Algorithms and Phase Transitions in Random Graph Alignment Problem

Shuyang Gong

School of Mathematical Sciences, Peking University

September, 2023

Joint work with Jian Ding(PKU), Hang Du(MIT) and Rundong Huang(PKU)

Motivations

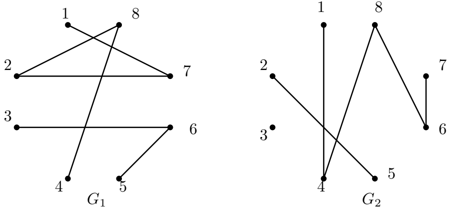
- **Random Graph Matching** is an extensively studied topic in recent years, which lies in the intersection of **probability, statistics and computer science**
- Goal: find a bijection between two vertex sets which maximizes the number of common edges(i.e. minimize the adjacency disagreements)
- **Quadratic Assignment Problem(QAP)**: $\max_{\pi \in S_n} \sum_{i<j} A_{i,j} B_{\pi(i),\pi(j)}$. Introduced by [Koopmans-Beckmann'57] NP-hard in the worst case.
  - $A \simeq B \Rightarrow$ Graph isomorphism problem(no noise)
  - But two graphs are usually not isomorphic.(statistical challenge)
  - $n!$ bijections (computationally expensive)
- Efforts from community on **average case** of graph matching: [Feizi at el.'16, Lyzinski at el'16, Cullina-Kiyavash'16,17, Ding-Ma-Wu-Xu'18, Barak-Chou-Lei-Schramm-Sheng'19, Fan-Mao-Wu-Xu'19a,19b, Ganassali-Massoulié'20, Hall-Massoulié'20, Ding-Du'22a,22b, Ding-Du-G'22, Ding-Li'22,23, Du-G-Huang'23...]

Mathematical model

- Erdős-Rényi graph $\mathrm{G}(n,p)$: Each edge in K_n is preserved with probability p independently.
- Sample two independent Erdős-Rényi graphs $\mathrm{G}_1(n,p)$ and $\mathrm{G}_2(n,p)$.
- **Core quantity** $\mathrm{O}(\pi)$: the number of **common edges** of these two graphs under π. Formally,

$$\mathrm{O}(\pi) := \sum_{i<j} \mathrm{G}^{(1)}_{i,j} \mathrm{G}^{(2)}_{\pi(i),\pi(j)} ,$$

  where $G^{(i)}$ are adjacency matrices.
- e.g. $\pi(1)=1, \pi(2)=8, \pi(3)=2, \pi(4)=7, \pi(5)=3, \pi(6)=5, \pi(7)=4, \pi(8)=6 \Rightarrow$ we have $\mathrm{O}(\pi)=5$



Our problem

- **Q1: what is the typical value of** $\max_{\pi \in S_n} O(\pi)$?
- A first moment computation on $\max_{\pi \in S_n} O(\pi)$ yields an **upper bound**, e.g. take $p = n^{-3/4}$, let $\gamma(n) := (1+\varepsilon)2n$,

$$
\begin{aligned}
&\mathbb{P}\left[\max_{\pi \in S_n} O(\pi) > 2(1+\varepsilon)n\right] \leq \sum_{\pi \in S_n} \mathbb{P}\left[O(\pi) \geq 2(1+\varepsilon)n\right] \\
=& n!\mathbb{P}\left[\mathbf{B}\left(\binom{n}{2}, p^2\right) > 2(1+\varepsilon)n\right] \\
\overset{\text{Chernoff}}{\leq}& n! \exp\left(-2(1+\varepsilon)n \log\left(\frac{2(1+\varepsilon)n}{\binom{n}{2} n^{-3/2}}\right) + 2(1+\varepsilon)n - \binom{n}{2} n^{-3/2}\right) \\
=& n! \exp(-(1+\varepsilon+o(1))n \log n) = o(1)\,.
\end{aligned}
$$

- The calculation for other p is similar.

