Tuyrr fq%Ijyjhyrts%tk%f%Ufsyji%R fyhmrsl%{rf% ymj%Hqzxyjw%J}ufsxrts

Cheng Mao

Ljtwlrf%Yjhm

joint work with **Timothy Wee**

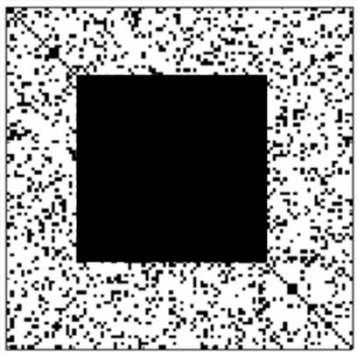
NR XNB%Hmmhflt

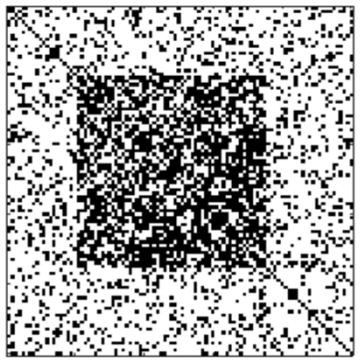
Ofszfw~%81%757;

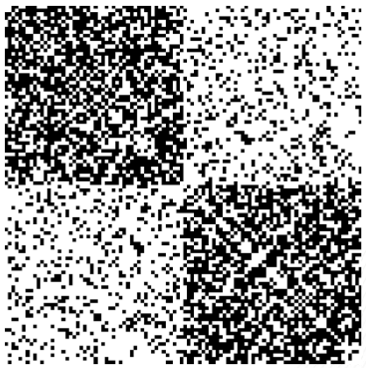
1. Background

Planted models of random graphs

- Planted clique [Jerrum ’92]
- Planted dense subgraph [Bhaskara et al ’10]
- Planted partition (stochastic block model) [Holland et al ’83]
- ...







Planted models of random graphs

- Planted (bipartite) matching [Chertkov et al ’10]
- Planted cycle (small world model) [Watts, Strogatz ’98]
- Planted trees [Massoulié et al ’18]
- Planted k-factors [Sicuro, Zdeborová ’20]
- …

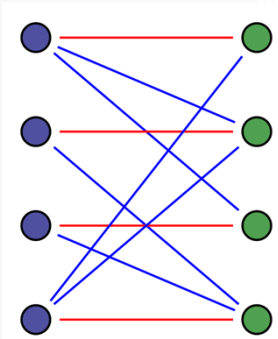
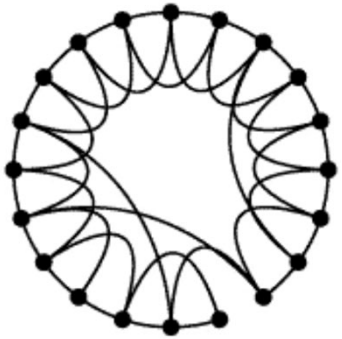
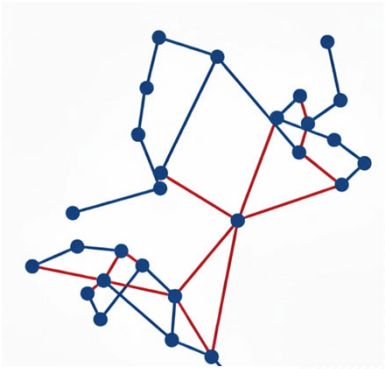


Figure from J. Xu’s slides





This talk: Planted matching (independent edge set)

- Statistical physics

  [Chertkov, Kroc, Krzakala, Vergassola, Zdeborová ’10]

  [Adomaityte, Toshniwal, Sicuro, Zdeborová ’22]

- Phase transitions for recovery

  [Semerjian, Sicuro, Zdeborová ’20]

