Fall 2018

S&DS 684: Statistical inference on graphs

Syllabus

Schedule:	Wednesday 2:30-5pm, 24 Hillhouse Rm 107
First lecture:	Wednesday, Aug 29 2018
Professor:	Yihong Wu yihong.wu@yale.edu, Rm 235 Dunham Lab (10 Hillhouse)
Office hours:	by appointment
Website:	http://stat.yale.edu/~yw562/684.html

1 Content

An emerging research thread in statistics and machine learning deals with finding latent structures from data represented in 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, as well as methodologies for attaining these limits, including spectral methods, semidefinite programming relaxations, message passing algorithms, etc. Specific topics will include spectral clustering, planted clique and partition problem, sparse PCA, community detection on stochastic block models, statistical-computational tradeoffs.

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. 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.

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, data processing principle, 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

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

9. **Ranking and sorting**

10. **Advanced topics** (TBD): Hidden Hamiltonian cycle problem, Noisy graph matching, small-world graphs, Gaussian graphical models

2 Administrivia

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

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

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

4. Materials: Lecture notes and additional reading materials will be posted online.