All-Something-Nothing Phase Transitions in Planted Subgraph Recovery

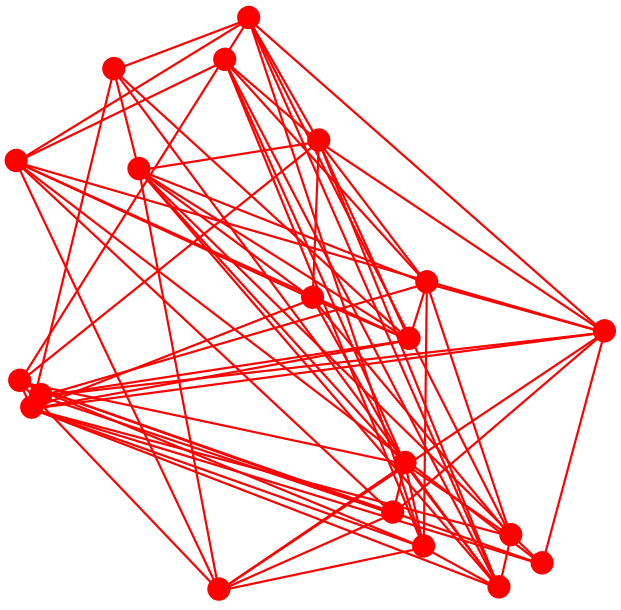
Jiaming Xu

The Fuqua School of Business
Duke University

Joint work with
Julia Gaudio (Northwestern), Colin Sandon (EPFL), Dana Yang (Cornell)

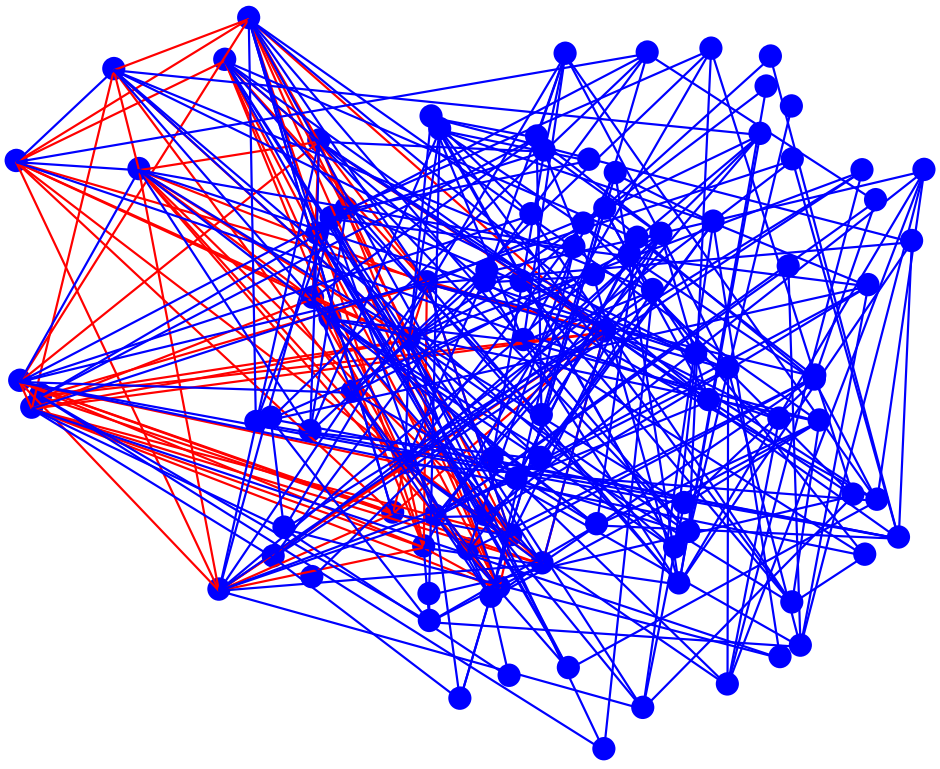
October 16, 2023
INFORMS Annual Meeting

The Planted subgraph recovery problem



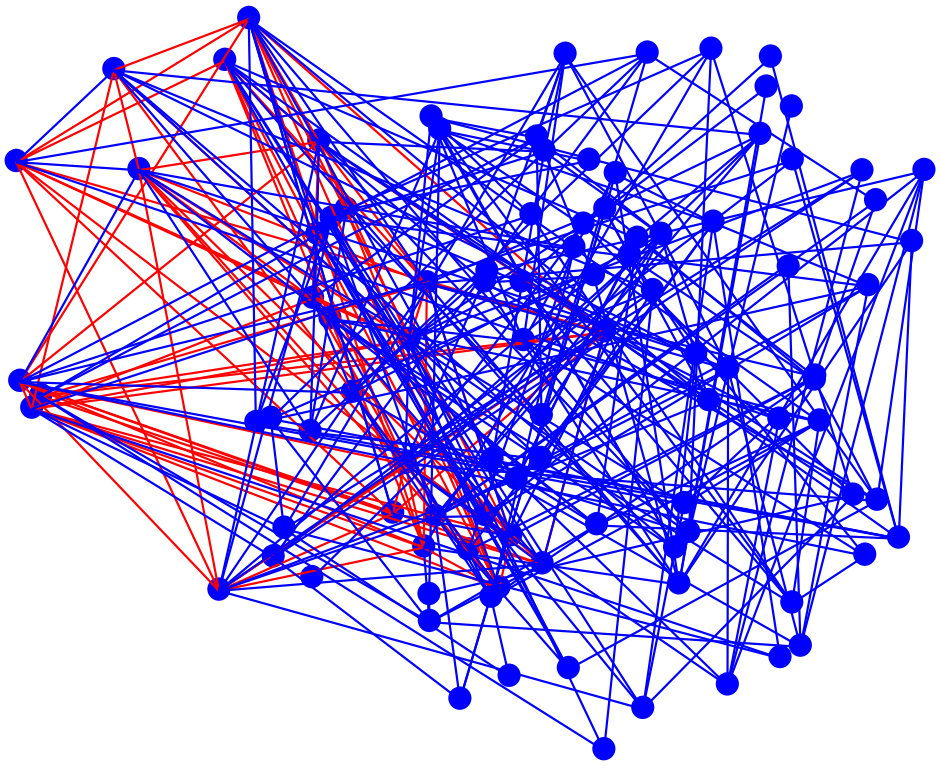
- A hidden subgraph H^*

The Planted subgraph recovery problem



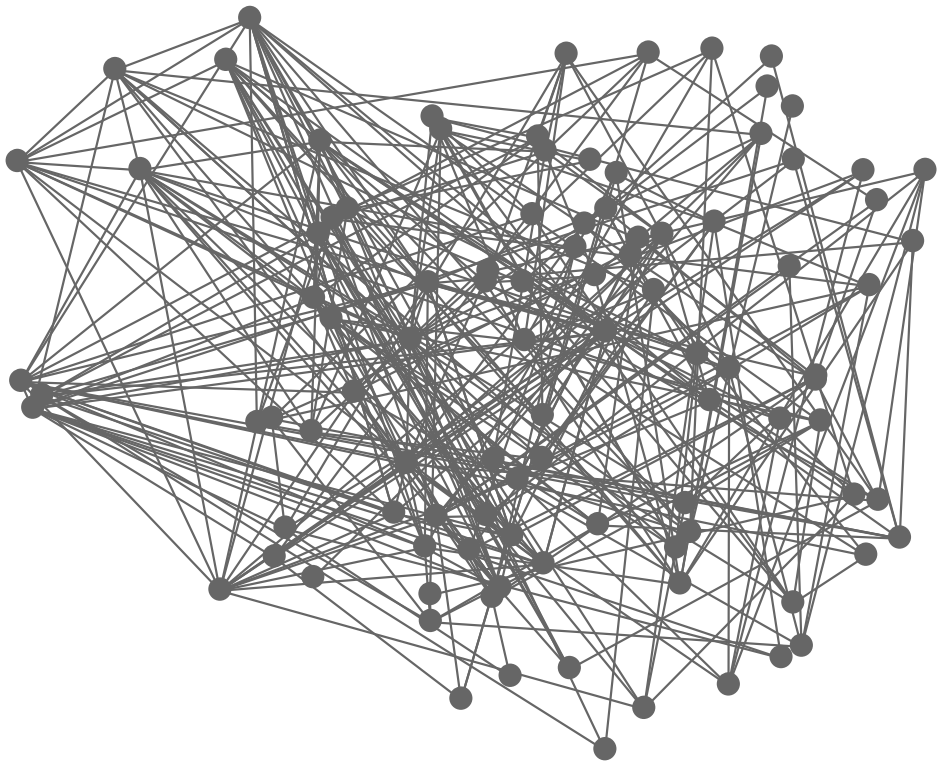
- A hidden subgraph H^*
- A background graph $G_0 \sim \mathcal{G}(n, p)$

The Planted subgraph recovery problem



- A hidden subgraph H^*
- A background graph $G_0 \sim \mathcal{G}(n, p)$
- Observe union graph $G = H^* \cup G_0$

The Planted subgraph recovery problem



- A hidden subgraph H^*
- A background graph $G_0 \sim \mathcal{G}(n, p)$
- Observe union graph $G = H^* \cup G_0$
- Goal: recover H^* from G

Encompasses many planted problems...