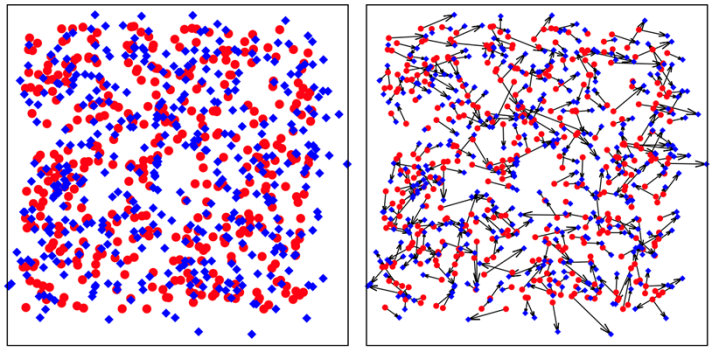
  [Moharrami, Moore, Xu ’21]

  [Ding, Wu, Xu, Yang ’23]

- Perfect matchings in random graphs

  [Janson ’94]

  analyzes likelihood ratio as an intermediate result



Particle tracking in turbulent flows

2. Models and results

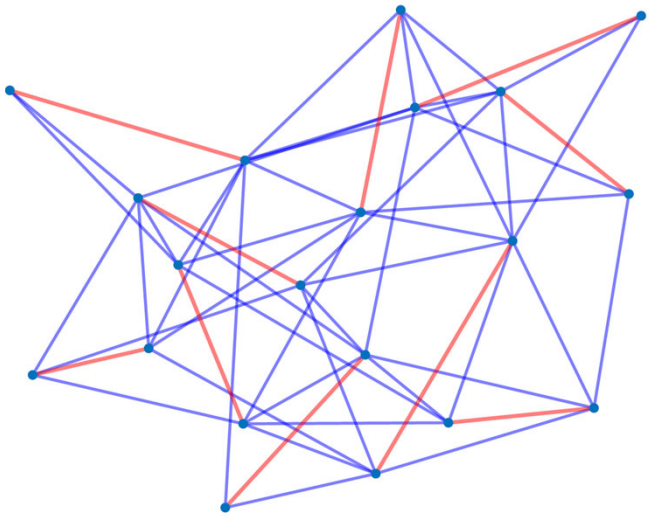
Model: Planted matching in an Erdős–Rényi graph

Graph with adjacency matrix $A \sim G(n,p) + M$

- $A_{ij} \sim \text{Bernoulli}(p)$ if $(i,j) \notin M$
- M a random matching

  $\mu(M) = \frac{1}{Z}\lambda^{|M|},\ Z = \sum_{M \subset K_n} \lambda^{|M|}$

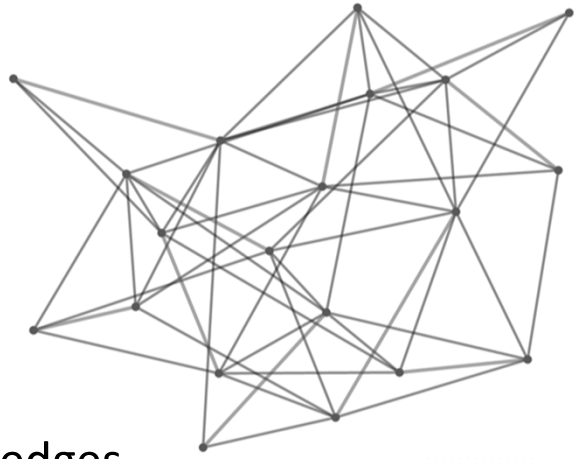
  - Monomer-dimer model
    [Heilmann, Lieb '72]
  - If $\lambda n = c_1$, then $|M| \approx c_2 n$
    [Alberici, Contucci, Mingione '14]



The detection problem: hypothesis testing

Given A, test $P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q)$

- $p \approx q$ defined so that $\mathbf{E}_P A_{ij} = \mathbf{E}_Q A_{ij}$



Phase transition: $|M| \approx cn$

Short answer: $?? = n^{-1/2}$, A has $\asymp n^{3/2}$ edges

Main result

$P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q)$

"Signed wedge count" $T(A) := \sum_{ijk}(A_{ij} - q)(A_{jk} - q)$

Theorem (Wee, M. '25)

Log-likelihood ratio dominated by $T(A)$ under Q:

$$\log \frac{dP}{dQ}(A) = -\frac{\sigma^2}{2} + \sigma \frac{T(A)}{\sqrt{\mathrm{Var}_Q\, T(A)}} + O_{\mathbf{P}}\left(\frac{1}{\sqrt{nq}}\right)$$

where $\sigma = \frac{1}{\sqrt{2nq}}\left(\frac{2\mathbf{E}|M|}{n}\right)^2$

- Le Cam's contiguity condition: mean $= -1/2 \times$ variance

Efficient test

$P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q)$, $|M| \approx cn$, $q\sqrt{n} \to \theta$

Proposition

The signed wedge count $T(A)$ is asymptotically normal under Q, P.

