Spring 2023

**S&DS 684: Statistical inference on graphs**

**Syllabus**

| | |
|---|---|
| Schedule: | Tuesday 4-550pm, 17 Hillhouse Rm 03 |
| Professor: | Yihong Wu yihong.wu@yale.edu, Rm 235 Dunham Lab (10 Hillhouse) |
| Office hours: | by appointment |
| Website: | http://www.stat.yale.edu/~yw562/teaching/684/ |

# 1 Content

An emerging research thread in statistics and machine learning deals with recovering latent structures from combinatorial data represented by graphs or matrices. This graduate-level course provides an introduction to the mathematical and algorithmic tools for studying such problems. We will discuss information-theoretic methods for determining the fundamental limits of various detection and recovery problems on random graphs. Complementing the objective of characterizing the information-theoretical limits, another significant direction is to develop computationally efficient algorithms (such as spectral methods, convex relaxation, and belief propagation) that attain the statistical optimality, or to understand the lack thereof (computational barriers).

Specific topics will include spectral clustering, planted clique and partition problem, community detection on stochastic block models, broadcasting on trees, planted matching problem (linear and quadratic assignment problems), statistical-computational tradeoffs.

We will be following this set lecture notes:

Yihong Wu and Jiaming Xu, "Statistical inference on graphs: Selected Topics", working draft, available at http://www.stat.yale.edu/~yw562/teaching/stats-graphs.pdf

Tentative outline

1. **Introduction**: detection-recovery-estimation, phase transition and sharp threshold, statistical-computational gaps
2. **Cliques in Erdös-Rényi graphs**: first and second moment methods, Grimmett-McDiarmid's greedy algorithm
3. **Spectral methods**: preliminaries from linear algebra, perturbation bound, application to clustering
4. **Basic Random Matrix Theory**: spectral norm and concentration of measure
5. **Planted clique**: degree test, spectral methods
6. **Semidefinite programming (SDP) relaxation I**: KKT conditions and dual certificates, convexified maximum likelihood
7. **Broadcasting on trees**: branching process, Kesten-Stigum bound, belief propagation (BP)

8. **Community detection**: stochastic block models, correlated recovery and mutual information, detection threshold and approximate cycle counting, truncated second moment method, from BP to non-backtracking matrices

9. **SDP relaxation II**: exact recovery threshold, Grothendieck inequality and consequences on clustering, robustness in semi-random models

10. **Planted assignment problem**: bipartite matching (linear assignment), graph matching (quadratic assignment), algorithms.

11. **Computational limits**: Polynomial-time randomized reduction, Planted dense subgraph problem, Sparse PCA

# 2 Administrivia

1. Course prerequisites: Maturity with probability theory. Familiarity with mathematical statistics.

2. This is a graduate-level course. Undergraduate students need express permission from the instructor to take the class.

3. Final project: submitting a report based on on either reading a paper or a standalone research project.

4. Grading: 30% participation, 30% homeworks, 40% final project.