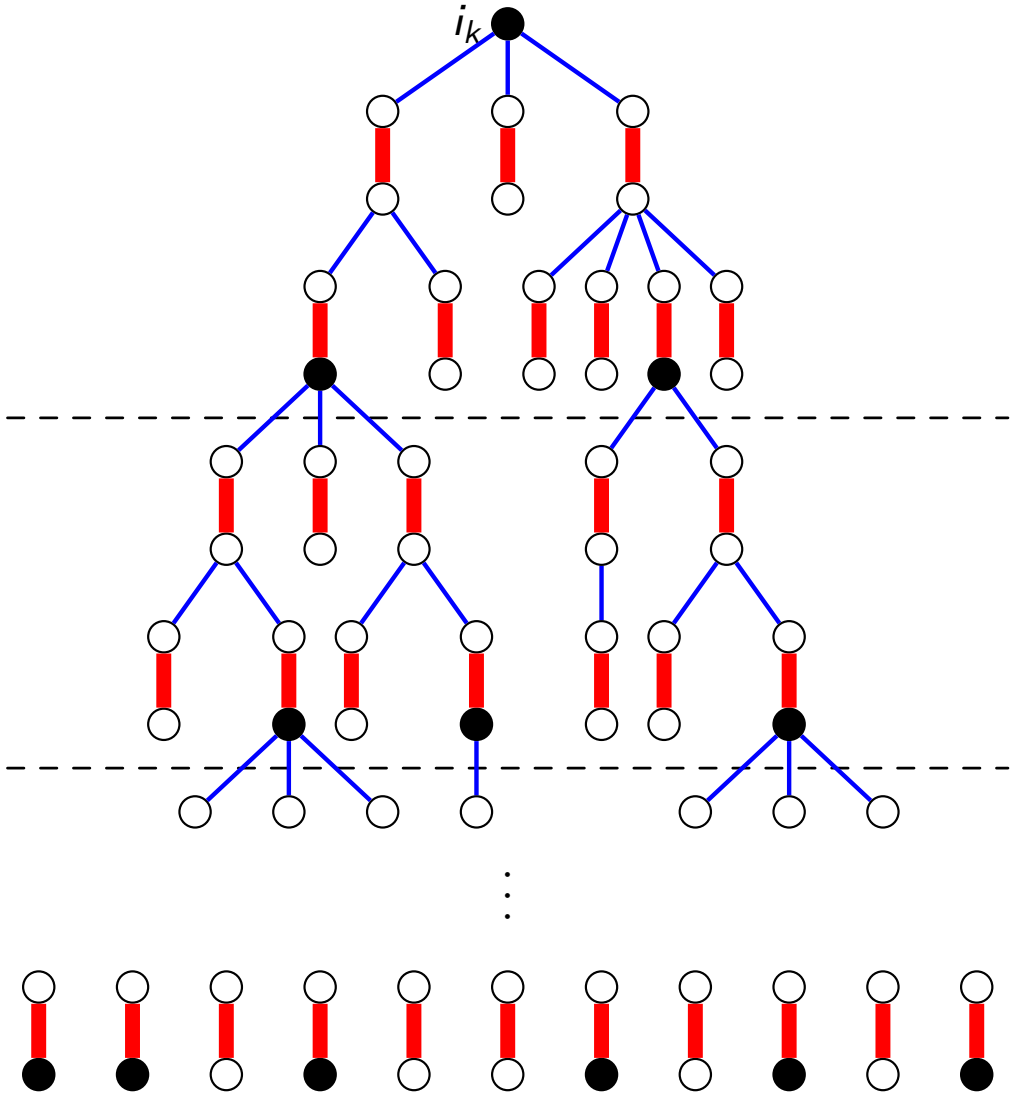


Starting from i_k, grow a tree, remove the inspected vertices, and then grow another tree from i'_k

Key challenge

How to ensure large L_k, R_k connected via augmenting alternating paths. while not using up too many vertices?

