Department of
Biostatistics

NYU | GPH

2022 Workshop on Statistical Network Analysis & Beyond (SNAB2022)

Wednesday, Aug 3
to Friday Aug 5, 8am ET

New York University
School of Public Health
708 Broadway, 3rd floor

2022 Workshop on Statistical Network Analysis & Beyond (SNAB2022)

WEDNESDAY, AUG 3 - FRIDAY, AUG 5 | 8AM ET

NEW YORK UNIVERSITY
SCHOOL OF GLOBAL
PUBLIC HEALTH

708 BROADWAY, 3RD FLOOR

Welcome

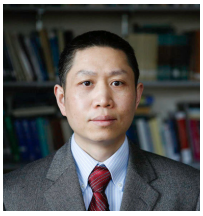
The 2022 Workshop on Statistical Network Analysis and Beyond (SNAB2022) is hosted by the NYU School of Global Public Health, Department of Biostatistics. It is the second workshop in the SNAB series with the first one held virtually in January 2021.
We are bringing together researchers on network analysis and beyond to exchange ideas and recent works for SNAB2022. The workshop will cover topics including analysis of social networks and biological networks, tensor analysis, and deep learning.

Organizing Committee

Yang Feng, Ph.D
Associate Professor
NYU GPH Department of Biostatistics

Jiashun Jin, Ph.D
Professor of Statistics
Carnegie Mellon University

Ji Zhu, Ph.D
Professor of Statistics
University of Michigan



Hui Zou, Ph.D
Professor of Statistics
University of Minnesota

Sponsors

NYU
SCHOOL OF GLOBAL PUBLIC HEALTH

Biostatistics
NYU | GPH

ims

Schedule Overview

Wednesday, August 3, 2022

8:00 AM	**Breakfast**
9:00 AM	**Session 1 (Jiashun Jin)** Short Course: *Analysis of Social Networks and Text Data (1)*
10:30 AM	**Break**
	Session 2 (Tracy Ke)
11:00 AM	Short Course: *Analysis of Social Networks and Text Data (2)*
12:30 PM	**Lunch**
	Session 3 (Jiashun Jin)
2:00 PM	Short Course: *Analysis of Social Networks and Text Data (3)*
3:30 PM	**Break**
	Session 4 (Tracy Ke)
4:00 PM	Short Course: *Analysis of Social Networks and Text Data (4)*

Thursday, August 4, 2022

8:50 AM	**Rebecca Betensky,** Chair of the Department of Biostatistics
	Session 1 (Chair: Yang Feng)
9:00 AM	**Cun-Hui Zhang:** *Tensor PCA in High Dimensional CP Models*
9:30 AM	**Tracy Ke:** *A Self-normalizing Cycle Count Statistic*
10:00 AM	**Stefan Wager:** *Random Graph Asymptotics for Treatment Effect Estimation under Network Interference*
10:30 AM	**Break**
	Session 2 (Chair: Jiashun Jin)
11:00 AM	**Carey Priebe:** *Discovering underlying dynamics in time series of networks*
11:30 AM	**Emma Zhang:** Gaussian Graphical Regression Models with Covariates
12:00 PM	**Yuan Zhang:** Edgeworth expansions for network moments
12:30 PM	**Lunch / Session 3 (Poster Session)**
	Session 4 Chair: Ji Zhu
2:00 PM	**Jianqing Fan:** *Optimal rates for Robust Deep ReLU networks*
2:30 PM	**Rajarshi Mukherjee:** *Global Testing Against Sparse Alternatives under Ising Models*
3:00 PM	**Wanjie Wang:** *Community Detection on Networks with Covariates*
3:30 PM	**Coffee Break**
	Session 5 Chair: Emma Zhang
4:00 PM	**Tian Zheng:** *Scalable Community Detection in Massive Networks using Aggregated Relational Data*
4:30 PM	**Patrick Rubin-Delanchy:** *Manifold structure in graph embeddings*
5:00 PM	**Daniel Sewell:** *Automated detection of edge clusters*
5:30 PM	**Arash Amini:** *Bayesian community detection for networks with covariates*

Schedule Overview

Friday, August 5, 2022

Session 1 Chair: Wanjie Wang

9:00 AM — **Eric Kolaczyk:** *Coevolving Latent Space Network with Attractors Models for Polarization*

9:30 AM — **Haolei Weng:** *Spectral clustering via adaptive layer aggregation for multi-layer networks*

10:00 AM — **Weijing Tang:** *Population-level Balance in Signed Networks*

10:30 AM — **Coffee Break**

Session 2 Chair: Haolei Weng

11:00 AM — **Elizaveta Levina:** *Resampling methods for networks*

11:30 AM — **Zuofeng Shang:** *Variational Nonparametric Testing in Functional Stochastic Block Model*

