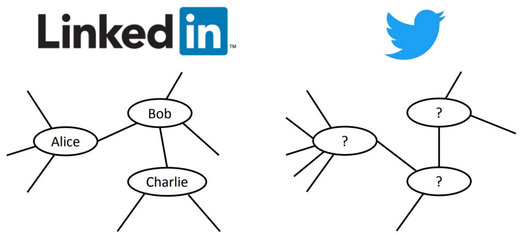
Sample Complexity of Correlation Detection in the Gaussian Wigner Model

Dong Huang Pengkun Yang

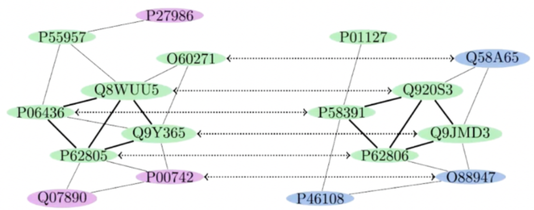
Tsinghua University

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.

Challenges

- **Lack of data.** The entire network is often unavailable due to API limitations in social network analysis.
- **Testing costs.** The data can be expensive in protein interaction networks.
- **Visualization.** The original graph is sometimes too large to be displayed on a screen.

A natural solution: graph sampling.

Problem Settings

Models

- Gaussian Wigner model: graph with n vertices and each weighted edges $\beta_{uv}(\mathbf{G})$ follow independent standard normals.
- Correlated Gaussian Wigner model:
  - $\mathbf{G}_1, \mathbf{G}_2$ follows Gaussian Wigner model;
  - Latent bijective mapping π^* on vertices set $\pi^* : V(\mathbf{G}_1) \mapsto V(\mathbf{G}_2)$;
  - $\mathrm{Corr}(\beta_{uv}(\mathbf{G}_1), \beta_{\pi^*(u)\pi^*(v)}(\mathbf{G}_2)) = \rho \in [0, 1]$ for any $u, v \in V(\mathbf{G}_1)$.

Graph Sampling

- Randomly sample two induced subgraphs $G_1 \subseteq \mathbf{G}_1$ and $G_2 \subseteq \mathbf{G}_2$ with $V(G_1) = V(G_2) = s$.

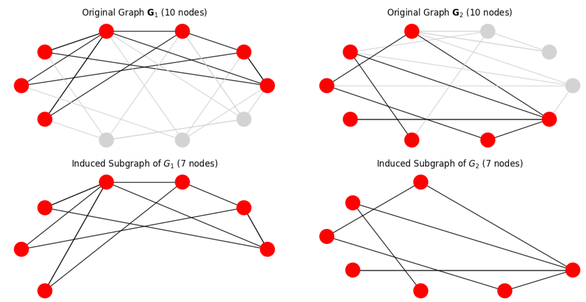


Figure 1. Induced subgraph: a subset of the vertices along with all the edges between them from the original graph.

Goal

- Hypothesis testing problem:
  - $\mathcal{H}_0$: $\mathbf{G}_1, \mathbf{G}_2$ follows independent Gaussian Wigner model;
  - $\mathcal{H}_1$: $\mathbf{G}_1, \mathbf{G}_2$ follows correlated Gaussian Wigner model with correlation ρ.
  - $\mathcal{Q}$: distribution of sampling subgraphs (G_1, G_2) under $\mathcal{H}_0$;
  - $\mathcal{P}$: distribution of sampling subgraphs (G_1, G_2) under $\mathcal{H}_1$.
- Detection criterion for a test statistic $\mathcal{T}(G_1, G_2)$ and a threshold τ:

$$\lim_{n\to\infty} \mathcal{P}(\mathcal{T}(G_1, G_2) < \tau) + \mathcal{Q}(\mathcal{T}(G_1, G_2) \geq \tau) \leq 0.05$$

Goal: determine the sample size $s = s(n, p, \rho)$ required for the hypothesis testing problem.

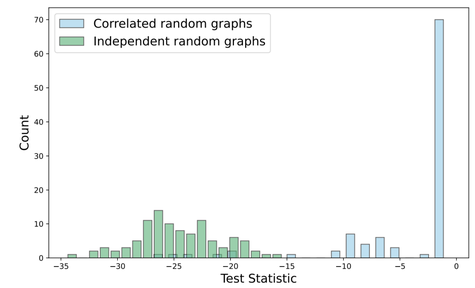
Our Results: Statistical Analysis

Theorem: There exist constants $\overline{C}, \underline{C}$ such that, for any $0 < \rho < 1$,

- if $s^2 \geq \overline{C}\left(\frac{n \log n}{\log(1/(1-\rho^2)} \vee n\right)$, then $\mathsf{TV}(\mathcal{P}, \mathcal{Q}) \geq 0.95$;
- if $s^2 \leq \underline{C}\left(\frac{n \log n}{\log(1/(1-\rho^2)} \vee n\right)$, then $\mathsf{TV}(\mathcal{P}, \mathcal{Q}) \leq 0.05$.

Our Results: Iterative Algorithm

- Step 1: Match pairs of small cliques of size K;
- Step 2: Aggregate matches from Step 1 to identify the seed set;
- Step 3: Iteratively expand the node mappings using the seed set obtained in Step 2.
- Overall time complexity: $O(s^{2K})$.



Future Directions

- **Extension to Erdős-Rényi model**: binary edge settings;
- **Theoretical analysis of the efficient algorithm**: the statistical analysis of the polynomial-time algorithm remains unknown;
- **Other graph models:** multiple correlated graphs, geometric graph, 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.