Let $\phi(A) = P$ if $T(A) < -c^2 q/\theta$ and $\phi(A) = Q$ otherwise. Then

$$\mathbf{P}_P\{\phi(A) = Q\} + \mathbf{P}_Q\{\phi(A) = P\} \to \operatorname{erfc}(c^2/\theta)$$

Asymptotic normality of log-likelihood ratio

$P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q)$, $|M| \approx cn$, $q\sqrt{n} \to \theta$

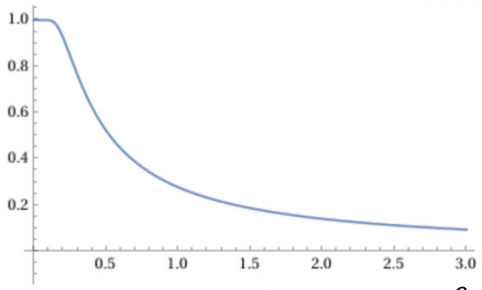
Corollary

Log-likelihood ratio is asymptotically $N(\pm 4c^4/\theta^2, 8c^4/\theta^2)$ under Q, P

Likelihood ratio test achieves the same testing error, and

$\mathrm{TV}(P,Q) \to 1 - \mathrm{erfc}(c^2/\theta)$.

- If $q \ll 1/\sqrt{n}$, then $\mathrm{TV}(P,Q) \to 1$ (possible)
- If $q \gg 1/\sqrt{n}$, then $\mathrm{TV}(P,Q) \to 0$ (impossible)

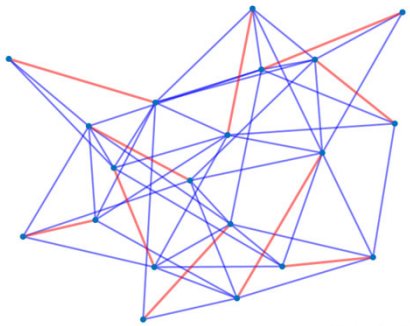


3. Testing threshold via wedge count

Signed wedge count

$P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q)$, $|M| \approx cn$, $q\sqrt{n} \to \theta$

$T(A) \coloneqq \sum_{i<j<k}(A_{ij} - q)(A_{jk} - q)$



Lemma $\mathbf{E}_Q T(A) = 0$, $\mathbf{E}_P T(A) = -2c^2 n$,

$$\mathrm{Var}_Q T(A) \approx \mathrm{Var}_P T(A) = n^3 q^2/2$$

- No wedge in a matching $\implies$ $\mathbf{E}_P T(A)$ negative
- $\left|\mathbf{E}_P - \mathbf{E}_Q\right| \asymp \sqrt{\mathrm{Var}}$ $\implies$ Threshold $q \asymp 1/\sqrt{n}$
- $T(A)$ asymptotically normal $\implies$ Precise testing error $\mathrm{erfc}(c^2/\theta)$

[Janson '95]

4. Cluster expansion of log-likelihood ratio

Log-likelihood ratio

$P: A \sim G(n,p) + M$ v.s. $Q: A \sim G(n,q),\ \mu(M) \propto \lambda^{|M|}$

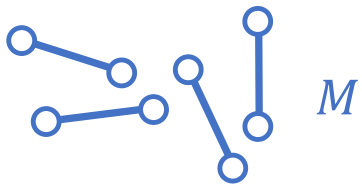
Lemma

$$\log \frac{dP}{dQ}(A) = F(A) + \log Z_A(\lambda/p) - \log Z_{K_n}(\lambda)$$

where $F(A) \coloneqq |A| \log \frac{p(1-q)}{q(1-p)} + \binom{n}{2} \log \frac{1-p}{1-q}$,