12:00 PM — **Xiaodong Li:** *Selecting the Number of Communities in Networks via Stepwise Matrix Scaling and Spectral Thresholding*

12:30 PM — **Lunch Break**

Session 3 Chair: Hai Shu

2:00 PM — **Ali Shojaie:** *Learning Causal Networks from Biological Data*

2:30 PM — **Lei Liu:** *Simultaneous Cluster Structure Learning and Estimation of Heterogeneous Graphs for Matrix-variate fMRI Data*

3:00 PM — **Wen Zhou:** *High Dimensional Clustering via Latent Semiparametric Mixture Models*

Session 4 Chair: Wen Zhou

4:00 PM — **Harrison Zhou:** *Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering*

4:30 PM — **David Choi:** Causal Inference in Experiments on Networks

5:00 PM — **Hai Shu:** *Orthogonal Common-Source and Distinctive-Source CoVariance Decomposition Between High-Dimensional Data Views*

5:30 PM — **Best Poster Award Presentation**

Full Program

Wednesday, August 3, 2022

Short Course:
Analysis of Social Networks and Text Data

Jiashun Jin, Carnegie Mellon University;
Tracy Ke, Harvard University

Abstract: This short course consists of four lectures, each is 80 minutes long and will be taught by both instructors. The lectures will cover recent models, methods and theory, and applications for analyzing social networks and text data.

For network models, we primarily focus on the Degree-Corrected Block Model (DCBM) and the more recent Degree-Corrected Mixed-Membership model (DCMM). DCBM and DCMM include the classical Stochastic Block Model (SBM) as a special case, but are much broader and so fit real networks better. We also discuss other network models. For text data, we focus on the probabilistic Latent Semantic Indexing (pLSI) model. We also discuss the popular Latent Dirichlet Allocation (LDA) model, which is a variant of the pLSI model. For methods and theory, we discuss several research topics, including but not limited to network community detection, mixed-membership estimation, network global testing, estimating the number of communities, dynamic networks, hierarchical community detection, and topic modeling, among others.

The data sets we use for applications have two parts. Part I consists of a number of network data sets frequently seen in the recent literature (e.g., ``karate", ``weblog", ``football"). Part II is the publication data set for statisticians. The data set contains the citation and bibtex information (e.g., title, abstract, author information) of 83,331 papers in 36 journals in statistics and related fields from 1975 to 2015. The data set can be used to construct many different kinds of networks, and the attributes of title and abstract can also be used for text learning. The data set is now freely available.

Lecture 1 (9:00 - 10:30am): Focuses on community detection. It concerns the setting where all nodes in the network do not have mixed-memberships. The goal is to use the network adjacency matrix to cluster all nodes of the network into K clusters or communities, where K is given.

Lecture 2 (11:00am - 12:30pm): Focuses on mixed-membership estimation. It deals with the more realistic setting where some nodes may have mixed-memberships, and the goal is to estimate the membership vector of each node (in connection to Lecture 1, community detection can be viewed as a special case of membership estimation). We introduce Mixed-SCORE as a recent approach to membership estimation. We also extend Mixed-SCORE to a dynamic version and use it to estimate the research trajectory of representative statisticians.

Lecture 3 (2:00 - 3:30pm): Focuses on estimation of the number of communities K. Most existing approaches to community detection and membership estimation require to know K a priori, but unfortunately, K is frequently unknown in real applications. We introduce a cycle count approach to estimating K. We show that this approach is optimal, broadly implementable, and adaptive to different levels of network sparsity. We also discuss network global testing and network goodness-of-fit.

Lecture 4 (4:00 - 5:30pm): Focuses on text learning. We introduce Topic-SCORE as a fast and easy-to-use approach to topic modeling. We also showcase the use of Topic-SCORE on the publication data of statisticians, treating the paper abstracts as text documents. The data analysis suggests 11 representative topics in the field of statistics. We develop an approach to ranking different research topics and use the estimated topic weights to help predict the future citations of a given paper.

Thursday, August 4, 2022

Session 1 (9:00 - 10:30am)

Tensor PCA in High Dimensional CP Models
Cun-Hui Zhang, Rutgers University

Abstract: The CP decomposition for high dimensional non-orthogonal spike tensors is an important problem with broad applications across many disciplines. However, previous works with theoretical guarantee typically assume restrictive incoherence conditions on the basis vectors for the CP components. We propose new computationally efficient composite PCA and concurrent orthogonalization algorithms for tensor CP decomposition with theoretical guarantees under mild incoherence conditions. The composite PCA applies the principal component or singular value decompositions twice, first to a matrix unfolding of the tensor data to obtain singular vectors and then to the matrix folding of the singular vectors obtained in the first step. It can be used as an initialization for any iterative optimization schemes for the tensor CP decomposition. The concurrent orthogonalization algorithm iteratively estimates the basis vector in each mode of the tensor by simultaneously applying projections to the orthogonal complements of the spaces generated by others CP components in other modes. It is designed to improve the alternating least squares estimator and other forms of the high order orthogonal iteration for tensors with low or moderately high CP ranks. Our theoretical investigation provides estimation accuracy and convergence rates for the two proposed algorithms. Our implementations on synthetic data demonstrate significant practical superiority of our approach over existing methods.

