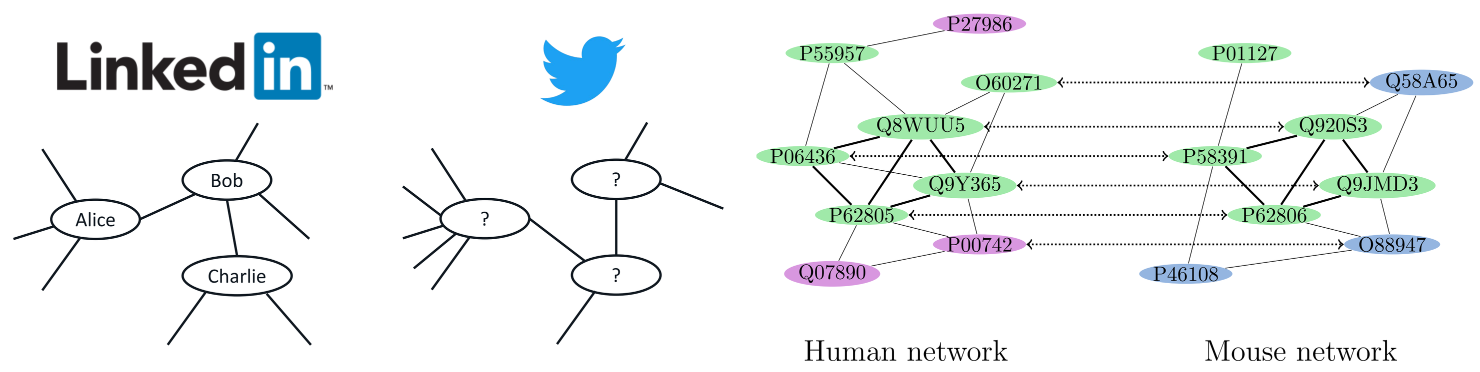
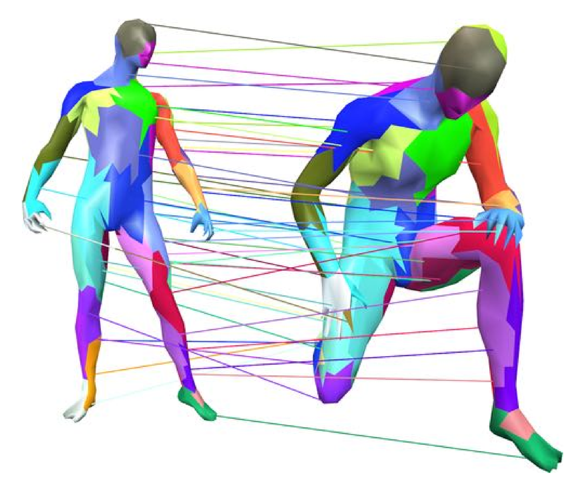
Optimal matching recovery of correlated Erdős-Rényi graphs

Hang Du (MIT)

