Frontiers of Statistical Network Analysis: Inference, Tensors and Beyond

Higher-order Networks and Tensors and Interface with other research areas
26–30 May 2025

National University of Singapore
Institute for Mathematical Sciences

ORGANIZERS

- Jialiang Li (National University of Singapore)
- Dong Xia (Hong Kong University of Science and Technology)
- Yuan Zhang (Ohio State University)

OVERVIEW

Network data present unique structures and distinct analytical challenges. Earlier efforts in statistical network analysis focused on modeling and parameter estimation. The past decade, known as a "golden age" for this field, has witnessed a surge in innovative methods and theoretical developments. The recent boom in data science also gives rise to numerous categories of complex networks where interactions among a set of entities are polyadic or non-linear, which are collectively referred to as tensors, or higher-order networks. Moreover, we witness the increasingly interdisciplinary collaborations between network analysis and applied fields. The goal of this workshop is three-fold: to advance the field by exchanging ideas and deepening understanding; to facilitate future research by discussing challenges and establishing collaboration opportunities; and to benefit the local Singaporean audience and other participants through lectures from leading experts and emerging star scholars.

Abstracts

Bhaswar B. Bhattacharya
University of Pennsylvania, USA

Higher-Order Graphon Theory: Fluctuations, Inference, and Degeneracies

Motifs (patterns of subgraphs), such as edges and triangles, encode important structural information about the geometry of a network. Consequently, counting motifs in a large network is an important statistical and computational problem. In this talk we will consider the problem of estimating motif densities and fluctuations of subgraph counts in an inhomogeneous random graph sampled from a graphon. We will show that the limiting distributions of subgraph counts can be Gaussian or non-Gaussian, depending on a notion of regularity of subgraphs with respect to the graphon. Using these results and a novel multiplier bootstrap for graphons, we will construct joint confidence sets for the motif densities. Finally, we will discuss various structure theorems and open questions about degeneracies of the limiting distribution and connections to quasirandom graphs.

Joint work with Anirban Chatterjee, Soham Dan, and Svante Janson

Back to Tables of Content

Yuxin Chen
University of Pennsylvania, USA

Heteroskedastic Tensor Clustering

Tensor clustering, which seeks to extract underlying cluster structures from noisy tensor observations, has gained increasing attention. One extensively studied model for tensor clustering is the tensor block model, which postulates the existence of clustering structures along each mode and has found broad applications in areas like multi-tissue gene expression analysis and multilayer network analysis. However, currently available computationally feasible methods for tensor clustering either are limited to handling i.i.d. sub-Gaussian noise or suffer from suboptimal statistical performance, which restrains their utility in applications that have to deal with heteroskedastic data and/or low signal-to-noise-ratio (SNR). To overcome these challenges, we propose a two-stage method, named High-order HeteroClustering (HHC), which starts by performing tensor subspace estimation via a novel spectral algorithm called Thresholded Deflated-HeteroPCA, followed by approximate k-means to obtain cluster nodes. Encouragingly, our algorithm provably achieves exact clustering as long as the SNR exceeds the computational limit (ignoring logarithmic factors); here, the SNR refers to the ratio of the pairwise disparity between nodes to the noise level, and the computational limit indicates the lowest SNR that enables exact clustering with polynomial runtime. Comprehensive simulation and real-data experiments suggest that our algorithm outperforms existing algorithms across various settings, delivering more reliable clustering performance.

Back to Table of Contents

David S. Choi
Carnegie Mellon University, USA

Agnostic Characterization of Interference in Randomized Experiments

In social network settings, the analysis of randomized experiments may be nontrivial due to the presence of interference between units, through diverse mechanisms such as peer influence, transmission of disease, market competition, and sharing of information. We give an approach for characterizing interference by lower bounding the number of units whose outcome depends on selected groups of treated individuals, such as depending on the treatment of others, or others who are at least a certain distance away. The approach is applicable to randomized experiments with binary-valued outcomes. Asymptotically conservative point estimates and one-sided confidence intervals may be constructed with no assumptions beyond the known randomization design, allowing the approach to be used when interference is poorly understood, or when an observed network might only be a crude proxy for the underlying social mechanisms. Point estimates are equal to Hajek-weighted comparisons of units with differing levels of treatment exposure. Empirically, we find that the width of our interval estimates is competitive with (and often smaller than) those of the EATE, an assumption-lean treatment effect, suggesting that the proposed estimands may be intrinsically easier to estimate than treatment effects.