- Planted clique model
- Planted tree model [Massoulié-Stephan-Towsley '18]
- Planted Hamiltonian cycle (TSP) model
  [Bagaria-Ding-Tse-Wu-X.'18]
- Planted k-NN graph model [Ding-Wu-X.-Yang '19]
- Planted matching [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10]
- ...

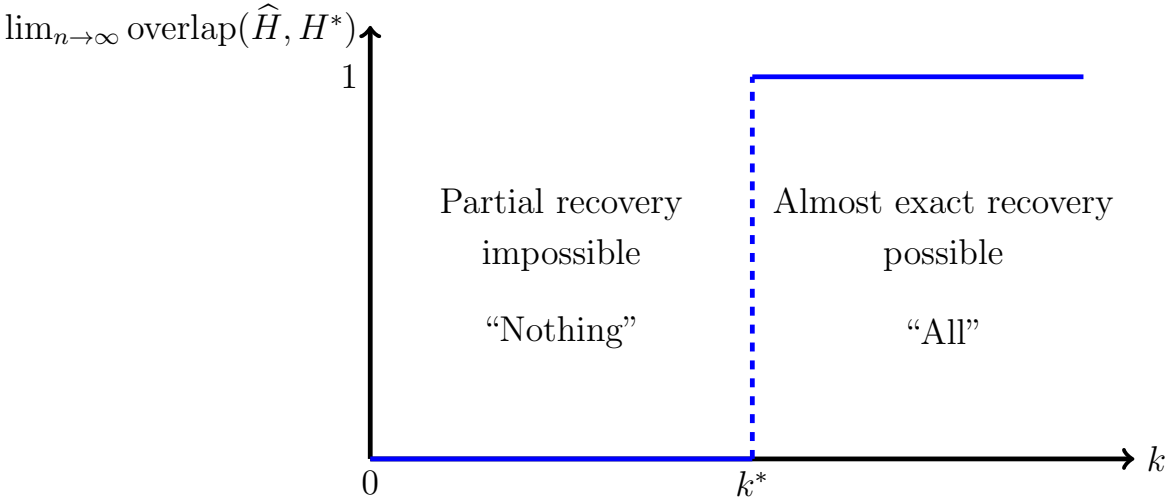
Encompasses many planted problems...

- Planted clique model
- Planted tree model [Massoulié-Stephan-Towsley '18]
- Planted Hamiltonian cycle (TSP) model
  [Bagaria-Ding-Tse-Wu-X.'18]
- Planted k-NN graph model [Ding-Wu-X.-Yang '19]
- Planted matching [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10]
- ...

Fruitful venue for studying statistical and computational aspects of network inference

Peculiar "All-or-Nothing" phase transitions

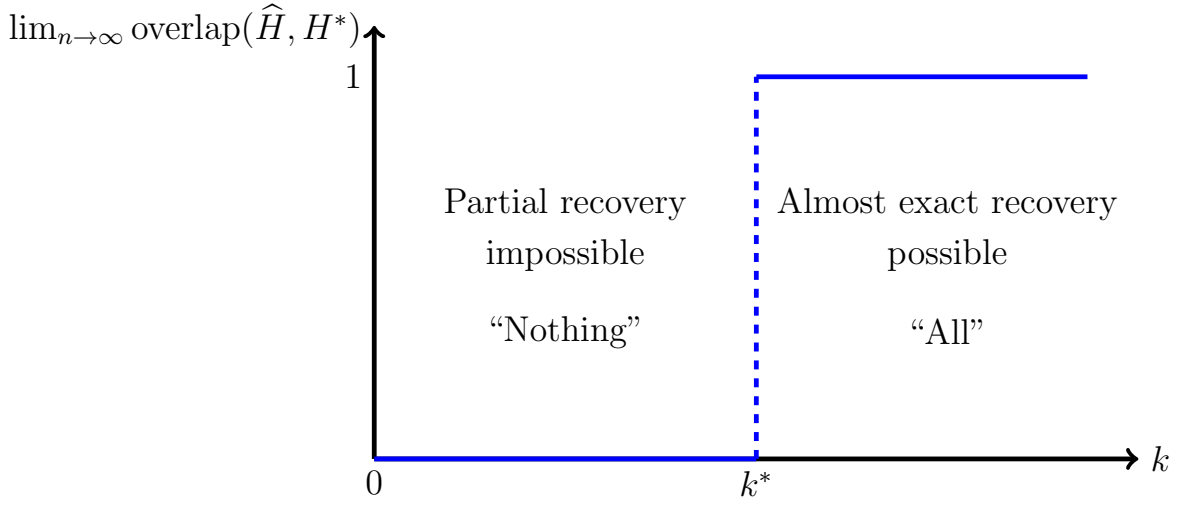
For both planted k-clique and k-tree model:



$$\text{overlap}(H, H^*) \triangleq \frac{|H \cap H^*|}{|H^*|}$$

Peculiar "All-or-Nothing" phase transitions

For both planted k-clique and k-tree model:



$$\text{overlap}(H, H^*) \triangleq \frac{|H \cap H^*|}{|H^*|}$$

$$k^* = \begin{cases} 2\log_2(n) & \text{planted clique in } G_0 \sim \mathcal{G}(n, 1/2) \\ \log_{1/\lambda}(n) & \text{planted tree in } G_0 \sim \mathcal{G}(n, \lambda/n) \end{cases}$$

coincides with the first-moment threshold k_{1M} in G_0,

at which the expected # of copies in G_0 equals 1

"All-or-Nothing" phase transitions are prevalent

Theorem (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

The planted subgraph recovery model exhibits AoN at p_{1M}, if

- *H is sufficiently dense and balanced:*

$$e(H) \gg v(H) \log v(H) \text{ and } \frac{e(H)}{v(H)} \geq \frac{e(J)}{v(J)}, \forall J \subset H$$

"All-or-Nothing" phase transitions are prevalent

Theorem (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

The planted subgraph recovery model exhibits AoN at p_{1M}, if

- *H is sufficiently dense and balanced:*

$$e(H) \gg v(H) \log v(H) \text{ and } \frac{e(H)}{v(H)} \geq \frac{e(J)}{v(J)}, \forall J \subset H$$

- *H is sufficiently small and strictly balanced for $c > 0$:*

$$e(H) + v(H) \leq \frac{c \log n}{3 \log \log n} \text{ and } \frac{e(H) - c}{v(H)} \geq \frac{e(J) - c}{v(J)}, \forall J \subset H$$

"All-or-Nothing" phase transitions are prevalent

Theorem (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

The planted subgraph recovery model exhibits AoN at p_{1M}, if

- *H is sufficiently dense and balanced:*

$$e(H) \gg v(H) \log v(H) \text{ and } \frac{e(H)}{v(H)} \geq \frac{e(J)}{v(J)}, \forall J \subset H$$

- *H is sufficiently small and strictly balanced for $c > 0$:*

$$e(H) + v(H) \leq \frac{c \log n}{3 \log \log n} \text{ and } \frac{e(H) - c}{v(H)} \geq \frac{e(J) - c}{v(J)}, \forall J \subset H$$

AoN was also established for many other models: sparse linear regression, sparse tensor PCA, group testing, graph alignment, ...