A Self-normalizing Cycle Count Statistic
Tracy Ke, Harvard University

Abstract: Consider a large binary data matrix. We introduce a self-normalizing cycle count (SCC) statistic, and show that it is asymptotically N(0,1) in many settings. We show that SCC can be useful for solving several recent problems in network analysis, including network global testing, estimating the number of communities, and goodness-of-fit. We present several optimality results on SCC and support our results with a number of real data examples.

Random Graph Asymptotics for Treatment Effect Estimation under Network Interference
Stefan Wager, Stanford University

Abstract: The network interference model for causal inference places all experimental units at the vertices of an undirected exposure graph, such that treatment assigned to one unit may affect the outcome of another unit if and only if these two units are connected by an edge. This model has recently gained popularity as means of incorporating interference effects into the Neyman--Rubin potential outcomes framework; and several authors have considered estimation of various causal targets, including the direct and indirect effects of treatment. In this paper, we consider large-sample asymptotics for treatment effect estimation under network interference in a setting where the exposure graph is a random draw from a graphon. When targeting the direct effect, we show that -- in our setting -- popular estimators are considerably more accurate than existing results suggest, and provide a central limit theorem in terms of moments of the graphon. Meanwhile, when targeting the indirect effect, we leverage our generative assumptions to propose a consistent estimator in a setting where no other consistent estimators are currently available. We also show how our results can be used to conduct a practical assessment of the sensitivity of randomized study inference to potential interference effects. Overall, our results highlight the promise of random graph asymptotics in understanding the practicality and limits of causal inference under network interference.

Thursday, August 4, 2022 (cont.)
Session 2 (11:00am - 12:30pm)

Discovering underlying dynamics in time series of networks
Carey Priebe, Johns Hopkins University

Abstract: Understanding dramatic changes in the evolution of networks is central to statistical network inference, as underscored by recent challenges of predicting and distinguishing pandemic-induced transformations in organizational and communication networks. We consider a joint network model in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits low-dimensional manifold structure under a suitable notion of distance. This distance can be approximated by a measure of separation between the observed networks themselves, and there exist consistent Euclidean representations for underlying network structure, as characterized by this distance, at any given time. These Euclidean representations permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classical setting. We illustrate our methodology with real and synthetic data, and identify change points corresponding to massive shifts in pandemic policies in a communication network of a large organization.

Gaussian Graphical Regression Models with Covariates
Emma Zhang, University of Miami

Abstract: Though Gaussian graphical models have been widely used in many scientific fields, relatively limited progress has been made to flexibly link graph structures to external covariates. In this talk, we describe a new Gaussian graphical regression model that relates the conditional dependence structure to covariates, discrete and continuous, of high dimensions. In the context of co-expression quantitative trait locus (QTL) studies, our method can determine how genetic variants and clinical conditions modulate the subject-level gene co-expressions and recover both the population-level and subject-level gene co-expression networks. Under the proposed framework, we address problems including efficient computation with a simultaneous sparsity structure, and error rate and variable selection consistency quantification. Finally, the utility of our proposed method is demonstrated through an application to a co-expression QTL study with brain cancer patients.

Edgeworth expansions for network moments
Yuan Zhang, The Ohio State University

Abstract: Network method of moments is an important tool for nonparametric network inference. However, there has been little investigation on accurate descriptions of the sampling distributions of network moment statistics. In this paper, we present the first higher-order accurate approximation to the sampling CDF of a studentized network moment by Edgeworth expansion. In sharp contrast to classical literature on noiseless U-statistics, we show that the Edgeworth expansion of a network moment statistic as a noisy U-statistic can achieve higher-order accuracy without nonlattice or smoothness assumptions but just requiring weak regularity conditions. Behind this result is our surprising discovery that the two typically-hated factors in network analysis, namely, sparsity and edgewise observational errors, jointly play a blessing role, contributing a crucial self-smoothing effect in the network moment statistic and making it analytically tractable. Our assumptions match the minimum requirements in related literature. For sparse networks, our theory shows that our empirical Edgeworth expansion and a simple normal approximation both achieve the same gradually depreciating Berry–Esseen-type bound as the network becomes sparser. This result also significantly refines the best previous theoretical result. For practitioners, our empirical Edgeworth expansion is highly accurate and computationally efficient. It is also easy to implement and convenient for parallel computing. We demonstrate the clear advantage of our method by several comprehensive simulation studies. As a byproduct, we also provide a finite-sample analysis of the network jackknife.

