1 Content

An emerging research thread in statistics and machine learning deals with finding latent structures from data represented by graphs or matrices. This course will provide an introduction to mathematical and algorithmic tools for studying such problems. We will discuss information-theoretic methods for determining the fundamental limits. Complementing this objective of understanding the fundamental limits, another significant direction is to develop computationally efficient procedures that attain the statistical optimality, or to understand the lack thereof. We will discuss computationally efficient procedures including spectral methods, semidefinite programming relaxations, Linear/quadratic programming relaxations, message passing (belief propagation) algorithms, etc. Towards the end we will also discuss the recent trend of combining the statistical and algorithmic perspectives and the computational barriers in a series of statistical problems on large matrices and random graphs.

Specific topics will include spectral clustering, planted clique and partition problem, community detection on stochastic block models, hidden Hamiltonian cycle problem, planted bipartite graph matching problem, and noisy graph isomorphism.

Tentative outline

1. Introduction: detection-recovery-estimation, sharp thresholds
2. Spectral methods: preliminaries from linear algebra, perturbation bound, application to clustering
3. Planted clique: degree test, spectral methods, message passing algorithms
4. Information-theoretic tools: Entropy and mutual information, total variation, Hellinger distance, Kullback-Leibler (KL) divergence and variational characterizations, (strong) data processing inequalities, Pinsker and related inequalities, rate-distortion function
5. Community detection: stochastic block models, correlated recovery and mutual information, almost exact and exact recovery, first and second moment methods
6. Semidefinite programming (SDP) relaxation: KKT conditions and exact recovery threshold, Grothendieck inequality and consequences on clustering, robustness in semi-random models
7. Broadcasting on trees: branching process, Kesten-Stigum bound, mutual information bound, connection to community detection, belief propagation algorithm
8. Hidden Hamiltonian cycle, bipartite graph matchings, and graph isomorphism problems: sharp information-theoretic limits, analysis of linear/quadratic programming relaxations, Stieltjes transform of random matrix, applications to DNA sequencing, particle tracking, and network alignment.

9. Computational limits: Polynomial-time randomized reduction, Planted dense subgraph problem

2 Administrivia

1. Course prerequisites: Maturity with probability theory and linear algebra. Familiarity with statistical theory, optimization, and algorithms.

2. Grading: 30% participation, 30% homework, 40% final project.

3. Class participation includes attending classes on time, raising good questions or providing good answers to questions, pointing out errors in the lecture notes, etc.

4. Final project: submit a report based on either reading a paper or a standalone research project. Some guidelines and project ideas will be posted later. Students are encouraged to apply the knowledge learned in this class to their application domains.
   - Team of 1-2 students
   - Mid-term project report due March 18 (after spring break)
   - Project presentation (15 mins each team) on April 10 and 15
   - Final project report due April 27

5. Materials: Lecture notes are available online https://faculty.fuqua.duke.edu/~jx77/Course_graphs.html and will be updated throughout the course. Additional reading materials will be posted online.

6. March 4 and 6 classes will be in RAND classroom just below seminar G.

7. Classes will be video recorded using Duke’s Panopto system.