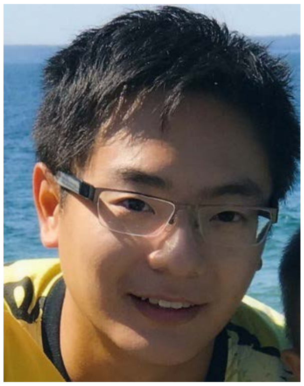
Biostatistics Seminars

Fall 2017

SPECIAL SEMINAR



**Tianxi Li, PhD candidate**
Department of Statistics
University of Michigan

# Statistical Tools for Analyzing Network-Linked Data

Thursday, December 14, 2017
3:00 pm – 4:00 pm
Purvis Hall, 1020 Pine Ave. West, Room 25

ALL ARE WELCOME

**Abstract**:
While classic statistical tools such as regression and graphical models have been well studied, they are no longer applicable when the observations are connected by a network, an increasingly common situation in modern complex datasets. We develop the analogue of loss-based prediction models and graphical models for such network-linked data, by a network-based penalty that can be combined with any number of existing techniques. We show, both empirically and theoretically, that incorporating network information improves performance on a variety of tasks under the assumption of network cohesion, the empirically observed phenomenon of linked nodes acting similarly. Computationally efficient algorithms are developed as well for implementing our proposal. We also consider the general question of how to perform cross-validation and bootstrapping on networks, a long-standing open problem in network analysis. Model selection and tuning for many tasks can be performed through cross-validation, but splitting network data is non-trivial, since removing links leads to a potential change in network structure. We propose a new general cross-validation strategy for networks, based on repeatedly removing edge values at random and then applying matrix completion to reconstruct the full network. We obtain theoretical guarantees for this method under a low rank assumption on the underlying edge probability matrix, and show that the method is computationally efficient and performs well for a wide range of network tasks, in contrast to previously developed approaches that only apply under specific models. Several real-world examples will be discussed throughout the talk, including the effect of friendship networks on adolescent marijuana usage, phrases that can be learned with the help of a collaboration network of statisticians as well as statistician communities extracted from a citation network.

# Biostatistics Seminars

## Fall 2017

### Bio:

Tianxi Li is a PhD candidate from the Department of Statistics at the University of Michigan. His main research interests include statistical network analysis and statistical machine learning, with focus on developing interpretable and computationally efficient methods to analyze complex data sets. Before his PhD student, he was an applied researcher in Microsoft Bing Search working on design and statistical inference of online randomized experiments.

...2/2