Exploration + selection



- Explore via BFS in epochs, each epoch has H steps
- At the end of each epoch, select leaves whose paths to root are augmenting and continue growing
- Behaves as a branching process with average number of offsprings $\left(dB^2(\mathcal{P}, \mathcal{Q})\right)^H > 1$

Analysis

- Proof of positive result via maximum likelihood
- Proof of negative result via analyzing posterior distribution
  - Two-stage cycle finding (path construction + sprinkling)
  - Path construction via neighborhood tree exploration process
- Proof of tight error lower bound under exponential model

Tight error lower bound under exponential model

- Recall exponential model: $d = n$, $\mathcal{P} = \exp(\lambda)$, $\mathcal{Q} = \exp(1/n)$
- Follow the two-stage cycle finding scheme
- However, the tree-based path construction is too wasteful (construct a fat tree, but ultimately uses one path)

Tight error lower bound under exponential model

- Recall exponential model: $d = n$, $\mathcal{P} = \exp(\lambda)$, $\mathcal{Q} = \exp(1/n)$
- Follow the two-stage cycle finding scheme
- However, the tree-based path construction is too wasteful (construct a fat tree, but ultimately uses one path)

Improved Path construction (exponential model)

1. Directly show the existence of many short augmenting alternating paths using first and second moment method
2. Extract a large collection of vertex-disjoint paths via Turán's Theorem

Follow the program in [Ding-Goswami '15] in a different context

First and second moment under exponential model

- Log-likelihood weight $\log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$ is scale-and-shift of $-W_e$:

  Alternating path P is augmenting $\Leftrightarrow \mathrm{wt}_{\mathsf{r}}(P) \geq \mathrm{wt}_{\mathsf{b}}(P)$

First and second moment under exponential model

- Log-likelihood weight $\log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$ is scale-and-shift of $-W_e$:

  Alternating path P is augmenting $\Leftrightarrow \mathsf{wt}_{\mathsf{r}}(P) \geq \mathsf{wt}_{\mathsf{b}}(P)$

- Separately control the total red and blue edge weights of P:

$$\mathsf{wt}_{\mathsf{r}}(P) \approx \frac{2}{\lambda} \cdot |\mathsf{r}(P)|, \quad \mathsf{wt}_{\mathsf{b}}(P) \approx \frac{2-\epsilon}{\lambda} \cdot |\mathsf{b}(P)|$$

First and second moment under exponential model

- Log-likelihood weight $\log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$ is scale-and-shift of $-W_e$:

  Alternating path P is augmenting $\Leftrightarrow \mathrm{wt_r}(P) \geq \mathrm{wt_b}(P)$

- Separately control the total red and blue edge weights of P:

$$\mathrm{wt_r}(P) \approx \frac{2}{\lambda} \cdot |\mathrm{r}(P)|, \quad \mathrm{wt_b}(P) \approx \frac{2-\epsilon}{\lambda} \cdot |\mathrm{b}(P)|$$

- Further need *uniformity* [Ding '13, Ding-Sun-Wilson'15] to reduce correlations among different P:

  Deviation of $\mathrm{wt_r}(Q)$ and $\mathrm{wt_b}(Q)$ in every subpath Q is $O(\frac{1}{\sqrt{\epsilon}})$

First and second moment under exponential model

- Log-likelihood weight $\log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$ is scale-and-shift of $-W_e$:

  Alternating path P is augmenting $\Leftrightarrow \mathrm{wt_r}(P) \geq \mathrm{wt_b}(P)$

- Separately control the total red and blue edge weights of P:

$$\mathrm{wt_r}(P) \approx \frac{2}{\lambda} \cdot |\mathrm{r}(P)|, \quad \mathrm{wt_b}(P) \approx \frac{2-\epsilon}{\lambda} \cdot |\mathrm{b}(P)|$$

- Further need *uniformity* [Ding '13, Ding-Sun-Wilson'15] to reduce correlations among different P:

  Deviation of $\mathrm{wt_r}(Q)$ and $\mathrm{wt_b}(Q)$ in every subpath Q is $O(\frac{1}{\sqrt{\epsilon}})$

- Let S_ℓ denote the set of such alternating paths of length ℓ:

$$\mathrm{Var}(|S_\ell|) \leq (\mathbb{E}[|S_\ell|])^2 \frac{\ell^2 e^{\Theta(1/\sqrt{\epsilon})}}{n}$$

Extract vertex-disjoint alternating paths via Turán

- Define graph H
  - Vertex: alternating path in S_ℓ
  - Edge: if two alternating paths share common vertices
- Independent set $\Leftrightarrow$ collection of vertex-disjoint alternating paths

Extract vertex-disjoint alternating paths via Turán

- Define graph H
  - Vertex: alternating path in S_ℓ
  - Edge: if two alternating paths share common vertices
- Independent set $\Leftrightarrow$ collection of vertex-disjoint alternating paths

Turán's Theorem

Let $H = (V, E)$ be any simple graph. Then H contains an independent subset of size at least $|V|^2/(2|E| + |V|)$.