Focus of this talk

Question

Is AoN universal in planted subgraph recovery?

Focus of this talk

Question

Is AoN universal in planted subgraph recovery?

Consider large, sparse, and balanced graphs

Planted factor model [Sicuro-Zdeborová '20]

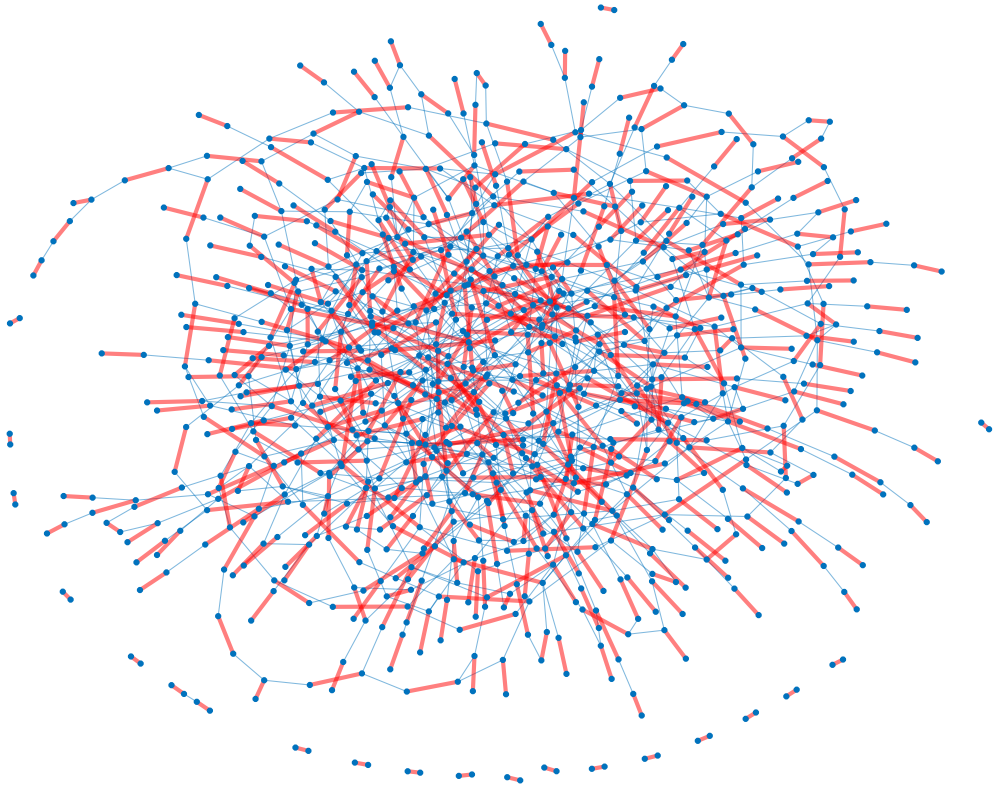
H is uniformly chosen from all labeled k-factors (spanning k-regular graphs) for a fixed constant k:

- $k = 1$: perfect matching
- $k = 2$: disjoint union of cycles (including Hamiltonian cycles)

Planted factor model [Sicuro-Zdeborová '20]

H is uniformly chosen from all labeled k-factors (spanning k-regular graphs) for a fixed constant k:

- $k = 1$: perfect matching
- $k = 2$: disjoint union of cycles (including Hamiltonian cycles)

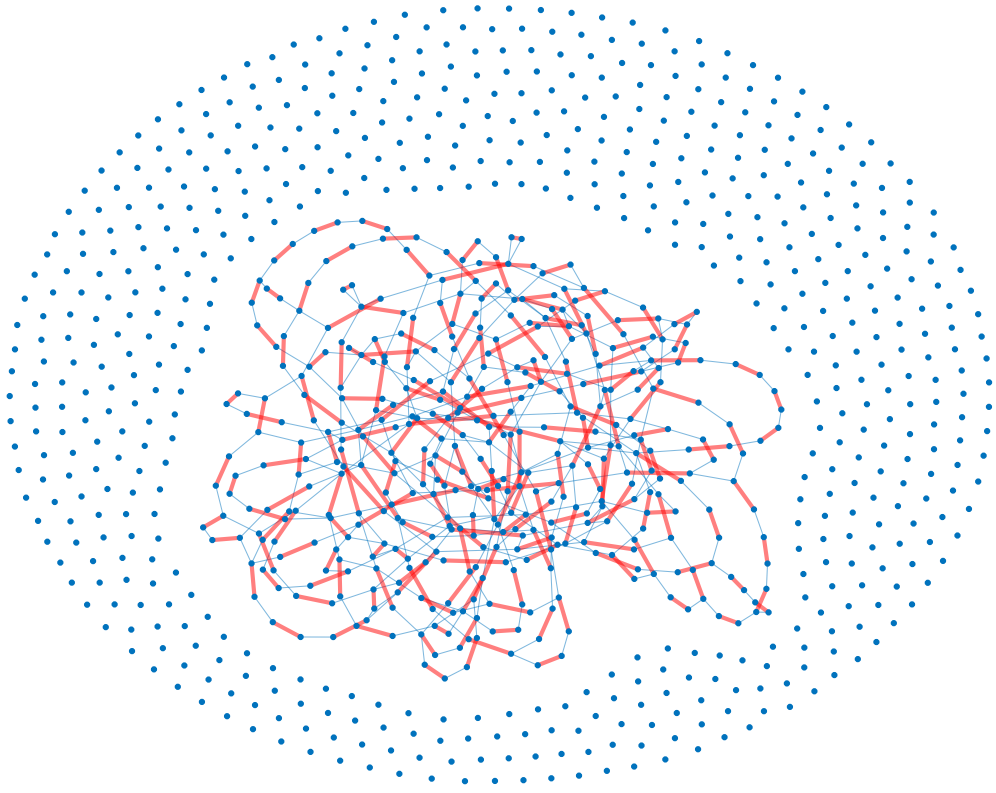


Planted matching model with $n = 1000$ and $\lambda = 1.5$

Planted factor model [Sicuro-Zdeborová '20]

H is uniformly chosen from all labeled k-factors (spanning k-regular graphs) for a fixed constant k:

- $k = 1$: perfect matching
- $k = 2$: disjoint union of cycles (including Hamiltonian cycles)



Planted matching model with $n = 1000$ and $\lambda = 1.5$

Posterior distribution under planted factor model

The posterior distribution is uniform over all labeled k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G.

Posterior distribution under planted factor model

The posterior distribution is uniform over all labeled k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H\in\mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G.

- Recall λ is the average degree in the background graph G_0

Posterior distribution under planted factor model

The posterior distribution is uniform over all labeled k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G.

- Recall λ is the average degree in the background graph G_0
- When $\lambda = 0$, μ_G is a delta mass on H^*

Posterior distribution under planted factor model

The posterior distribution is uniform over all labeled k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G.

- Recall λ is the average degree in the background graph G_0
- When $\lambda = 0$, μ_G is a delta mass on H^*
- As λ increases, we expect to observe more k-factors in G

Posterior distribution under planted factor model

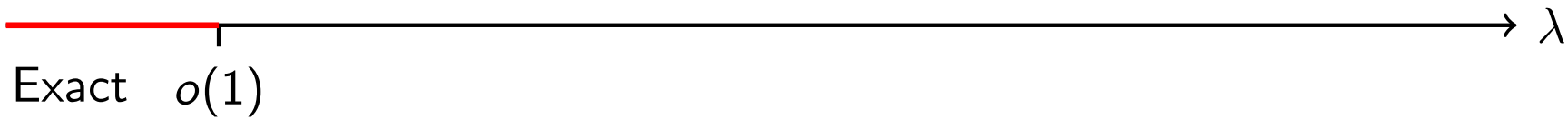
The posterior distribution is uniform over all labeled k-factors in G:

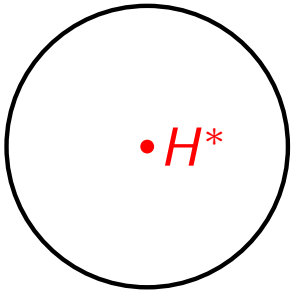
$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H\in\mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G.

- Recall λ is the average degree in the background graph G_0
- When $\lambda = 0$, μ_G is a delta mass on H^*
- As λ increases, we expect to observe more k-factors in G
- But, how μ_G exactly behaves?