Program - Student Posters

Thursday, August 4, 2022 (cont.)

Session 3 - Student Posters (12:30 - 2:00pm)

Yongkai Chen (University of Georgia):
Fluid Correlation: A Novel Way to Assess the Dynamic Association

Huimin Cheng (University of Georgia):
Graphon convolutional network: A highly efficient learner for random graph

Nate Josephs (Yale University):
Communication network dynamics in a large organizational hierarchy

Hancong Pan (Boston University):
Understanding Social Dynamics Through Coevolving Latent Space Networks With Attractors

Jiajin Sun (Columbia University):
Longitudinal Latent Space Model

Paxton Turner (Harvard University):
Detection of Heterogeneous Documents in a Corpus

Jingming Wang (Harvard University):
Optimal Network Membership Estimation Under Severe Degree Heterogeneity

Shuoyang Wang (Yale University):
Robust Deep Neural Network Estimation for Multi-dimensional Functional Data

William Wang (Massachusetts Institute of Technology):
Graphon Estimation from Degree-Censored Network Data

Owen Ward (Columbia University):
Online Community Detection for Events on Networks

Shushan Wu (University of Georgia):
Personalized Risk Score Prediction of a Covid-19 Population-based Contact Tracing Network

Yisha Yao (Yale University):
Contraction of a quasi-Bayesian model with shrinkage priors in precision matrix estimation

Thursday, August 4, 2022 (cont.)

Session 4 (2:00-3:30pm)

Optimal rates for Robust Deep ReLU networks
Jianqing Fan, Princeton University

Abstract: This talk is on the stability of deep ReLU neural networks for nonparametric regression under the assumption that the noise has only a finite p-th moment. We unveil how the optimal rate of convergence depends on p, the degree of smoothness and the intrinsic dimension in a class of nonparametric regression functions with hierarchical composition structure when both the adaptive Huber loss and deep ReLU neural networks are used. This optimal rate of convergence cannot be obtained by the ordinary least squares but can be achieved by the Huber loss with a properly chosen parameter that adapts to the sample size, smoothness, and moment parameters. A concentration inequality for the adaptive Huber ReLU neural network estimators with allowable optimization errors is also derived. To establish a matching lower bound within the class of neural network estimators using the Huber loss, we employ a different strategy from the traditional route: constructing a deep ReLU network estimator that has a better empirical loss than the true function and the difference between these two functions furnishes a low bound. This step is related to the Huberization bias, yet more critically to the approximability of deep ReLU networks. As a result, we also contribute some new results on the approximation theory of deep ReLU neural networks.
(*Joint work with Yihong Gu and Wenxin Zhou)*

Global Testing Against Sparse Alternatives under Ising Models
Rajarshi Mukherjee, Harvard University

Abstract: In this talk, I will discuss the effect of dependence on detecting sparse signals in a concrete class of models. In particular, we will focus on mean-type signals in Ising models and establish how the interplay between the strength, structure, and sparsity of signals determine their detectability under various levels of dependence. The impact of dependence is best illustrated under Mean-Field type models where we observe the effect of a "thermodynamic" phase transition. In particular, critical states of these models exhibit a subtle "blessing of dependence" phenomenon in that one can detect much weaker signals at criticality than otherwise. We also argue that similar results are valid for non-Mean-Field models as well by explicitly analyzing Ising models on lattices in arbitrary but fixed dimensions. Moreover, we develop testing procedures that are broadly applicable to account for dependence and show their asymptotic minimax optimality. Finally, I will also discuss the behavior of sharp constants of detection boundaries for a class of structured signals and explore how one can pinpoint the precise benefits and perils of dependence on inference in Ising models. This talk is based on past and ongoing projects with Sohom Bhattacharya, Nabarun Deb, Sumit Mukherjee, Gourab Ray, and Ming Yuan.

Community Detection on Networks with Covariates
Wanjie Wang, National University of Singapore

Abstract: In social networks, besides the connection information, the nodes also have some node-specific covariates or attributes. These covariates also provide some information to detect the connection community of nodes, but community detection based on the covariates only may not be coincident with the network communities. In this talk, I will present a general model CA-DCSBM for social networks with covariates, which allows flexible relationships between the covariate classes and the communities. Based on CA-DCSBM, we developed the community detection method, CA-SCORE, which can recover the community labels exactly under some regular conditions. We present the generic theoretical results, and the conditions for two special cases that the adjacency matrix is informative or the covariate matrix is informative. The results are supported by numerical analysis.