Extract vertex-disjoint alternating paths via Turán

- Define graph H
  - Vertex: alternating path in S_ℓ
  - Edge: if two alternating paths share common vertices
- Independent set $\Leftrightarrow$ collection of vertex-disjoint alternating paths

Turán's Theorem

Let $H = (V, E)$ be any simple graph. Then H contains an independent subset of size at least $|V|^2/(2|E| + |V|)$.

- Apply Turán's Theorem with $|V| \approx \mathbb{E}[|S_\ell|]$ and $|E| \approx \mathrm{Var}(|S_\ell|)$

$\Rightarrow$ There exist $\dfrac{n}{\ell^2 e^{\Theta(1/\sqrt{\epsilon})}}$ vertex-disjoint alternating paths of length ℓ

Extract vertex-disjoint alternating paths via Turán

- Define graph H
  - Vertex: alternating path in S_ℓ
  - Edge: if two alternating paths share common vertices
- Independent set $\Leftrightarrow$ collection of vertex-disjoint alternating paths

Turán's Theorem

Let $H = (V, E)$ be any simple graph. Then H contains an independent subset of size at least $|V|^2/(2|E| + |V|)$.

- Apply Turán's Theorem with $|V| \approx \mathbb{E}[|S_\ell|]$ and $|E| \approx \mathrm{Var}(|S_\ell|)$

  $\Rightarrow$ There exist $\dfrac{n}{\ell^2 e^{\Theta(1/\sqrt{\epsilon})}}$ vertex-disjoint alternating paths of length ℓ

- Choose $\ell = e^{\Theta(1/\sqrt{\epsilon})}$ and get desired augmenting alternating cycles of length $ne^{-\Theta(1/\sqrt{\epsilon})}$ via sprinkling

Conclusion

- Sharp threshold for almost perfect recovery: $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) = 1$
- Infinite-order phase transition under the exponential model: Optimal reconstruction error is $\exp\left(-\Theta(1/\sqrt{\epsilon})\right)$ when $\lambda = 4 - \epsilon$
- Key idea: two-stage cycle finding (path construction + sprinkling)
- Under exponential model, further need to impose uniformity

Conclusion

- Sharp threshold for almost perfect recovery: $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) = 1$
- Infinite-order phase transition under the exponential model: Optimal reconstruction error is $\exp\left(-\Theta(1/\sqrt{\epsilon})\right)$ when $\lambda = 4 - \epsilon$
- Key idea: two-stage cycle finding (path construction + sprinkling)
- Under exponential model, further need to impose uniformity

Open problems:

1. Optimal error for general distributions? in entire parameter range?
2. Extension to planted k-factor model?
   Conjecture: $\sqrt{kd}\, B(\mathcal{P}, \mathcal{Q}) = 1$ [Sicuro-Zdeborová '20]

Conclusion

- Sharp threshold for almost perfect recovery: $\sqrt{d}\, B(\mathcal{P}, \mathcal{Q}) = 1$
- Infinite-order phase transition under the exponential model: Optimal reconstruction error is $\exp\left(-\Theta(1/\sqrt{\epsilon})\right)$ when $\lambda = 4 - \epsilon$
- Key idea: two-stage cycle finding (path construction + sprinkling)
- Under exponential model, further need to impose uniformity

Open problems:

1. Optimal error for general distributions? in entire parameter range?
2. Extension to planted k-factor model?
   Conjecture: $\sqrt{kd}\, B(\mathcal{P}, \mathcal{Q}) = 1$ [Sicuro-Zdeborová '20]

Reference

D. Jian, Y. Wu, J. Xu, & D. Yang *The planted matching problem: Sharp threshold and infinite-order phase transition*. arXiv:2103.09383.