第九届北大–清华统计论坛

The 9th Peking–Tsinghua Joint Statistics Colloquium (2025)

Testing Correlation in Graphs by Counting Bounded Degree Motifs

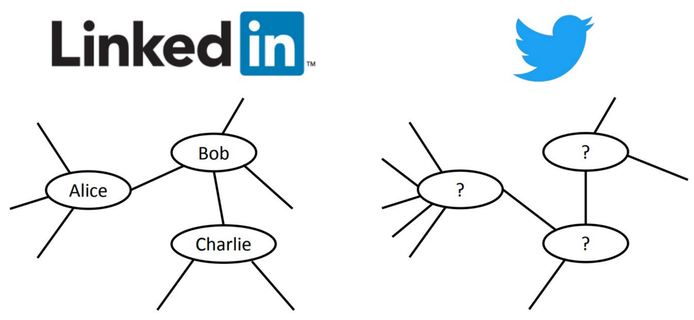
Dong Huang[1] and Pengkun Yang[1]

Presented by **Dong Huang** Email: hd23@mails.tsinghua.edu.cn

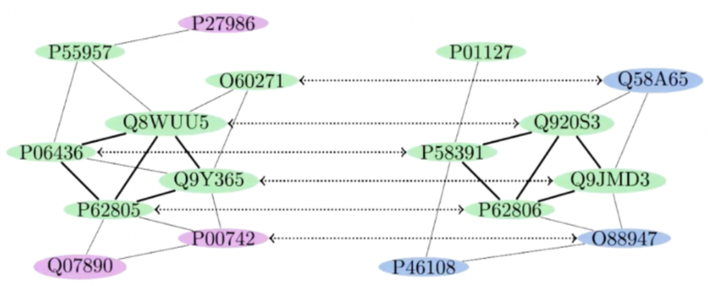
Department of Statistics and Data Science, Tsinghua University[1]

Backgrounds and Motivations

- Questions arise from various fields:
  - **Social network de-anonymization:** deciding whether two friendship networks on different social platforms have similarities [1].



  - **Protein interaction network:** detecting protein-protein interactions between two species [2].



Human network Mouse network

  - **Computer vision:** determining whether two graphs represent the same object under different rotations [3].
  - **Natural language processing:** the ontology alignment problem refers to uncovering the correlation between knowledge graphs that are in different languages [4].
- We are interested in testing correlation between two graphs.

Problem Settings

Models

- Erdős-Rényi graph $\mathcal{G}(n,p)$: graph on n vertices where each edge connects with probability $0<p<1$ independently.
- **Correlated Erdős-Rényi graph** $\mathcal{G}(n,p,\rho)$:
  - $G_1, G_2 \sim \mathcal{G}(n,p)$;
  - Latent bijective mapping π^* on vertices set: $\pi^*: V(G_1) \mapsto V(G_2)$;
  - $\mathrm{Corr}(uv, \pi^*(u)\pi^*(v)) = \rho \in [0,1]$ for any $u,v \in V(G_1)$.

Goals

- Hypothesis testing problem:
  - $\mathcal{H}_0$: $G_1, G_2 \overset{\text{i.i.d.}}{\sim} \mathcal{G}(n,p)$ (corresponding distribution: $\mathcal{P}_0$).
  - $\mathcal{H}_1$: $G_1, G_2 \sim \mathcal{G}(n,p,\rho)$ (corresponding distribution: $\mathcal{P}_1$).
- Detection criteria: there exists some testing statistic $\mathcal{T}(G_1,G_2)$ and a threshold τ such that
$$\mathcal{P}_0\left(\mathcal{T}(G_1,G_2)\geq\tau\right)+\mathcal{P}_1\left(\mathcal{T}(G_1,G_2)<\tau\right)\leq 0.05.$$
- **Goals:** test $\mathcal{H}_0$ against $\mathcal{H}_1$ in **polynomial time**.

Challenges

- Polynomial-time algorithm design: likelihood ratio test requires search over $n!$ potential permutations.
- Single observation: only one pair of (G_1,G_2) is observed.
- Type I and II errors control: simultaneously control of two kinds of errors.