Thursday, August 4, 2022 (cont.)

Session 5 (4:00-6:00pm)

Scalable Community Detection in Massive Networks using Aggregated Relational Data
Tian Zheng, Columbia University

Abstract: Fitting large Bayesian network models quickly become computationally infeasible when the number of nodes grows into the hundred of thousands and millions. In particular, the mixed membership stochastic blockmodel (MMSB) is a popular Bayesian network model used for community detection. In this paper, we introduce a scalable inference method that leverages nodal information that often accompanies real-world networks. Conditioning on this extra information leads to a model that admits a parallel variational inference algorithm. We apply our method to a citation network with over two million nodes and 25 million edges.

Manifold structure in graph embeddings
Patrick Rubin-Delanchy, University of Bristol

Abstract: Graph embedding refers to a collection of methods for representing the nodes of a network as a set of points in space, for purposes of visualisation, exploratory data analysis, prediction, synthetic network generation, and more. In this talk, I will give some arguments supporting a manifold hypothesis: the point representations of the nodes might reasonably be expected to be close to a lower-dimensional structure whose geometric properties (e.g. topology, geodesic distances) can be connected to true quantities of interest. Hence we can use the tools of topological data analysis and manifold estimation to recover information about the nodes that might have seemed lost forever.

Automated detection of edge clusters
Daniel Sewell, The University of Iowa

Abstract: An important facet of network analysis that receives significant attention is community detection. However, while most community detection algorithms focus on clustering the actors of the network, it is very intuitive to cluster the edges. Connections exist because they were formed within some latent environment such as, in the case of a social network, a workplace or religious group, and hence by clustering the edges of a network we may gain some insight into these latent environments. We propose a model-based approach to clustering the edges of a network using a latent space model describing the features of both actors and latent environments. Within a Bayesian framework, we use a sparse mixture prior that supports automated selection of the number of clusters. Estimation for our automated latent space edge clustering (aLSEC) model is obtained efficiently via a variational Bayes generalized expectation-maximization approach which has a computational cost that grows linearly with the number of actors in the network, making it scalable to large sparse networks. We demonstrate the potential impact of our proposed approach on a patient transfer network, verifying these results by running simple epidemic simulations, and on a real friendship network among faculty members at a university in the United Kingdom.

Bayesian community detection for networks with covariates
Arash Amini, The University of California, Los Angeles

Abstract: Among the various learning tasks with network data, community detection, the discovery of node clusters or “communities,” has arguably received the most attention in the scientific community. In many real-world applications, the network data often come with additional information in the form of node or edge covariates that should ideally be leveraged for inference. We add to a limited literature on community detection for networks with covariates by proposing a Bayesian stochastic block model with a covariate-dependent random partition prior. Under our prior, the covariates are explicitly expressed in specifying the prior distribution on the cluster membership. Our model has the flexibility of modeling uncertainties of all the parameter estimates including the community membership. Importantly, and unlike the majority of existing methods, our model has the ability to learn the number of the communities via posterior inference. Our model can be applied to community detection in both dense and sparse networks, with both categorical and continuous covariates, and our MCMC algorithm is very efficient with good mixing properties. We demonstrate the superior performance of our model over existing models in a comprehensive simulation study and an application to two real datasets.

Friday, August 5, 2022
Session 1 (9:00 - 10:30am)

Coevolving Latent Space Network with Attractors Models for Polarization
Eric Kolaczyk, McGill University

Abstract: We develop a broadly applicable class of coevolving latent space network with attractors (CLSNA) models, where nodes represent individual social actors assumed to lie in an unknown latent space, edges represent the presence of a specified interaction between actors, and attractors are added in the latent level to capture the notion of attractive and repulsive forces. We apply the CLSNA models to understand the dynamics of partisan polarization on social media, where we expect US Republicans and Democrats to increasingly interact with their own party and disengage with the opposing party. Using longitudinal social networks from the social media platforms Twitter and Reddit, we investigate the relative contributions of positive (attractive) and negative (repulsive) forces among political elites and the public, respectively. Our goals are to disentangle the positive and negative forces within and between parties and explore if and how they change over time. Our analysis confirms the existence of partisan polarization in social media interactions among both political elites and the public. Moreover, while positive partisanship is the driving force of interactions across the full periods of study for both the public and Democratic elites, negative partisanship has come to dominate Republican elites' interactions since the run-up to the 2016 presidential election. *This is joint work with Xiaojing Zhu, Cantay Caliskan, Dino P. Christenson, Kostas Spiliopoulos, and Dylan Walker.*

Spectral clustering via adaptive layer aggregation for multi-layer networks
Haolei Weng, Michigan State University

Abstract: One of the fundamental problems in network analysis is detecting community structure in multi-layer networks, of which each layer represents one type of edge information among the nodes. We propose integrative spectral clustering approaches based on effective convex layer aggregations. Our aggregation methods are strongly motivated by a delicate asymptotic analysis of the spectral embedding of weighted adjacency matrices and the downstream k-means clustering, in a challenging regime where community detection consistency is impossible. In fact, the methods are shown to estimate the optimal convex aggregation, which minimizes the mis-clustering error under some specialized multi-layer network models. Our analysis further suggests that clustering using Gaussian mixture models is generally superior to the commonly used k-means in spectral clustering. Extensive numerical studies demonstrate that our adaptive aggregation techniques, together with Gaussian mixture model clustering, make the new spectral clustering remarkably competitive compared to several popularly used methods.