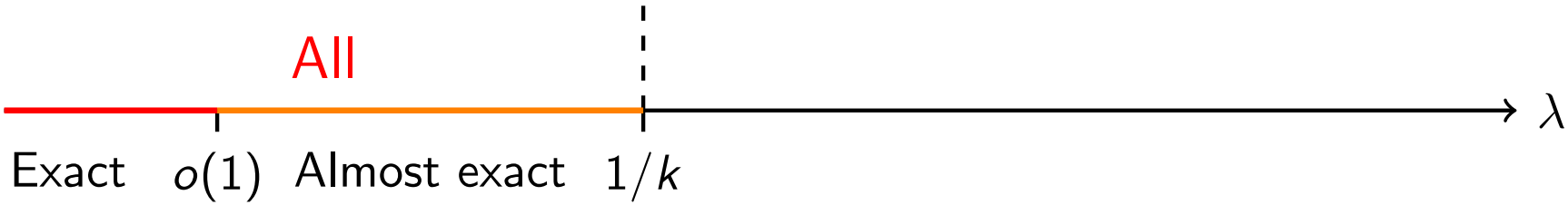
Our result [Gaudio-Sandon-X.-Yang '23]

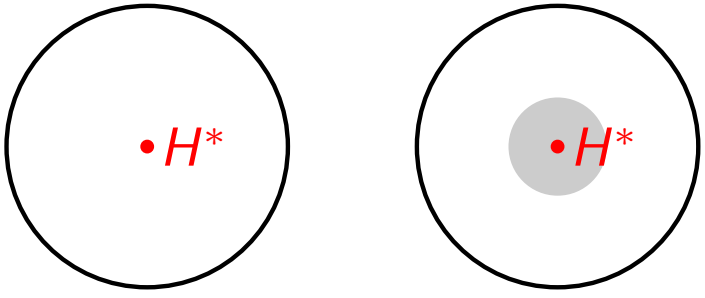




1. H^* is the unique k-factor in G

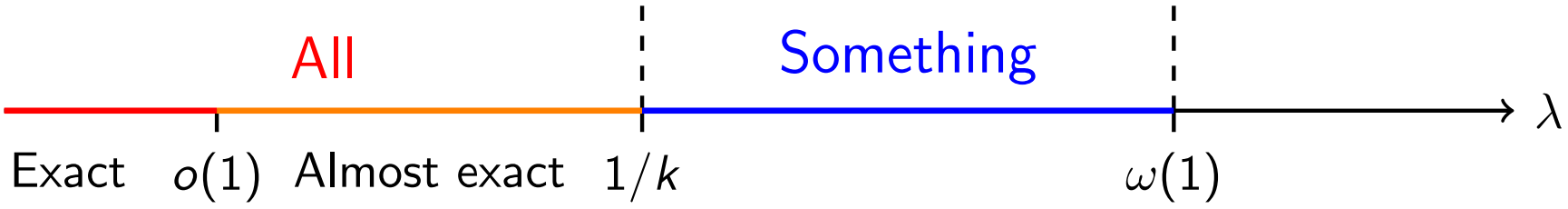
Our result [Gaudio-Sandon-X.-Yang '23]

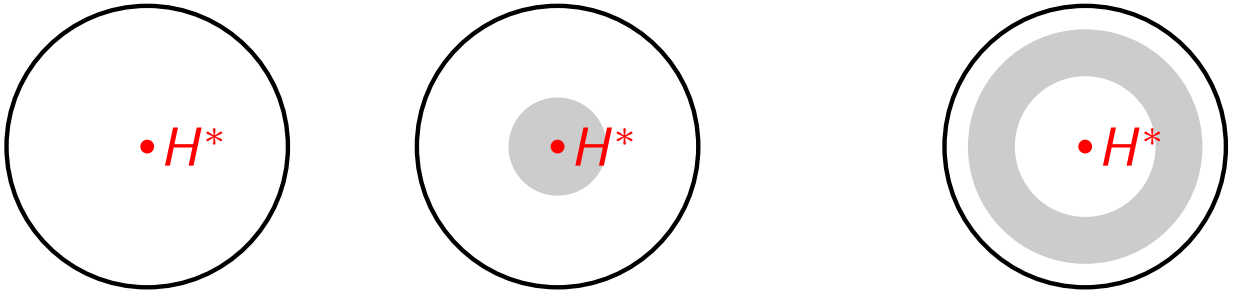




1. H^* is the unique k-factor in G
2. $\text{overlap}(H, H^*) \to 1$ for all $H \in \mathcal{H}(G)$

Our result [Gaudio-Sandon-X.-Yang '23]

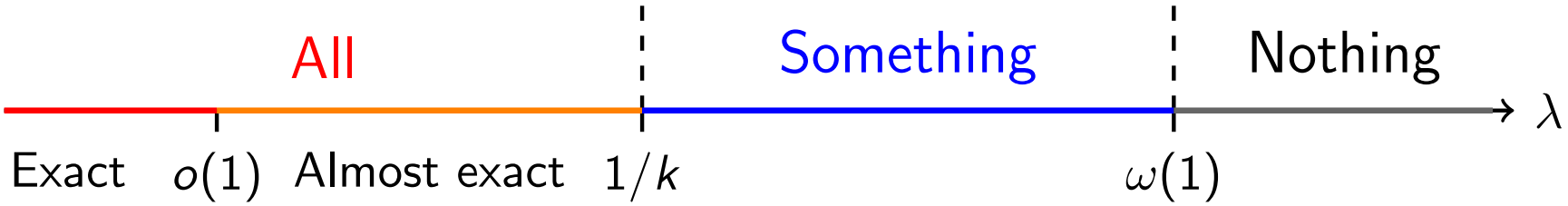


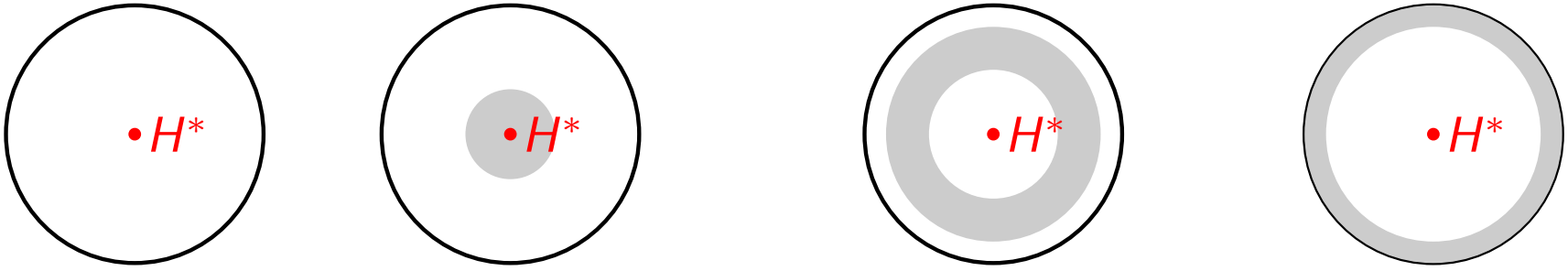


1. H^* is the unique k-factor in G
2. $\text{overlap}(H, H^*) \to 1$ for all $H \in \mathcal{H}(G)$
3. $\text{overlap}(H, H^*) \in [\Omega(1), 1 - \Omega(1)]$ for almost all $H \in \mathcal{H}(G)$

$\lambda = 1/k$ is sharp threshold for almost exact recovery: resolves conjecture in [Sicuro-Zdeborová '20]