Back to Table of Contents

Yingying Fan
University of Southern California, USA

HNCI: high-dimensional Network Causal Inference

The problem of evaluating the effectiveness of a treatment or policy commonly appears in causal inference applications under network interference. In this paper, we suggest the new method of high-dimensional network causal inference (HNCI) that provides both valid confidence interval on the average direct treatment effect on the treated (ADET) and valid confidence set for the neighbourhood size for interference effect. We exploit the model setting in Belloni et al. (2022) and allow certain type of heterogeneity in node interference neighbourhood sizes. We propose a linear regression formulation of potential outcomes, where the regression coefficients correspond to the underlying true interference function values of nodes and exhibit a latent homogeneous structure. Such a formulation allows us to leverage existing literature from linear regression and homogeneity pursuit to conduct valid statistical inferences with theoretical guarantees. The resulting confidence intervals for the ADET are formally justified through asymptotic normalities with estimable variances. We further provide the confidence set for the neighbourhood size with theoretical guarantees exploiting the repro samples approach. The practical utilities of the newly suggested methods are demonstrated through simulation and real data examples.

Back to Table of Contents

Yang Feng
New York University, USA

Semiparametric Modelling and Analysis for Longitudinal Network Data

We introduce a semiparametric latent space model for analysing longitudinal network data. The model consists of a static latent space component and a time-varying node-specific baseline component. We develop a semiparametric efficient score equation for the latent space parameter by adjusting for the baseline nuisance component. Estimation is accomplished through a one-step update estimator and an appropriately penalized maximum likelihood estimator. We derive oracle error bounds for the two estimators and address identifiability concerns from a quotient manifold perspective. Our approach is demonstrated using the New York Citi Bike Dataset.

Back to Table of Contents

Niels Richard Hansen
University of Copenhagen, Denmark

Identification and Inference from Cross-sectional Data via Higher Order Cumulants

Drawing inference from single cell data about the dynamics of the chemical reactions in the cell is the epitome of the problem I will address in the talk. The cell is killed during measurement, and we are thus only able to see a snapshot of the cell constituents at a single timepoint. This is an example of cross-sectional data from a multivariate dynamical system, which is obtained with the purpose of inferring properties of the dynamics of the process. This is only possible by making modelling assumptions and/or by obtaining data under various perturbations. Within a framework of linear (non-Gaussian) steady-state models, I will present a new characterization of all cumulant tensors as solutions of higher order Lyapunov equations. I will outline how this result can be used for practical estimation and inference with applications to single cell gene expression data. I will also present recent theoretical results on generic identifiability of model parameters related to the dynamics of the system from the third order cumulant tensor in the non-Gaussian setting.

Back to Table of Contents

Kengo Kato
Cornell University, USA

Limit Laws for Gromov-Wasserstein Alignment with Applications to Testing Graph Isomorphisms

---

The Gromov-Wasserstein (GW) distance enables comparing metric measure spaces based solely on their internal structure, making it invariant to isomorphic transformations. This property is particularly useful for comparing datasets that naturally admit isomorphic representations, such as unlabelled graphs or objects embedded in space. However, apart from the recently derived empirical convergence rates for the quadratic GW problem, a statistical theory for valid estimation and inference remains largely obscure. Pushing the frontier of statistical GW further, this work derives the first limit laws for the empirical GW distance across several settings of interest: (i)~discrete, (ii)~semi-discrete, and (iii)~general distributions under moment constraints under the entropically regularized GW distance. The derivations rely on a novel stability analysis of the GW functional in the marginal distributions. The limit laws then follow by an adaptation of the functional delta method. As asymptotic normality fails to hold in most cases, we establish the consistency of an efficient estimation procedure for the limiting law in the discrete case, bypassing the need for computationally intensive resampling methods. We apply these findings to testing whether collections of unlabelled graphs are generated from distributions that are isomorphic to each other.