Our problem

- For other p(divide into sparse/dense by $\sqrt{\log n/n}$),

regime	$\max_{\pi \in S_n} O(\pi)$
sparse: $\frac{\log n}{n} \ll p \ll \sqrt{\frac{\log n}{n}}$	$n \cdot \frac{\log n}{\log(\log n/np^2)}$
dense: $\sqrt{\frac{\log n}{n}} \ll p \le \frac{1}{(\log n)^4}$	$\binom{n}{2}p^2 + \sqrt{n^3p^2 \log n}$

- First moment computation $\Rightarrow$ Upper bound w.h.p.
- Right asymptotics?—True.
- **Q2: Find a polynomial time algorithm for** $\arg\max_\pi O(\pi)$**?**(sparse: yes, dense: no)
- **Information-Computation gap?**(sparse: no, dense: yes)

Sparse regime

Theorem (Ding-Du-G. 22)

For $p = n^{-\alpha+o(1)}$, $1/2 < \alpha \le 1$, there exists a polynomial-time algorithm s.t.

$$\mathbb{P}\left[\mathrm{O}(\pi^*) \ge \frac{1-\epsilon}{2\alpha-1}n\right] = 1 - o(1).$$

Theorem (Du-G.-Huang 23)

For $p = n^{-1/2+o(1)}$ and $p \ll \sqrt{\log n/n}$, for any $\varepsilon > 0$, there exists an $O(n^3)$-time algorithm such that

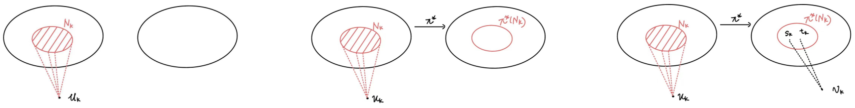
$$\mathbb{P}\left[\mathrm{O}(\pi^*) \ge \frac{(1-\varepsilon)n\log n}{\log\left(\log n/np^2\right)}\right] = 1 - o(1)\,.$$

- $n/(2\alpha-1) = (1+o(1))n\log n/\log\left(\log n/np^2\right)$ for $p = n^{-\alpha}$.
- The **constructive lower bound** matches the $\gamma(n)$ derived in the first moment computation.
- **No information-computation gap** in the sparse regime.

Algorithm

The algorithm in [Ding-Du-G'22], let $\alpha = 3/4 - \delta$, $\frac{1}{2\alpha-1}n \approx 2n$

- Match the first εn vertices arbitrarily.
- In step $k+1$, select unmatched u_k in G_1. Neighbor of u_k in matched part N_k
- Map N_k by π^*
- For each such pair (s_k, t_k), check if there exists unmatched v_k in G_2
- If succeed, let $\pi^*(u_k) = v_k$.



(i) (ii) (iii)

Dense regime—information

Theorem (Du-G.-Huang 23, informational result)

For p in the dense regime, we have

$$\frac{\max_{\pi\in S_n} O(\pi) - \binom{n}{2}p^2}{\sqrt{n^3p^2\log n}} \xrightarrow{prob.} 1\,.$$

- **Second moment method:** $X_\varepsilon := \sum_{\pi\in S_n} \mathbf{1}_{O(\pi)>\binom{n}{2}p^2+\sqrt{(1-\varepsilon)n^3p^2\log n}}$.

$$\mathbb{P}\left[\max_{\pi\in S_n} O(\pi) > \binom{n}{2}p^2 + \sqrt{(1-\varepsilon)n^3p^2\log n}\right] \geq \frac{(\mathbb{E}X_\varepsilon)^2}{\mathbb{E}X_\varepsilon^2} = \exp(-o(n\log n))\,.$$

- **Talagrand's concentration:**

$$\mathbb{P}\left[\left|\max_{\pi\in S_n} O(\pi) - \mathbb{E}\max_{\pi\in S_n} O(\pi)\right| \geq \sqrt{\varepsilon n^3p^2\log n}\right] \leq \exp\left(-c(\varepsilon)n\log n\right)\,.$$

- **Idea:** concentration of maximum
- (Gaussian) Talagrand's concentration $\Rightarrow$ Borell–TIS inequality