$Z_G(\lambda) \coloneqq \sum_{M \subset G} \lambda^{|M|}$ partition function of monomer-dimer model

matching M in graph G



M

Cluster expansion

- Series expansion of log of “polymer” partition function
  [Gruber, Kunz ’71], [Brydges ’84], [Kotecky, Preiss ’86],
  [Scott, Sokal ’05], [Friedli, Velenik ’17] Chapter 5

- Applications in combinatorics, statistical physics, random graphs
  [Balogh, Treglown, Wagner ’16], [Dey, Wu ’23],
  [Helmuth, Jenssen, Perkins ’23], [Jenssen, Perkins, Potukuchi ’25],
  [Bangachev, Bresler ’24], [Quitmann ’24], …

Cluster expansion: general theory

Partition function: finite set G of "polymers" γ

$$Z = \sum_{M \subset G} \left(\prod_{\gamma \in M} w(\gamma) \right) \left(\prod_{\gamma, \gamma' \in M} \delta(\gamma, \gamma') \right)$$

weight pairwise interaction

Cluster expansion (formal series):

$$\log Z = \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \ldots, \gamma_m) \prod_{i=1}^{m} w(\gamma_i)$$

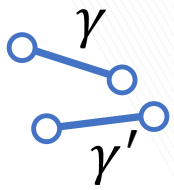
where $\varphi(\gamma_1, \ldots, \gamma_m) = \frac{1}{m!} \sum_H \prod_{(i,j) \in H} \left(\delta(\gamma_i, \gamma_j) - 1 \right)$ "Ursell function"

connected, spanning graph H on vertex set $\{1, \ldots m\}$

Apply the cluster expansion

Partition function:

$$Z = \sum_{M \subset G} \left(\prod_{\gamma \in M} w(\gamma) \right) \left(\prod_{\gamma, \gamma' \in M} \delta(\gamma, \gamma') \right)$$



In our case $\gamma = (i, j)$, $w(\gamma) = \lambda$, $\delta(\gamma, \gamma') = \mathbf{1}\{\gamma, \gamma' \text{ not adjacent}\}$

"hard-core repulsion"

$$Z_G(\lambda) := \sum_{\substack{\text{matching} \\ M \subset G}} \prod_{(i,j) \in M} \lambda = \sum_{M \subset G} \lambda^{|M|}$$

Apply the cluster expansion

Cluster expansion:

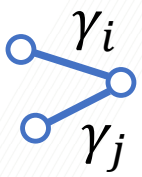
$$\log Z_G(\lambda) = \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \ldots, \gamma_m)\, \lambda^m$$

where, since $\mathbf{1}\{\gamma_i, \gamma_j \text{ not adjacent}\} - 1 = -\mathbf{1}\{\gamma_i, \gamma_j \text{ adjacent}\}$,

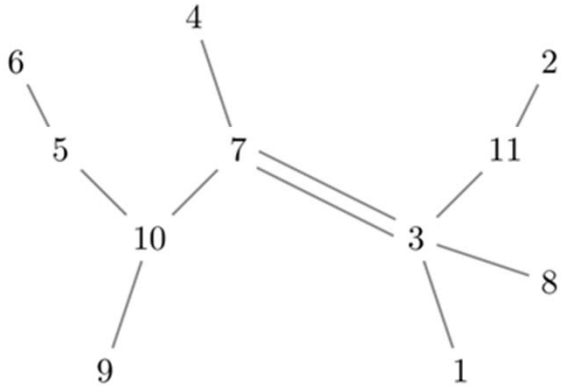
$$\varphi(\gamma_1, \ldots, \gamma_m) = \frac{1}{m!} \underbrace{\sum_H} \prod_{(i,j) \in H} \left(-\mathbf{1}\{\gamma_i, \gamma_j \text{ adjacent}\}\right)$$