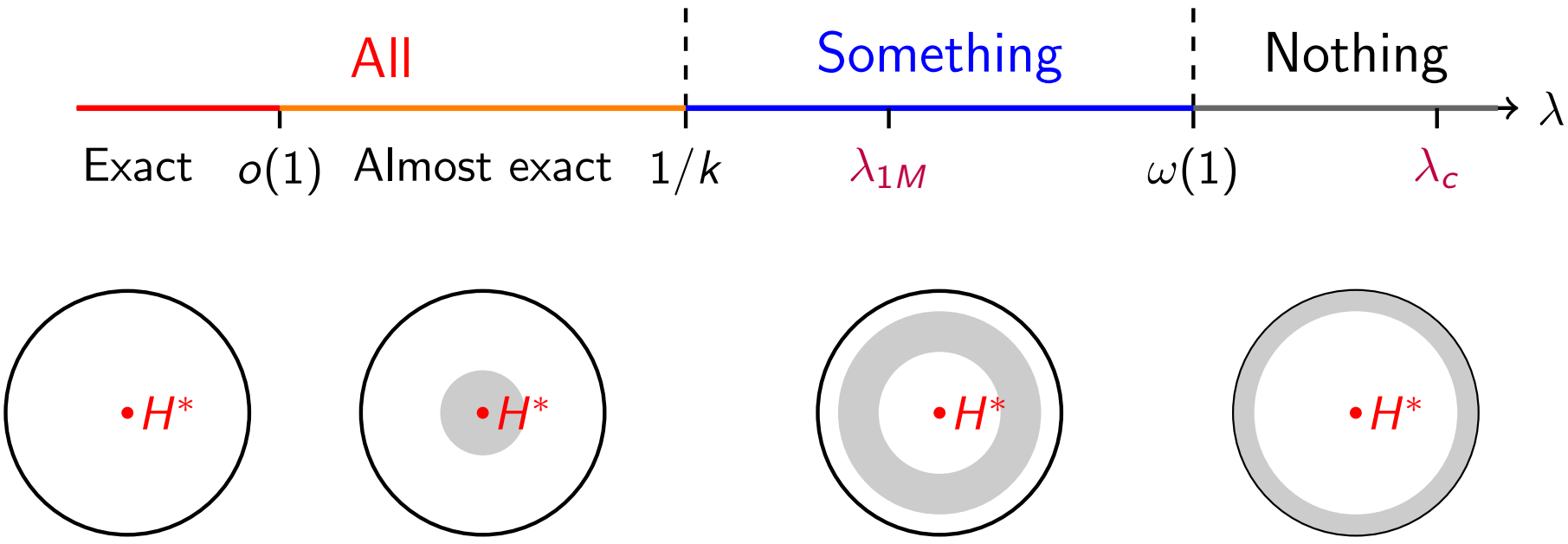
Our result [Gaudio-Sandon-X.-Yang '23]





1. H^* is the unique k-factor in G
2. $\text{overlap}(H, H^*) \to 1$ for all $H \in \mathcal{H}(G)$
3. $\text{overlap}(H, H^*) \in [\Omega(1), 1 - \Omega(1)]$ for almost all $H \in \mathcal{H}(G)$
4. $\text{overlap}(H, H^*) \to 0$ for almost all $H \in \mathcal{H}(G)$

Our result [Gaudio-Sandon-X.-Yang '23]



1. H^* is the unique k-factor in G
2. $\mathrm{overlap}(H, H^*) \to 1$ for all $H \in \mathcal{H}(G)$
3. $\mathrm{overlap}(H, H^*) \in [\Omega(1), 1 - \Omega(1)]$ for almost all $H \in \mathcal{H}(G)$
4. $\mathrm{overlap}(H, H^*) \to 0$ for almost all $H \in \mathcal{H}(G)$

$\lambda_{1M} = e(k!)^{2/k}/k$ and $\lambda_c = \log n + (k-1)\log\log n + \omega(1)$

Proof sketch

- All phase:
  - $\lambda = o(1)$: $\text{overlap}(H, H^*) = 1$
  - $\lambda \leq 1/k$: $\text{overlap}(H, H^*) \to 1$
- Something phase: $1/k < \lambda \leq O(1)$:
  - $\text{overlap}(H, H^*) \leq 1 - \Omega(1)$
  - $\text{overlap}(H, H^*) \geq \Omega(1)$
- Nothing phase: $\lambda = \omega(1)$:
  - $\text{overlap}(H, H^*) \to 0$

Proof sketch

- All phase:
  - $\lambda = o(1)$: $\text{overlap}(H, H^*) = 1$ (first-moment method)
  - $\lambda \leq 1/k$: $\text{overlap}(H, H^*) \to 1$ (first-moment method)
- Something phase: $1/k < \lambda \leq O(1)$:
  - $\text{overlap}(H, H^*) \leq 1 - \Omega(1)$
  - $\text{overlap}(H, H^*) \geq \Omega(1)$
- Nothing phase: $\lambda = \omega(1)$:
  - $\text{overlap}(H, H^*) \to 0$

Proof sketch

- All phase:
  - $\lambda = o(1)$: $\text{overlap}(H, H^*) = 1$ (first-moment method)
  - $\lambda \leq 1/k$: $\text{overlap}(H, H^*) \to 1$ (first-moment method)
- Something phase: $1/k < \lambda \leq O(1)$:
  - $\text{overlap}(H, H^*) \leq 1 - \Omega(1)$
  - $\text{overlap}(H, H^*) \geq \Omega(1)$ ($\exists \Theta(n)$ isolated nodes in G_0)
- Nothing phase: $\lambda = \omega(1)$:
  - $\text{overlap}(H, H^*) \to 0$

Proof sketch

- All phase:
  - $\lambda = o(1)$: $\text{overlap}(H, H^*) = 1$ (first-moment method)
  - $\lambda \leq 1/k$: $\text{overlap}(H, H^*) \to 1$ (first-moment method)
- Something phase: $1/k < \lambda \leq O(1)$:
  - $\text{overlap}(H, H^*) \leq 1 - \Omega(1) \longrightarrow$ Focus
  - $\text{overlap}(H, H^*) \geq \Omega(1)$ ($\exists \Theta(n)$ isolated nodes in G_0)
- Nothing phase: $\lambda = \omega(1)$:
  - $\text{overlap}(H, H^*) \to 0 \longrightarrow$ Focus

Analyzing posterior distribution

Recall: posterior distribution is uniform over k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

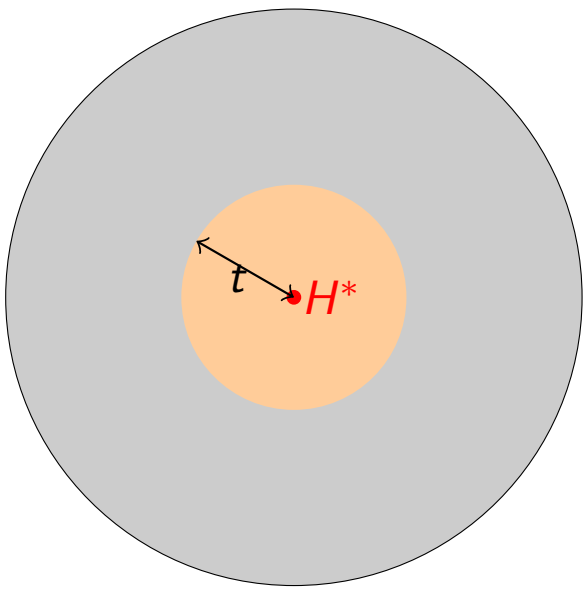
where $\mathcal{H}(G)$ is the set of k-factors in G

Analyzing posterior distribution

Recall: posterior distribution is uniform over k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G



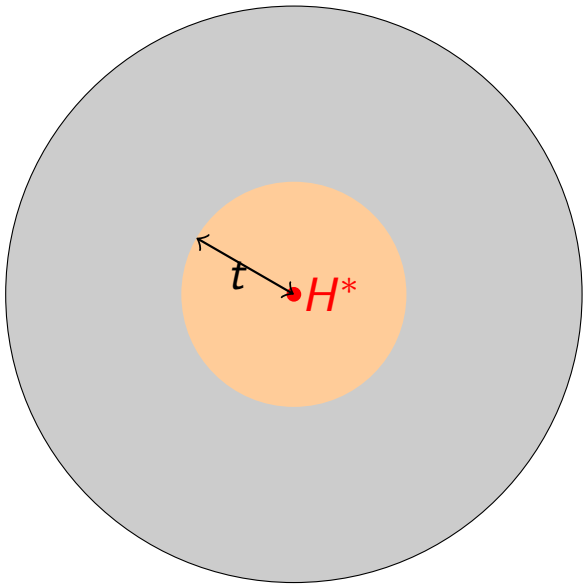
- Upper bound number of k-factors in G near H^*

Analyzing posterior distribution

Recall: posterior distribution is uniform over k-factors in G:

$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G



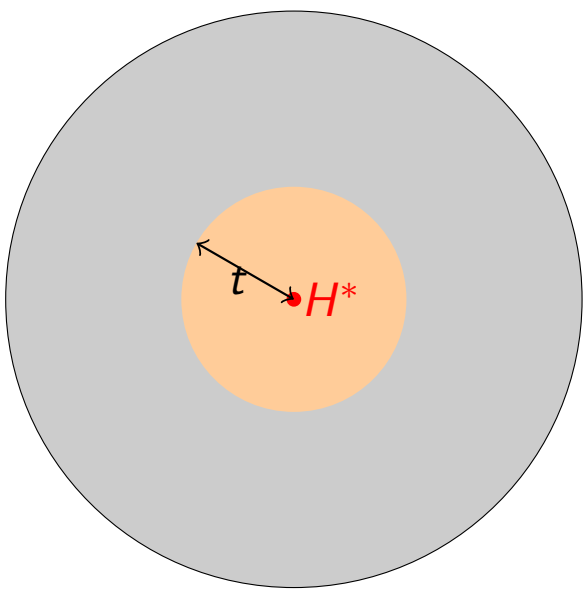
- Upper bound number of k-factors in G near H^*
- Lower bound number of k-factors in G far away from H^*

Analyzing posterior distribution

Recall: posterior distribution is uniform over k-factors in G:

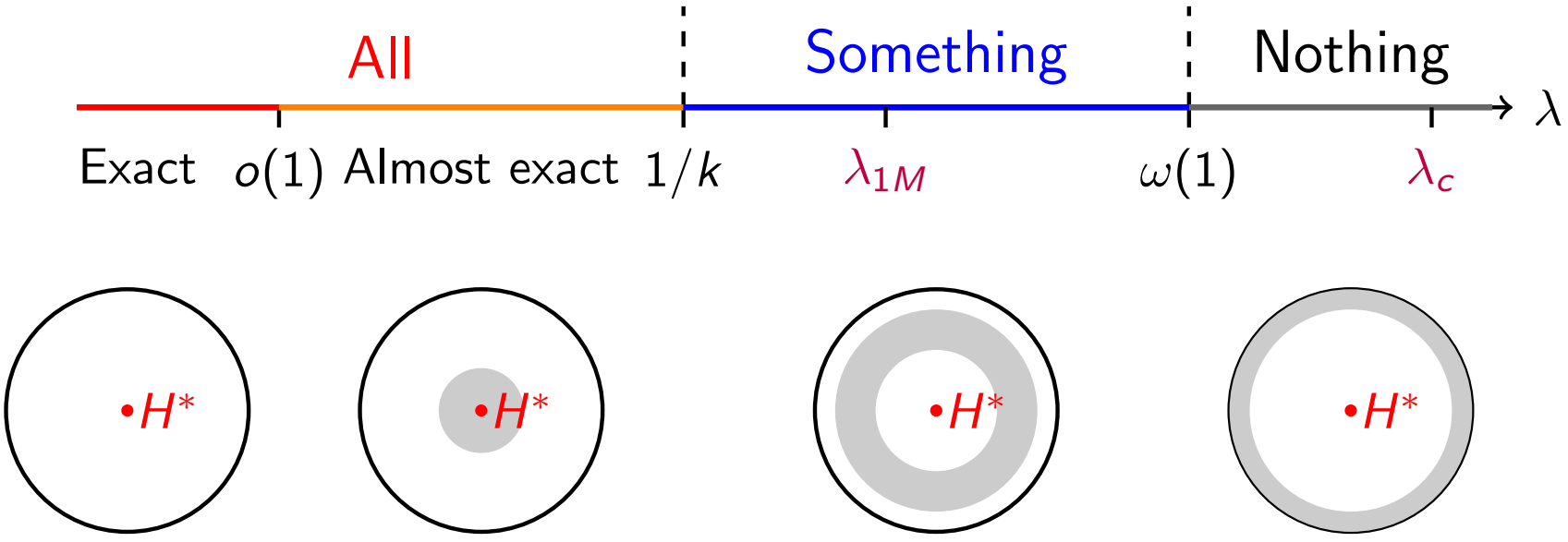
$$\mu_G(H) = \frac{1}{|\mathcal{H}(G)|}\mathbf{1}_{\{H \in \mathcal{H}(G)\}},$$

where $\mathcal{H}(G)$ is the set of k-factors in G

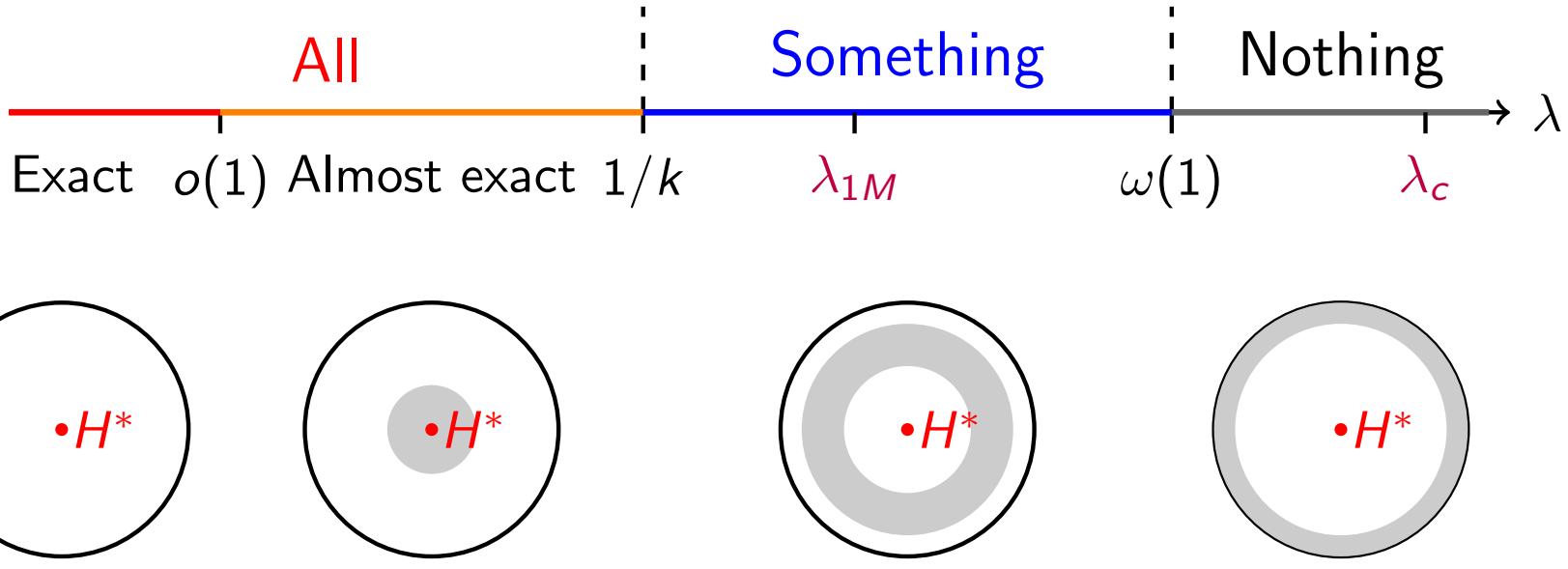


- Upper bound number of k-factors in G near H^*
- Lower bound number of k-factors in G far away from H^*
- $|\mathcal{H}_{\text{near}}(G)| \ll |\mathcal{H}_{\text{far}}(G)| \Rightarrow$ overlap$(H, H^*) \leq 1 - t$
  - Something phase: $t = \Omega(1)$
  - Nothing phase: $t = 1 - o(1)$

Conclusion


All
Something
Nothing
λ
Exact
$o(1)$
Almost exact
$1/k$
λ_{1M}
$\omega(1)$
λ_c
H^*
H^*
H^*
H^*

Conclusion



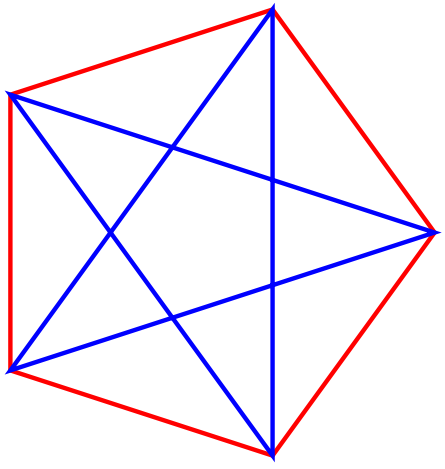
Open problems:

1. Characterize the overlap as a function of λ in something phase
2. Planted k-factor model for growing $k \equiv k(n)$
3. What causes "something" phase to emerge/disappear?
   - For dense or small subgraphs, we observe AoN
   - For sparse, large subgraphs, we observe ASN

Backup Slides

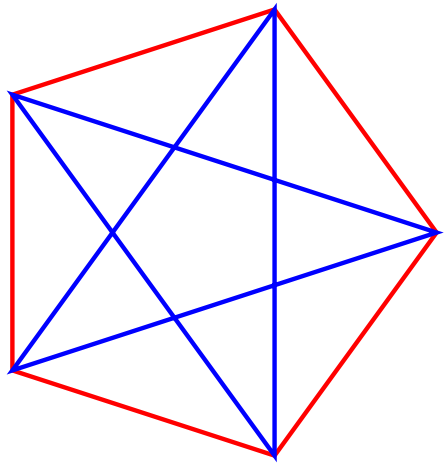
Upper bound: Enumerating k-factors near H^*

For k-factor H, $H \triangle H^* =$ disjoint union of alternating Eulerian circuits



Upper bound: Enumerating k-factors near H^*

For k-factor H, $H \triangle H^* =$ disjoint union of alternating Eulerian circuits



As a consequence,

$$|\{k\text{-factor } H : |H \Delta H^*| = 2t\}| \leq \binom{kn/2}{t}(2t-1)!! \leq (kn)^t$$

$$\Rightarrow \mathbb{E}[|\{k\text{-factor } H : |H \Delta H^*| = 2t, H \subset G\}|] \leq (kn)^t \left(\frac{\lambda}{n}\right)^t = (k\lambda)^t$$

A generic, non-constructive lower bound

- Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distribution of the planted k-factor model and $\mathcal{G}(n, \lambda/n)$, respectively
- Recall $\mathcal{H}(G)$ is the set of k-factors in G

Lemma (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

For any $\epsilon > 0$,

$$\mathbb{P}\left\{|\mathcal{H}(G)| \geq \epsilon \mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|\right\} \geq 1 - \epsilon$$

A generic, non-constructive lower bound

- Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distribution of the planted k-factor model and $\mathcal{G}(n, \lambda/n)$, respectively
- Recall $\mathcal{H}(G)$ is the set of k-factors in G

Lemma (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

For any $\epsilon > 0$,

$$\mathbb{P}\left\{|\mathcal{H}(G)| \geq \epsilon \mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|\right\} \geq 1 - \epsilon$$

Proof: $\mathbb{P}(G)/\mathbb{Q}(G) = |\mathcal{H}(G)|/\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|$ and change of measure

A generic, non-constructive lower bound

- Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distribution of the planted k-factor model and $\mathcal{G}(n, \lambda/n)$, respectively
- Recall $\mathcal{H}(G)$ is the set of k-factors in G

Lemma (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

For any $\epsilon > 0$,

$$\mathbb{P}\left\{|\mathcal{H}(G)| \geq \epsilon \mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|\right\} \geq 1 - \epsilon$$

Proof: $\mathbb{P}(G)/\mathbb{Q}(G) = |\mathcal{H}(G)|/\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|$ and change of measure

- When $\lambda k = \omega(1)$, $\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)| \gg |\mathcal{H}_{\text{near}}(G)|$ and suffices for proving overlap$(H, H^*) \to 0$ in "Nothing" phase

A generic, non-constructive lower bound

- Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distribution of the planted k-factor model and $\mathcal{G}(n, \lambda/n)$, respectively
- Recall $\mathcal{H}(G)$ is the set of k-factors in G

Lemma (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

For any $\epsilon > 0$,

$$\mathbb{P}\{|\mathcal{H}(G)| \geq \epsilon \mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|\} \geq 1 - \epsilon$$

Proof: $\mathbb{P}(G)/\mathbb{Q}(G) = |\mathcal{H}(G)|/\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|$ and change of measure

- When $\lambda k = \omega(1)$, $\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)| \gg |\mathcal{H}_{\text{near}}(G)|$ and suffices for proving overlap$(H, H^*) \to 0$ in "Nothing" phase
- When $\lambda k = O(1)$, $\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)| \ll |\mathcal{H}_{\text{near}}(G)|$ and falls short for proving overlap$(H, H^*) \leq 1 - \Omega(1)$ in "Something" phase

A generic, non-constructive lower bound

- Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distribution of the planted k-factor model and $\mathcal{G}(n, \lambda/n)$, respectively
- Recall $\mathcal{H}(G)$ is the set of k-factors in G

Lemma (Mossel-Niles-Weed-Sohn-Sun-Zadik '23)

For any $\epsilon > 0$,

$$\mathbb{P}\{|\mathcal{H}(G)| \geq \epsilon \mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|\} \geq 1 - \epsilon$$

Proof: $\mathbb{P}(G)/\mathbb{Q}(G) = |\mathcal{H}(G)|/\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)|$ and change of measure

- When $\lambda k = \omega(1)$, $\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)| \gg |\mathcal{H}_{\text{near}}(G)|$ and suffices for proving overlap$(H, H^*) \to 0$ in "Nothing" phase
- When $\lambda k = O(1)$, $\mathbb{E}_{\mathbb{Q}}|\mathcal{H}(G)| \ll |\mathcal{H}_{\text{near}}(G)|$ and falls short for proving overlap$(H, H^*) \leq 1 - \Omega(1)$ in "Something" phase
  $\longrightarrow$ need a tighter lower bound

Constructing k-factors far away from H^*

Goal: Find exponentially many long alternating Eulerian circuits (AEC) in G

- long AEC are rare;

Constructing k-factors far away from H^*

Goal: Find exponentially many long alternating Eulerian circuits (AEC) in G

- long AEC are rare; but there are many possibilities to consider

Constructing k-factors far away from H^*

Goal: Find exponentially many long alternating Eulerian circuits (AEC) in G

- long AEC are rare; but there are many possibilities to consider
- Natural second-moment argument doesn't work due to excessive correlation between long AEC.

Constructing k-factors far away from H^*

Goal: Find exponentially many long alternating Eulerian circuits (AEC) in G

- long AEC are rare; but there are many possibilities to consider
- Natural second-moment argument doesn't work due to excessive correlation between long AEC.

Key idea: Sprinkling

1. Reserve a small fraction of vertices
2. Greedily construct many disjoint short alternating paths using non-reserved vertices
3. Connect the paths into long alternating cycles via reserved vertices

Inspired by [Aldous '98, Ding '13, Ding-Wu-X.-Yang '21]

Existence of many long augmenting alternating cycles

Two-stage cycle-finding scheme

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

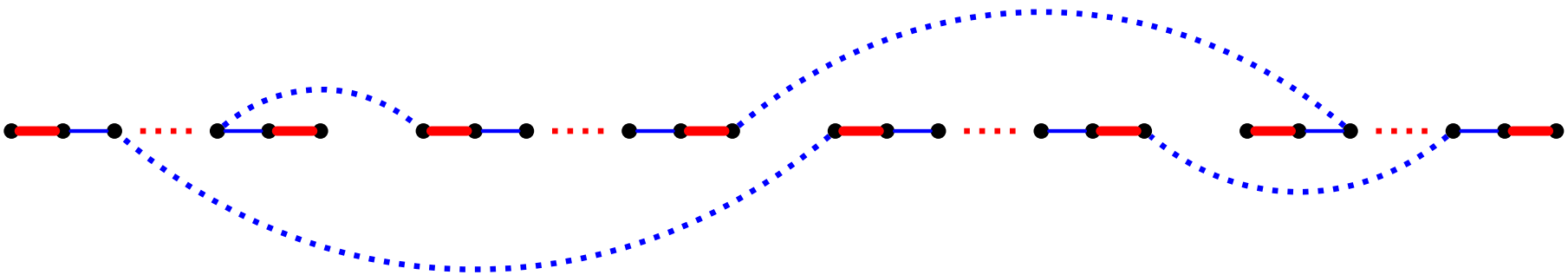
1. Stage 1 (path construction): Find $\Theta(n)$ disjoint short (constant length) alternating paths, using vertices in V^c.