Population-level Balance in Signed Networks
Weijing Tang, The University of Michigan

Abstract: Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for signed networks have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. In this talk, we introduce a statistically principled latent space approach for modeling signed networks and accommodating the well-known balance theory, i.e., ``the enemy of my enemy is my friend'' and ``the friend of my friend is my friend''. The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates, which are further verified through simulation studies. In addition, we apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II.

Friday, August 5, 2022 (cont.)

Session 2 (11:00am - 12:30pm)

Resampling methods for networks
Elizaveta Levina, University of Michigan

Abstract: With network data becoming ubiquitous in many applications, many models and algorithms for network analysis have been proposed, yet methods for providing uncertainty estimates are much less common. Bootstrap and other resampling procedures have been an effective general tool for estimating uncertainty from i.i.d. samples, but resampling network data is substantially more complicated, since we typically only observe one network. This talk will provide an overview of several recent resampling methods we have developed for networks

Variational Nonparametric Testing in Functional Stochastic Block Model
Zuofeng Shang, New Jersey Institute of Technology

Abstract: Abstract: We consider an extension of the classic stochastic block models whose vertices involve functional data information, and name the new model as functional stochastic block model (FSBM). The FSBM finds applications in many real-world networks in which the nodal information could appear as functional curves. Examples include world GDP data in which each network vertex (a country) is associated with the annual or quarterly GDP over certain time period. The statistical task is to test the significance of the nodal functional information in FSBM. We propose a penalized likelihood ratio test and show that it is asymptotically chi-square with diverging degrees of freedom, and propose a confidence band for the slop function based on asymptotic normality, both approaches involving variational MLE, and hence, are computationally tractable. A weakly consistent spectral algorithm is proposed for estimating the vertex labels. A simulation study is provided to support our methods.

Selecting the Number of Communities in Networks via Stepwise Matrix Scaling and Spectral Thresholding
Xiaodong Li, University of California, Davis

Abstract: This work aims to study how to select the number of communities in a network with unweighted edges or count-weighted edges. We consider the standard degree-corrected stochastic block model (DCSBM) for unweighted networks and Poisson DCSBM for count-weighted networks. For both cases, we proposed a stepwise procedure with the candidate number of communities $m=2,3,\ldots$. In each step, we first cluster the nodes into m groups with SCORE, and then fit the standard or Poisson DCSBM. Then we normalize the adjacency matrix, where the normalizing factors are determined by the estimated degree-correction parameters as well as scaling factors that are obtained from a step of Sinkorn's matrix scaling. The eigenvalues of the resultant normalized adjacency matrix are truncated with the threshold $2.01\sqrt{n}$ in magnitude. The stepwise procedure continues if the number of remaining eigenvalues is greater than m, and stops otherwise. Under mild conditions for the standard or Poisson DCSBM, we show that the proposed procedure can lead to consistent estimate of the true number of communities with high probability. Nonsplitting properties of under-fitting spectral clustering and spectral radii of inhomogeneous random graphs play essential roles in the analysis. Extensive numerical experiments on simulated and real data have also been conducted to illustrate our theoretical results.

Friday, August 5, 2022 (cont.)
Session 3 (2:00 - 3:30pm)

Learning Causal Networks from Biological Data
Ali Shojaie, University of Washington

Abstract: Learning high-dimensional directed networks is critical in many biological applications. An important application area is the study of biomolecular systems using various -omics measurements. While several existing algorithms can be used for this task, these general-purpose algorithms do not account for features of biological networks and do not take advantage of the properties of biological networks and data. In this talk we discuss alternative algorithms that address this limitation and can offer reliable estimates under potentially less stringent assumptions.