Previous Results

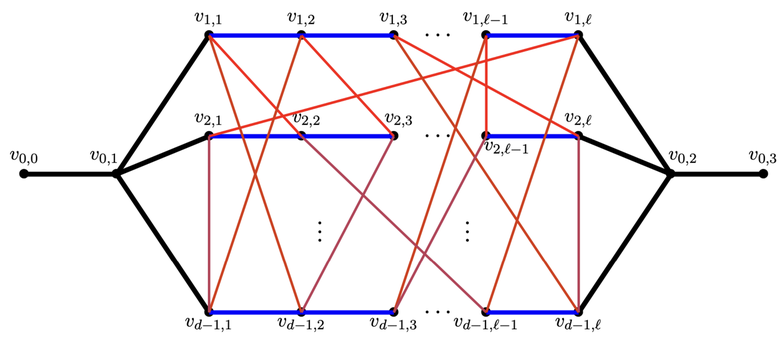
- **Counting balanced graphs:** succeeds for correlation detection when $p \in [n^{-1+\epsilon}, n^{-1+1/153}] \cup [n^{-1/3}, n^{-\epsilon}]$ for any constant ρ in polynomial time [5].
- **Counting trees:** succeeds for correlation detection when $p \geq n^{-1+o(1)}$ and $\rho^2 \geq \alpha \approx 0.338$ in polynomial time [6].

Our Results: Arbitrary Constant

For any constant correlation $\rho \in (0,1)$, when $p \geq n^{-2/3}$, there exists an estimator $\mathcal{T}(G_1,G_2)$ computable in polynomial time and a threshold τ such that
$$\mathcal{P}_1(\mathcal{T}(G_1,G_2)<\tau)+\mathcal{P}_0(\mathcal{T}(G_1,G_2)\geq\tau)\leq 0.05.$$

Key Components

- Graph homomorphism:
  - $\bar{G}$: weighted graph with weight edge $\beta_{uv}(\bar{G}) = \mathbf{1}_{\{uv\in E(G)\}} - p$.
  - $$\mathsf{hom}_\varphi(\mathsf{M},\bar{G}) = \prod_{uv\in E(\mathsf{M})} \beta_{\varphi(u)\varphi(v)}(\bar{G}).$$
  - $$\mathsf{inj}(\mathsf{M},\bar{G}) = \sum_{\substack{\varphi: V(\mathsf{M})\mapsto V(G)\\ \varphi \text{ injective}}} \mathsf{hom}_\varphi(\mathsf{M},\bar{G})$$
- Bounded degree motif $\mathcal{M}(N_\mathsf{v}, N_\mathsf{e}, d)$: a special motif family with N_v vertices, N_e edges and bounded degree d.



- Testing statistic:
$$\mathcal{T}(G_1,G_2) = \sum_{\mathsf{M}\in\mathcal{M}(N_\mathsf{v},N_\mathsf{e},d)} \omega_\mathsf{M}\mathsf{inj}(\mathsf{M},\bar{G}_1)\mathsf{inj}(\mathsf{M},\bar{G}_2),$$
where ω_M is a weight and the testing statistic can be regarded as a weighted inner product between vectors $\left[\mathsf{inj}(\mathsf{M},\bar{G}_1)\right]_{\mathsf{M}\in\mathcal{M}}$ and $\left[\mathsf{inj}(\mathsf{M},\bar{G}_2)\right]_{\mathsf{M}\in\mathcal{M}}$.
- Errors control: pick $\tau = \frac{1}{2}\mathbb{E}_{\mathcal{P}_1}\left[\mathcal{T}(G_1,G_2)\right]$ and use Chebyshev's inequality.
- Time complexity: $n^{C\rho^{-2d/(d-2)}}$; no $n^{-o(\rho^{-1})}$-time algorithm conjectured [7].

Future Directions

- **General settings:** inhomogeneous and correlated connection probability in a single graph; time-varying structures.
- **Graph alignment problem:** using the bounded degree estimator to recover latent permutation when given $G_1, G_2 \sim \mathcal{G}(n,p,\rho)$.
- **General graph models:** extensions to multiple graphs, random geometric model, graphon model.

References

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE, 2008.

[2] E. Kazemi, H. Hassani, M. Grossglauser, and H. Pezeshgi Modarres, "Proper: global protein interaction network alignment through percolation matching," *BMC bioinformatics*, vol. 17, pp. 1–16, 2016.

[3] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 26–33, IEEE, 2005.

[4] A. Haghighi, A. Y. Ng, and C. D. Manning, "Robust textual inference via graph matching," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 387–394, 2005.

[5] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng, "(Nearly) efficient algorithms for the graph matching problem on correlated random graphs," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Testing network correlation efficiently via counting trees," *The Annals of Statistics*, vol. 52, no. 6, pp. 2483–2505, 2024.

[7] J. Ding, H. Du, and Z. Li, "Low-degree hardness of detection for correlated Erdős-Rényi graphs," *arXiv preprint arXiv:2311.15931*, 2023.