Dense regime—computation

Theorem (Du-G.-Huang 23, computational result)

There exists an $O(n^3)$-time algorithm $\mathcal{A}$ which outputs a π^* such that

$$\mathbb{P}\left[\frac{\mathsf{O}(\pi^*)-\binom{n}{2}p^2}{\sqrt{n^3p^2\log n}}\geq \sqrt{\mathbf{8/9}}-\varepsilon\right]=1-o(1)\,.$$

Theorem (Du-G.-Huang 23, hardness result)

For p in the dense regime, for all $\varepsilon>0$, there exists a constant $c=c(\varepsilon)>0$ such that for any online algorithm $\mathcal{A}$,

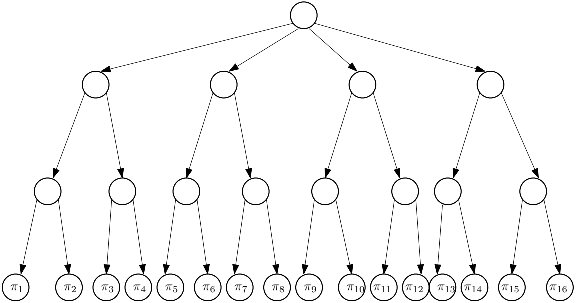
$$\mathbb{P}\left[\frac{\mathsf{O}(\mathcal{A}(\mathsf{G}_1,\mathsf{G}_2))-\binom{n}{2}p^2}{\sqrt{n^3p^2\log n}}\geq \sqrt{\mathbf{8/9}}+\varepsilon\right]=o(1)\,.$$

No online algorithm above $\sqrt{8/9}$! — Hardness result.

- Main tool: **Branching-OGP structure**[Huang-Sellke'22].

The Branching OGP

- Define tree $\mathbb{T}$ with leave set $\mathbb{L}$.
- Construct leave-indexed correlated instances $\{(\mathsf{G}_1^{(u)}, \mathsf{G}_2)\}_{u\in\mathbb{L}}$.
- Each ray represents an instance. $G_1^{(u)}$ and $G_1^{(v)}$ share $\rho_{|u\wedge v|}$ Bernoulli variables.($\rho_1 < \rho_2 < \rho_3$)
- Impossible for all $\mathsf{O}(\pi_i)$ above $\sqrt{8/9} + \varepsilon$.
- Run online algorithm on all instances. Prove by contradiction.



Summary

Take-home messages

- In sparse regime, no information computation gap.
- In dense regime, information computation gap emerges with a threshold $\sqrt{8/9}$.

Related papers:

DDG22 Jian Ding, Hang Du and Shuyang Gong, A Polynomial-time Approximation Scheme for the Maximal Overlap Between Two Independent Erdős-Rényi Graphs, *arXiv:2210.07823*.

DGH23 Hang Du, Shuyang Gong and Rundong Huang, The Algorithmic Phase Transition of Random Graph Alignment Problem, *arXiv:2307.06590*.

Thank you for your listening!

References

[BCPP98] Rainer E Burkard, Eranda Cela, Panos M Pardalos, and Leonidas S Pitsoulis. The quadratic assignment problem. In handbook of combinatorial optimization, pages 1713-1809. *Springer*, 1998.
[DDG22] Jian Ding, Hang Du and Shuyang Gong, A Polynomial-time Approximation Scheme for the Maximal Overlap Between Two Independent Erdős-Rényi Graphs, *arXiv:2210.07823*.
[DGH23+]Hang Du, Shuyang Gong and Rundong Huang, The Algorithmic Phase Transition of Random Graph Alignment Problem, *arXiv:2307.06590*.
[HS22] B. Huang and M. Sellke, "Tight Lipschitz Hardness for optimizing Mean Field Spin Glasses," *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), Denver, CO, USA, 2022, pp. 312-322, doi: 10.1109/FOCS54457.2022.00037*.
[PRW94] Panos M. Pardalos, Franz Rendl, and Henry Wolkowicz. The quadratic assignment problem: A survey and recent developments.