Simultaneous Cluster Structure Learning and Estimation of Heterogeneous Graphs for Matrix-variate fMRI Data
Lei Liu, Washington University at St. Louis

Abstract: Graphical models play an important role in neuroscience studies, particularly in brain connectivity analysis. Typically, observations/samples are from several heterogeneous groups and the group membership of each observation/sample is unavailable, which poses a great challenge for graph structure learning. In this article, we propose a method which can achieve Simultaneous Clustering and Estimation of Heterogeneous Graphs (briefly denoted as SCEHG) for matrix-variate function Magnetic Resonance Imaging (fMRI) data. Unlike the conventional clustering methods which rely on the mean differences of various groups, the proposed SCEHG method fully exploits the group differences of conditional dependence relationships among brain regions for learning cluster structure. In essence, by constructing individual-level between-region network measures, we formulate clustering as penalized regression with grouping and sparsity pursuit, which transforms the unsupervised learning into supervised learning. An ADMM algorithm is proposed to solve the corresponding optimization problem. We also propose a generalized criterion to specify the number of clusters. Extensive simulation studies illustrate the superiority of the SCEHG method over some state-of-the-art methods in terms of both clustering and graph recovery accuracy. We also apply the SCEHG procedure to analyze fMRI data associated with ADHD (abbreviated for Attention Deficit Hyperactivity Disorder), which illustrate its empirical usefulness.

High Dimensional Clustering via Latent Semiparametric Mixture Models
Wen Zhou, Colorado State University

Abstract: Cluster analysis is a fundamental task in machine learning. Several clustering algorithms have been extended to handle high-dimensional data by incorporating a sparsity constraint in the estimation of a mixture of Gaussian models. Though it makes some neat theoretical analysis possible, this type of approach is arguably restrictive for many applications. In this work we propose a novel latent variable transformation mixture model for clustering in which we assume that after some unknown monotone transformations the data follows a mixture of Gaussians. Under the assumption that the optimal clustering admits a sparsity structure, we develop a new clustering algorithm named CESME for high-dimensional clustering. The use of unspecified transformation makes the model far more flexible than the classical mixture of Gaussians. On the other hand, the transformation also brings quite a few technical challenges to the model estimation as well as the theoretical analysis of CESME. We offer a comprehensive analysis of CESME including identifiability, initialization, algorithmic convergence, and statistical guarantees on clustering. In addition, the convergence analysis has revealed an interesting algorithmic phase transition for CESME, which has also been noted for the EM algorithm in literature. Leveraging such a transition, a data-adaptive procedure is developed and substantially improves the computational efficiency of CESME. Extensive numerical study and real data analysis show that CESME outperforms the existing high-dimensional clustering algorithms including CHIME, sparse spectral clustering, sparse K-means, sparse convex clustering, and IF-PCA.

Friday, August 5, 2022 (cont.)
Session 4 (4:00 - 5:30pm)

Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering
Harrison Zhou, Yale University

Abstract: The singular subspaces perturbation theory is of fundamental importance in probability and statistics. It has various applications across different fields. We consider two arbitrary matrices where one is a leave-one-column-out submatrix of the other one and establish a novel perturbation upper bound for the distance between two corresponding singular subspaces. It is well-suited for mixture models and results in a sharper and finer statistical analysis than classical perturbation bounds such as Wedin's Theorem. Powered by this leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the performance of the spectral clustering under mixture models. Our analysis leads to an explicit exponential error rate for the clustering of sub-Gaussian mixture models. For the mixture of isotropic Gaussians, the rate is optimal under a weaker signal-to-noise condition than that of Löffler et al. (2021).

Causal Inference in Experiments on Networks
David Choi, Carnegie Mellon University

Abstract: In experiments that study social phenomena, such as peer influence or herd immunity,
the treatment of one unit may influence the outcomes of others. Such ""interference between units"" violates traditional approaches for causal inference, so that additional assumptions are often imposed to model or limit the underlying social mechanism; for example, one might assume that the units can be partitioned into non-interfering groups, or that an underlying dependency graph is unknown but sparse so that most units do not interfere with each other. For binary outcomes, we propose a randomization-based approach that does not require such assumptions, allowing for interference that is both unmodeled and arbitrarily strong. However, the estimates will have wider confidence intervals and weaker causal implications than those attainable under stronger assumptions, essentially showing only that effects exist and are correlated with specified measures of treatment exposure, such as the number of treated friends or neighborhood treatment rate. The approach allows for the usage of regression, matching, or weighting, as may best fit the application at hand. Novel computation methods are required for inference.

Orthogonal Common-Source and Distinctive-Source CoVariance Decomposition Between High-Dimensional Data Views
Hai Shu, New York University

Abstract: Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of two high-dimensional data views/sets is to decompose each data matrix into three parts: a low-rank common-source matrix that captures the shared information across data views, a low-rank distinctive-source matrix that characterizes the individual information within each single data view, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common-source and distinctive-source matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive-source matrices. The latter guarantees that no more shared information is extractable from the distinctive-source matrices. We propose a novel decomposition method that defines the common-source and distinctive-source matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with a careful construction of the orthogonal relationship between distinctive-source matrices. The proposed estimators of common-source and distinctive-source matrices are shown to be asymptotically consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and the real data analysis.