Back to Table of Contents

Mladen Kolar
University of Southern California, USA

A Transfer Learning Approach to Precision Matrix Estimation

Many real-world systems—ranging from gene regulatory interactions in biology to financial asset dependencies—can be represented by networks, whose edges correspond to conditional relationships among variables. These relationships are succinctly captured by the precision matrix of a multivariate distribution. Estimating the precision matrix is thus fundamental to uncovering the underlying network structure. However, this task can be challenging when the available data for the target domain are limited, undermining accurate inference.

In this talk, I will present Trans-Glasso, a novel two-step transfer learning framework for precision matrix estimation that leverages data from source studies to improve estimates in the target study. First, Trans-Glasso identifies shared and unique features across studies via a multi-task learning objective. Then, it refines these initial estimates through differential network estimation to account for structural differences between the target and source precision matrices. Assuming that most entries of the target precision matrix are shared with at least one source matrix, we derive non-asymptotic error bounds and show that Trans-Glasso achieves minimax optimality under certain conditions.

Through extensive simulations, Trans-Glasso demonstrates improved performance over standard methods, especially in small-sample settings. Applications to gene regulatory networks across multiple brain tissues and protein networks in various cancer subtypes confirm its practical effectiveness in biological contexts, where understanding network structures can provide insights into disease mechanisms and potential interventions. Beyond biology, these techniques are broadly applicable wherever precision matrix estimation and network inference play a crucial role, including neuroscience, finance, and social science.

This is joint work with Boxin Zhao and Cong Ma.

Back to Table of Contents

Lexin Li
University of California, Berkeley, USA

Tensor Data Analysis and Some Applications in Neuroscience

Multidimensional arrays, or tensors, are becoming increasingly prevalent in a wide range of scientific applications. In this talk, I will present two case studies from neuroscience, where tensor decomposition proves particularly useful. The first study is a cross-area neuronal spike trains analysis, which we formulate as the problem of regressing a multivariate point process on another multivariate point process. We model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We then organize the corresponding transferring coefficients in the form of a three-way tensor, and impose the low-rank, sparsity, and subgroup structures on this coefficient tensor. The second study is a multimodal neuroimaging analysis for Alzheimer's disease, which we formulate as the problem of modelling the correlations of two sets of variables conditioning on the third set of variables. We propose a generalized liquid association analysis method to study such three-way associations. We establish a population dimension reduction model, and transform the problem to sparse decomposition of a three-way tensor.

Back to Table of Contents

Wei-Yin Loh
University of Wisconsin-Madison, USA

A Regression Tree Approach to Missing Data and Explainable AI

The problem of dealing with missing values in data is arguably the most difficult one in statistics. Although imputation is a popular solution, there are everyday situations where imputation makes no sense. Worse yet, unless the variables are well understood, it may be impossible to know whether imputation makes sense or not. In the first part of this talk, we introduce a regression tree algorithm called GUIDE that can fit classification and regression models to data without requiring imputation of missing values in the predictor variables. Unlike all other regression tree methods that perform imputation implicitly, GUIDE will identify the variables with informative missing values by highlighting them explicitly in the tree structures. In the second part of the talk, we show how GUIDE can be used to explain in simple terms the predictions of any machine learning model.

Back to Table of Contents

Aaron Potechin
University of Chicago, USA

Graph Matrices and Tensor Networks

Graph matrices are a type of matrix which plays a key role in proving sum of squares lower bounds on average case problems and is a powerful tool for analyzing problems on random inputs. In this talk, I will describe graph matrices and what is known about them, illustrate how they can be used, and describe how they are related to tensor networks.

Back to Table of Contents

Annie Qu
University of California, Irvine, USA

Representation Retrieval Learning for Heterogeneous Data Integration

In this presentation, I will showcase advanced statistical machine learning techniques and tools designed for the seamless integration of information from multi-source datasets. These datasets may originate from various sources, encompass distinct studies with different variables, and exhibit unique dependent structures. One of the greatest challenges in investigating research findings is the systematic heterogeneity across individuals, which could significantly undermine the power of existing machine learning methods to identify the underlying true signals. This talk will investigate the advantages and drawbacks of current data integration methods such as multi-task learning, optimal transport, missing data imputations, matrix completions and transfer learning. Additionally, we will introduce a new representation retriever learning aimed at mapping heterogeneous observed data to a latent space, facilitating the extraction of shared information and knowledge, and disentanglement of source-specific information and knowledge. The key idea is to project heterogeneous raw observations to representation retriever library, and the novelty of our method is that we can retrieve partial representations from the library for a target study. The main advantages of the proposed method are that it can increase statistical power through borrowing partially shared representation retrievers from multiple sources of data. This approach ultimately allows one to extract information from heterogeneous data sources and transfer generalizable knowledge beyond observed data and enhance the accuracy of prediction and statistical inference.