Existence of many long augmenting alternating cycles

Two-stage cycle-finding scheme

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

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

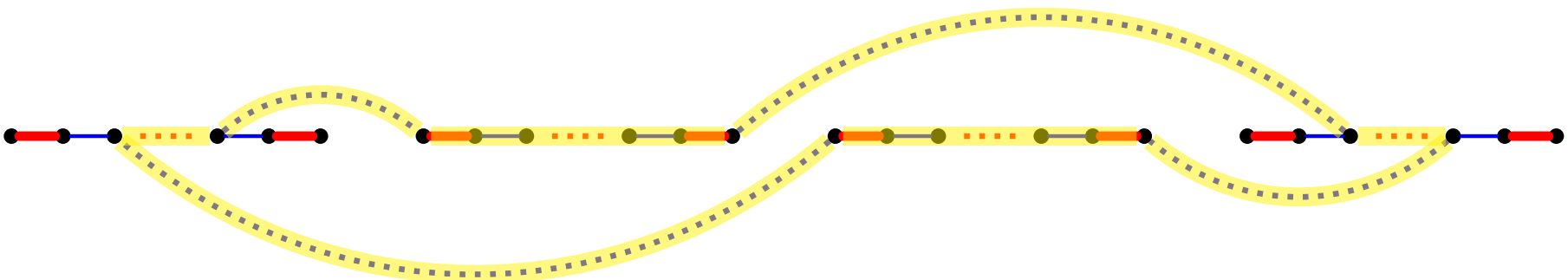


Existence of many long augmenting alternating cycles

Two-stage cycle-finding scheme

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

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

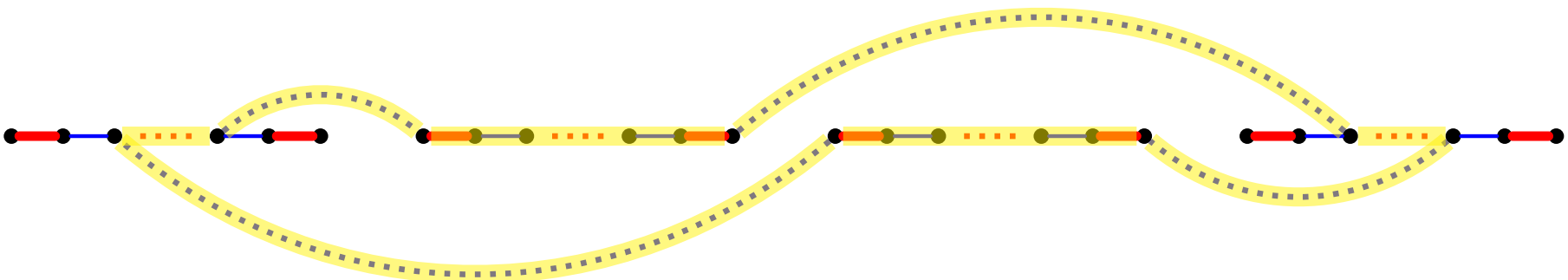


Existence of many long augmenting alternating cycles

Two-stage cycle-finding scheme

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

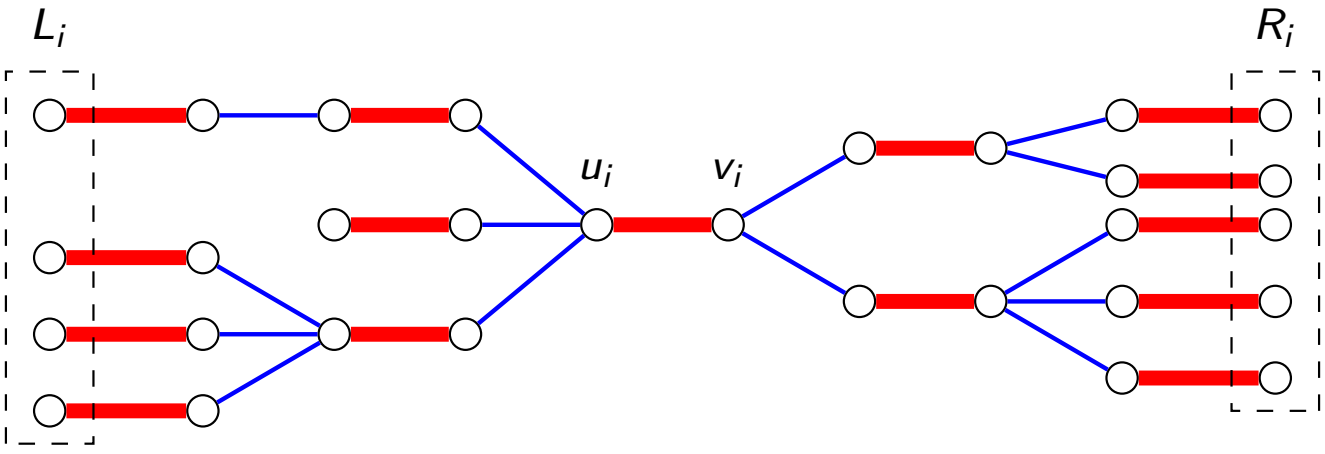
1. Stage 1 (path construction): Find $\Theta(n)$ disjoint short (constant length) alternating paths, using vertices in V^c.
2. Stage 2 (sprinkling): Connect the paths into long cycles, using vertices in V.



Caution: need to ensure alternating colors in sprinkling

Path construction via neighborhood exploration process

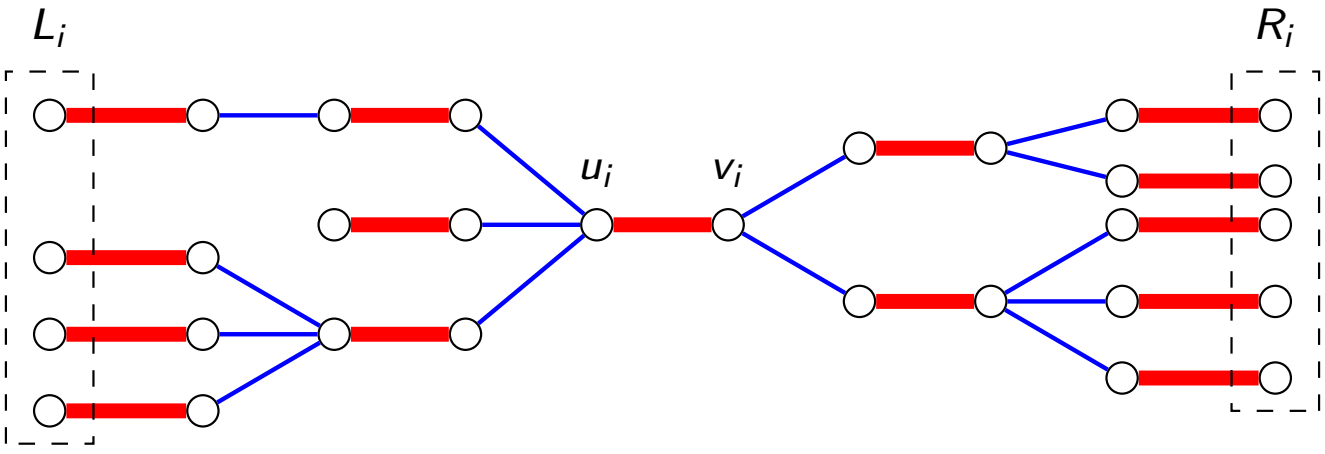
Pick a planted edge (u_i, v_i), grow a left tree starting from u_i, remove the inspected vertices, and then grow the right tree from v_i



An illustration for $k = 1$. For $k > 1$, need to branch out all k red edges

Path construction via neighborhood exploration process

Pick a planted edge (u_i, v_i), grow a left tree starting from u_i, remove the inspected vertices, and then grow the right tree from v_i



An illustration for $k = 1$. For $k > 1$, need to branch out all k red edges

When $\lambda k > 1$, the branching processes survive with constant probability. Thus the BFS returns $K = \Omega(n)$ two-sided trees with many leaf nodes.