Speakers

Arash Amini
University of California, Los Angeles

David Choi
Carnegie Mellon University

Jianqing Fan
Princeton University

Tracy Ke
Harvard University

Eric Kolaczyk
Boston University

Liza Levina
University of Michigan

Xiaodong Li
University of California, Davis

Lei Liu
Washington University, St. Louis

Rajarshi Mukherjee
Harvard University

Carey Priebe
Johns Hopkins University

Patrick Rubin-Delanchy
University of Bristol

Purnamrita Sarkar
University of Texas, Austin

Daniel Sewell
University of Iowa

Zuofeng Shang
New Jersey Institute of Technology

Ali Shojaie
University of Washington

Hai Shu
New York University

Weijing Tang
University of Michigan

Stefan Wager
Stanford University

Wanjie Wang
National University, Singapore

Haolei Weng
Michigan State University

Yihong Wu
Yale University

Cui-Hui Zhang
Rutgers University

Emma Zhang
University of Miami

Yuan Zhang
Ohio State University

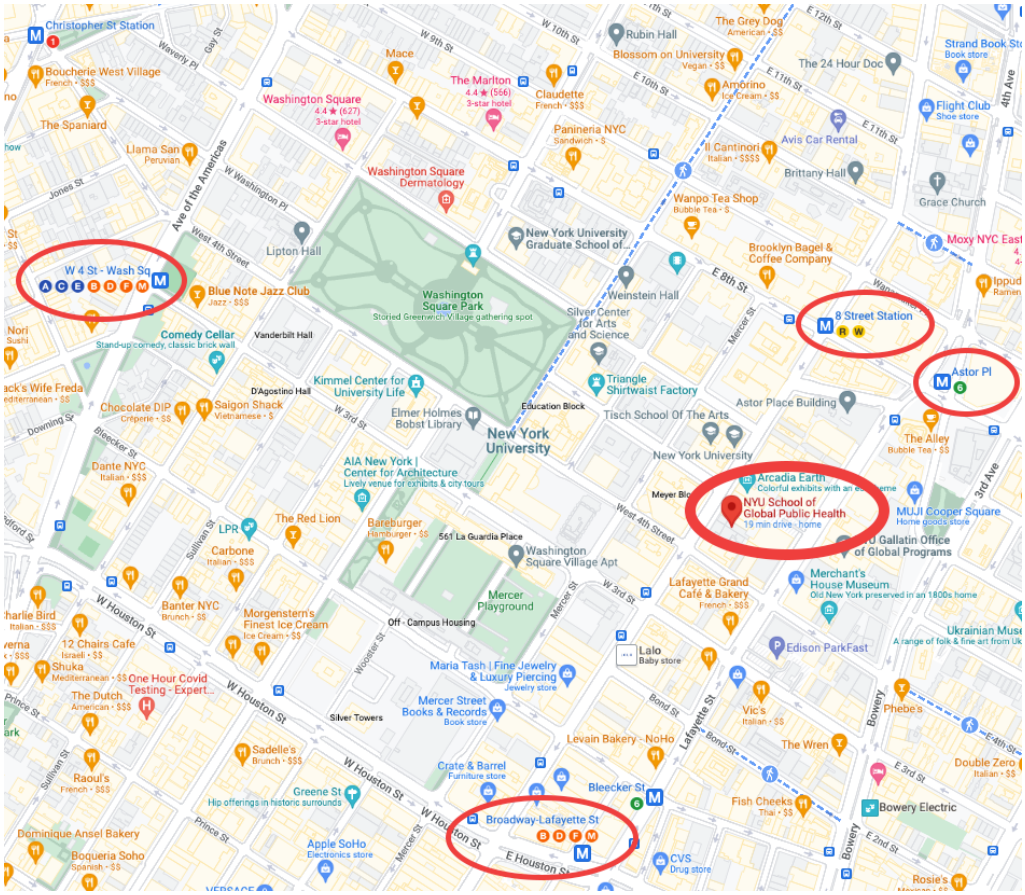
Tian Zheng
Columbia University

Harrison Zhou
Yale University

Wen Zhou
Colorado State University

Venue Map

708 Broadway, 3rd Floor / New York, NY, 10003



NOTE: The entrance is on Broadway. You can also scan the following QR code using your phone to get directions.

SCAN ME

2022 Workshop on Statistical Network Analysis & Beyond (SNAB2022)

WEDNESDAY, AUG 3 - FRIDAY, AUG 5 | **8AM ET**

NEW YORK UNIVERSITY
SCHOOL OF GLOBAL
PUBLIC HEALTH

708 BROADWAY, 3RD FLOOR