Back to Table of Contents

Garvesh Raskutti
University of Wisconsin-Madison, USA

Context-dependent and Model-agnostic Network Estimation

Social networks often present data in the form of events/posts by multiple people over time (e.g. tweets, posts, memes, etc..). One of the underlying challenges is determining who is influencing whom and how the influence is occurring. In this talk, I address this question by posing this problem as learning a context-dependent network structure by posing this problem as convex optimization problem with a tensor network parameter. The model has connections to multinomial auto-regressive models and compositional time series. Our approach is validated with meme-tracker and political tweet data. The main novelty is the scalability, theoretical guarantees and connection to model-agnostic variable importance networks. Building on this connection to model-agnostic variable importance networks, I briefly touch on recent work on scalable variable importance estimation and present some open questions on how this may be connects back to influence network estimation in the model-agnostic setting.

Back to Table of Contents

Will Wei Sun
Purdue University, USA

Online Statistical Inference for Low-Rank Reinforcement Learning

Reinforcement learning (RL) enables intelligent agents to make data-driven decisions in uncertain environments by leveraging contextual information to maximize cumulative rewards. Modern applications often involve high-dimensional tensor contexts, requiring low-rank structures for sample efficient RL models. While most RL algorithms focus on minimizing regret or selecting actions based on oracle policies, statistical inference for adaptively collected RL data remains underexplored. Such inference is crucial in domains like personalized medicine, mobile health, and autonomous driving, where understanding the statistical uncertainty of policy evaluations is essential. This talk presents online inferential tools designed for low-rank RL models, providing provable measures of uncertainty for safer and more reliable decision-making.

Back to Table of Contents

Anru Zhang
Duke University, USA

Theoretical Guarantees for Alternative Least Square Algorithm in Tensor CP Decomposition

We introduce a statistical and computational framework for tensor Canonical Polyadic (CP) decomposition, with a focus on statistical theory, convergence, and algorithmic improvements. First, we show that the Alternating Least Squares (ALS) algorithm achieves the desired error rate within three iterations when $R = 1$. Second, for the more general case where $R > 1$, we derive statistical bounds for ALS, showing that the estimation error exhibits an initial phase of quadratic convergence followed by linear convergence until reaching the desired accuracy. Third, we propose a novel warm-start procedure for ALS in the $R > 1$ setting, which integrates tensor Tucker decomposition with simultaneous diagonalization (Jennrich's algorithm) to significantly enhance performance over existing benchmark methods. Numerical experiments support our theoretical findings, demonstrating the practical advantages of our approach.

Back to Table of Contents

Harrison Zhou
Yale University, USA

From Score Estimation to Sampling

Recent impressive advances in the algorithmic generation of high-fidelity images, audio, and video can be largely attributed to the success of score-based diffusion models. A crucial step in their implementation is score matching, which involves estimating the score function of the forward diffusion process from training data. In this work, we establish the rate-optimal estimation of the score function for smooth, compactly supported densities and explore its applications to estimation of density, transport, and optimal transport.

Back to Table of Contents

Wen Zhou
New York University, USA

Nonparametric Inference on Network Effects with Dependent Edges: Optimality, Two-sample, Multiple Strata

Testing network effects in weighted directed networks is a foundational problem in econometrics, sociology, and psychology. Yet, the prevalent edge dependency poses a significant methodological challenge. Most existing methods are model-based and come with stringent assumptions, limiting their applicability. In response, we introduce a novel, fully nonparametric framework that requires only minimal regularity assumptions. While inspired by recent developments in U-statistic literature, our approach notably broadens their scopes. Specifically, we identified and carefully addressed the challenge of indeterminate degeneracy in the test statistics -- a problem that aforementioned tools do not handle. We established Berry-Esseen type bounds for the accuracy of type-I error rate control. With original analysis, we also proved the minimax power optimality of our test. Simulations underscore the superiority of our method in computation speed, accuracy, and numerical robustness compared to competing methods.

Back to Table of Contents