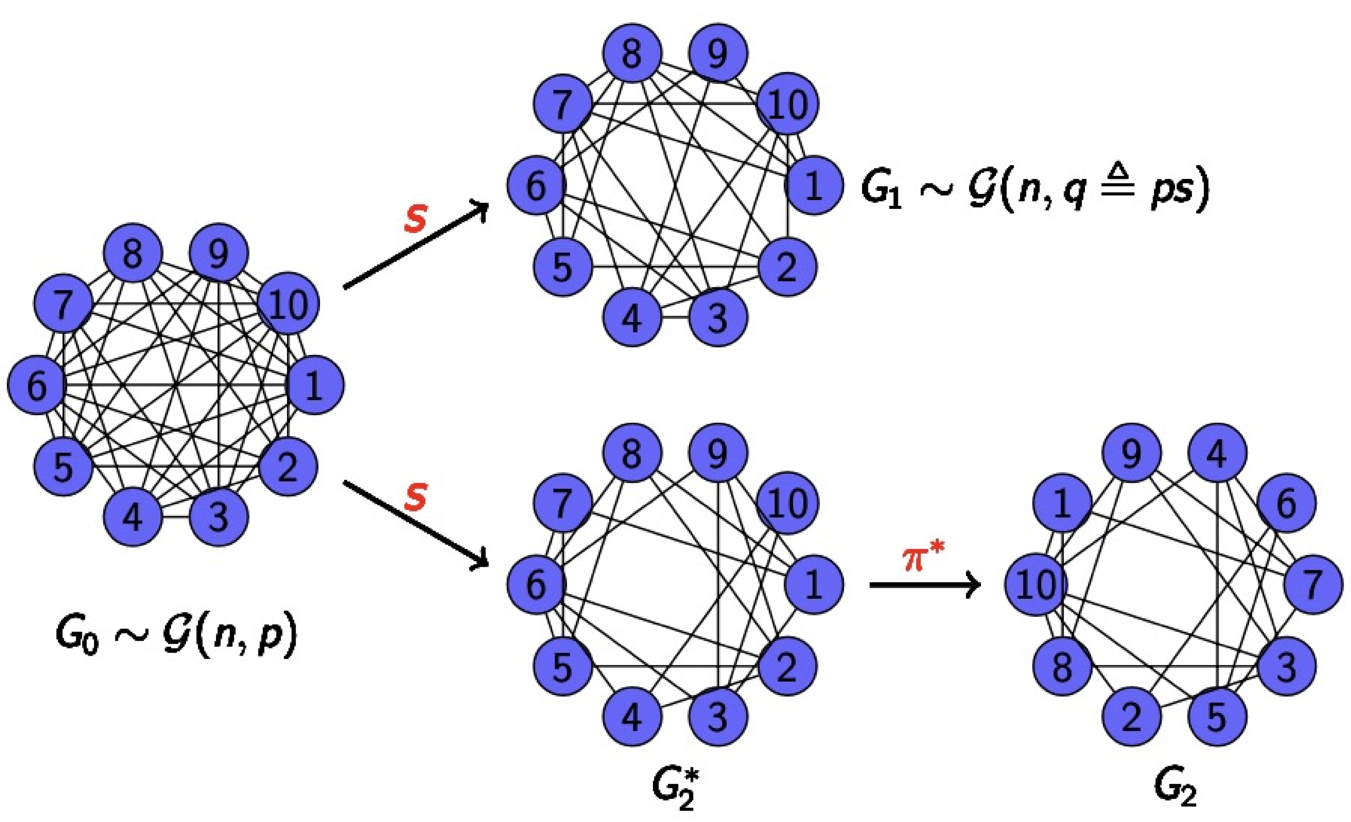
1.Motivations and mathematical settings

- **The correlated Erdős-Rényi graph model** is an extensively studied model that is motivated by various applied fields: network de-anonymization, protein-protein interaction, computer vision...





- Given $n \in \mathbb{N}$ and $p, s \in (0,1)$, define a pair of **latently correlated** Erdős-Rényi graphs (G_1, G_2) on $[n]$ as follows:
(i) Sample $G_0 \sim \mathcal{G}(, n, p)$;
(ii) Independently subsample G_1, G_2^* from G_0 by keeping each edge with probability s, independent of each other.
(iii) Relabel G_2^* by a uniform random permutation π^* to get G_2.



- **Main goal**: recover the matching π^* **as good as possible** based on the sole observation of (G_1, G_2).

2.Previous work

- There are three types of matching recovery studied in the literature:
(i) **Exact recovery**: recover the entire π^*;
(ii) **Almost exact recovery**: recover a $1 - o(1)$ fraction of π^*;
(iii) **Partial recovery**: recover a positive fraction of π^*.
- Previous work mostly focus on **determining the thresholds** for the above types of recovery, and the transitions have been well-understood [WXY22, DD23].
- In the dense regime $p = n^{-o(1)}$, there is a **sharp phase transition** in s that impossibility of **partial recovery** *suddenly* transits to achievablity of **almost exact recovery** (the **All-or-Nothing** phenomenon).
- In the sparse regime $p = n^{-\alpha+o(1)}$ where $0 < \alpha \le 1$ is a fixed constant, There is a non-trivial regime $nps^2 = \Theta(1)$ that **partial recovery** is *achievable* while **almost exact recovery** is *impossible*.

Preprint of this work available at: arXiv:2502.12077.