connected, spanning graph H on $\{1, \ldots m\}$

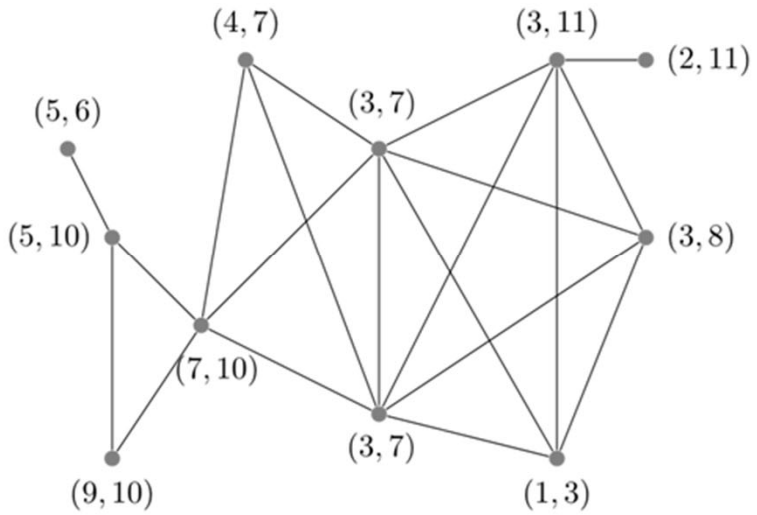
so $\gamma_1, \ldots, \gamma_m$ form a connected multigraph called a **"cluster"**



Example of a cluster and its line graph



Cluster $\gamma_1, \ldots, \gamma_m$



H is a connected, spanning subgraph of the line graph

Cluster expansion of the log-likelihood ratio

Log-likelihood ratio:

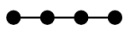
$$\log \frac{dP}{dQ}(A) = F(A) + \log Z_A(\lambda/p) - \log Z_{K_n}(\lambda)$$
$$= F(A) + \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \cdots, \gamma_m) \left(\prod_{i=1}^{m} \left(A_{\gamma_i} \frac{\lambda}{p} \right) - \lambda^m \right)$$

where edges $\gamma_1, \ldots, \gamma_m$ form a cluster (connected multigraph)

- Not just a formal series, converges if $\lambda \leq c_1/n$, $\mathbf{E}|M| \leq c_2 n$

First few terms in the cluster expansion for $p = q$

$$\log \frac{dP}{dQ}(A) = \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \ldots, \gamma_m) \left(\prod_{i \in [m]} \left(A_{\gamma_i} \frac{\lambda}{p} \right) - \lambda^m \right)$$

Cluster	Size	Count	Ursell
K_2	1	$\binom{n}{2}$	1
K_2 with rep. edge	2	$\binom{n}{2}$	$-\frac{1}{2}$
 P_2	2	$2! \cdot 3 \cdot \binom{n}{3}$	$-\frac{1}{2}$
K_2 with two rep. edge	3	$\binom{n}{2}$	$\frac{1}{3}$
 P_2 with one rep. edge	3	$\frac{3!}{2!} \cdot 2 \cdot 3 \cdot \binom{n}{3}$	$\frac{1}{3}$
 K_3	3	$3!\binom{n}{3}$	$\frac{1}{3}$
 S_3	3	$3!4\binom{n}{4}$	$\frac{1}{3}$
 P_3	3	$3!\frac{4!}{2}\binom{n}{4}$	$\frac{1}{6}$

 $\lambda \left[\frac{K_2(A)}{p} - \binom{n}{2} \right]$

 $-\frac{\lambda^2}{2} \left[\frac{K_2(A)}{p^2} - \binom{n}{2} \right]$

 $-\lambda^2 \left[\frac{P_2(A)}{p^2} - 3\binom{n}{3} \right]$

 $\frac{\lambda^3}{3} \left[\frac{K_2(A)}{p^3} - \binom{n}{2} \right]$

 $2\lambda^3 \left[\frac{P_2(A)}{p^3} - 3\binom{n}{3} \right]$

 $2\lambda^3 \left[\frac{K_3(A)}{p^3} - \binom{n}{3} \right]$

 $2\lambda^3 \left[\frac{S_3(A)}{p^3} - 4\binom{n}{4} \right]$

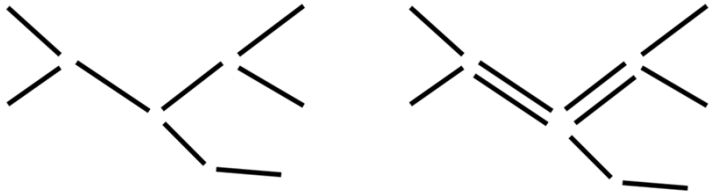
 $\lambda^3 \left[\frac{P_3(A)}{p^3} - \frac{4!}{2}\binom{n}{4} \right]$

Log-likelihood ratio dominated by signed wedge count

$$\log \frac{dP}{dQ}(A) = F(A) + \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \ldots, \gamma_m) \left(\prod_{i \in [m]} \left(A_{\gamma_i} \frac{\lambda}{p} \right) - \lambda^m \right)$$

Step 1: $\log \frac{dP}{dQ}(A)$ is dominated by clusters that are **trees** with at most two repeated edges

Step 2: The dominating terms are either (essentially) deterministic or $\big(1 - o(1)\big)$-corrected with the signed wedge count $T(A)$



Variance Mean

Log-likelihood ratio dominated by signed wedge count

$$\log \frac{dP}{dQ}(A) = F(A) + \sum_{m \geq 1} \sum_{\gamma_1, \ldots, \gamma_m} \varphi(\gamma_1, \ldots, \gamma_m) \left(\prod_{i \in [m]} \left(A_{\gamma_i} \frac{\lambda}{p} \right) - \lambda^m \right)$$

Theorem For $A \sim G(n, q)$ and $\lambda \leq c_1/n$, $\mathbf{E}|M| \leq c_2 n$,

$$\log \frac{dP}{dQ}(A) \approx -\frac{\sigma^2}{2} + \sigma \frac{T(A)}{\sqrt{\operatorname{Var} T(A)}} \longrightarrow N\left(-\frac{\sigma^2}{2}, \sigma^2\right)$$

where $\sigma = \frac{1}{\sqrt{2nq}} \left(\frac{2\mathbf{E}|M|}{n} \right)^2$

- [Janson '94]: Asymptotic normality for perfect matching $|M| = n/2$

5. Cluster expansion v.s. orthogonal decomposition

Planted matching

Cluster expansion:

$$\log \frac{dP}{dQ}(A) = F(A) + \sum_{m \geq 1} \sum_{\substack{\gamma_1,\dots,\gamma_m \\ \text{connected}}} \varphi(\gamma_1, \dots, \gamma_m) \lambda^m \left(\prod_{i \in [m]} \frac{A_{\gamma_i}}{p} - 1 \right)$$

log-likelihood, infinite series, cumulants, not unique, subgraph count

Orthogonal decomposition [Janson '94-95, Hopkins '18]:

$$\frac{dP}{dQ}(A) = 1 + \sum_{m=1}^{\binom{n}{2}} \sum_{\substack{\gamma_1,\dots,\gamma_m \\ \text{distinct}}} \psi(\gamma_1, \dots, \gamma_m) \left(\prod_{i \in [m]} (A_{\gamma_i} - p) \right)$$

likelihood, finite sum, moments, unique, signed subgraph count

Georgia Tech
CREATING THE NEXT

Planted k-clique in $G(n, 1/2)$: a curious case

Cluster expansion (truncated):

$$\left(\log \frac{dP}{dQ}(A)\right)_{\text{trun.}} = \sum_{\substack{\alpha \subset K_n,\ \alpha \neq \emptyset \\ \text{connected}}} \left(\frac{k}{n-k}\right)^{|V(\alpha)|} \prod_{\gamma \in \alpha} (2A_\gamma - 1)$$

Orthogonal decomposition:

$$\frac{dP}{dQ}(A) = 1 + \sum_{\alpha \subset K_n,\ \alpha \neq \emptyset} \left(\frac{k}{n}\right)^{|V(\alpha)|} \prod_{\gamma \in \alpha} (2A_\gamma - 1)$$

- **Good**: can see statistical & computational thresholds from $\left(\log \frac{dP}{dQ}\right)_{\text{trun.}}$
- **Bad**: cluster expansion doesn't converge unless $k = O(1)$

Summary

- Detecting a planted matching in a random graph
- Phase transition at the critical threshold
- Efficient statistic: signed wedge count
- Analysis of the log-likelihood ratio via the cluster expansion

Question: Other planted models (e.g. k-regular graphs)?

Thank you!

Georgia Tech
CREATING THE NEXT