3.Our focus

- We focus on the regime that only a fraction of π^* can be recovered and study the **optimal recovery fraction** of π^*.
- Specifically, we assume $p = n^{-\alpha+o(1)}$ for a constant $0 < \alpha \le 1$, and s satisfies $nps^2 = \Theta(1)$,

4. Intuitions and preliminaries

- Intuitively, the part of π^* that can be recovered is the **"dense part"** in the intersection graph $\mathcal{H}_{\pi^*}$ of G_1, G_2 through π^*.
- To formulate the above intuition in a more quantitative and rigorous way, we need the following concepts.
- **Balanced load**. For a finite graph $G = (V, E)$, let $\vec{E}$ be the directed edge set. A balanced allocation is a function $\theta : \vec{E} \to [0,1]$ satisfying the following two properties:
(i) $\theta(x \to y) + \theta(y \to x) = 1, \forall (x,y) \in G$.
(ii) Let $\partial\theta(x) = \sum_{(x,y)\in E} \theta(y \to x)$, then for any $(x,y) \in E$,
$$\partial\theta(x) < \partial\theta(y) \quad \Rightarrow \quad \theta(x \to y) = 0\,.$$
- **Fact** [Haj90]. For any finite graph G, ballanced allocations exist and the induced funtion $\partial\theta$ is *unique*. We call the function $\partial\theta : V \to \mathbb{R}$ the **balanced load function**.
- **Fact**. For any finite graph $G = (V, E)$ and any $t > 0$, define
$$f_t(H) = t|E(H)| - |H|\,, \quad H \subset V\,,$$
where $E(H)$ is the induced edge set of G in H. Then,
$$\arg\max f_t(H) = \{v \in V : \partial\theta(v) \ge t^{-1}\}\,.$$

5.Main results

- Consider the intersection graph $\mathcal{H}_{\pi^*} = (\mathcal{V}, \mathcal{E}) \sim \mathcal{G}(n, ps^2)$.
- Fix any $\varepsilon > 0$. For a vertex $v \in V$, we call it *heavy* if $\partial\theta(v) \ge \alpha^{-1} + \varepsilon$, and we call it light if $\partial\theta(v) \le \alpha^{-1} - \varepsilon$.

Theorem. For any $\varepsilon > 0$, $0 < \alpha < 1$ and $\lambda > 1$, assume $p = n^{-\alpha+o(1)}$ and $nps^2 = \lambda$, the following holds:
(i) There exists $\widetilde{\pi} = \widetilde{\pi}(G_1, G_2)$ s.t. with high probability,
$$\#\{v \text{ is a heavy vertex}, \widetilde{\pi}(v) \ne \pi^*(v)\} \le \varepsilon n\,.$$
(ii) There is no $\hat{\pi} = \hat{\pi}(G_1, G_2)$ s.t. with non-vanishing probability,
$$\#\{v \text{ is a light vertex}, \hat{\pi}(v) = \pi^*(v)\} \le \varepsilon n\,.$$

- **Fact** [AS16]. The empirical measure of the balanced load function of $\mathcal{H} \sim \mathcal{G}(n, \lambda/n)$ converges weakly to a limiting measure μ_λ.
- Combining with the above fact, sending $\varepsilon \downarrow 0$ in the theorem yields the following corollary.

Corollary. Under the same assumptions, the optimal recovery fraction is lies in between $\mu_\lambda((\alpha^{-1}, \infty))$ and $\mu_\lambda([\alpha^{-1}, \infty])$.
In particular, when $\mu_\lambda(\{\alpha^{-1}\}) = 0$ (which happens for *all but countably many* α), the optimal recovery fraction is $\mu_\lambda((\alpha^{-1}, \infty))$.

References

[Haj90] B. Hajek, Performance of global load balancing by local adjustment, *IEEE Trans. Info. Theory*, 1990.
AS16 V. Anantharam and J. Salez. The densest subgraph problem in sparse random graphs. *Ann. Appl. Prob.*, 2016.
[WXY22] Y. Wu, J. Xu, and S. H. Yu. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Trans. Info. Theory*, 2022.
[DD23] J. Ding and H. Du, Partial recovery threshold for correlated random graphs, *Ann. Stat.*, 2023.