Frontiers of Statistical Network Analysis: Inference, Tensors and Beyond

Parametric and Nonparametric Network Inference

19–23 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

Jinyuan Chang
Southwestern University of Finance and Economics, China

Autoregressive Networks with Dependent Edges

We propose an autoregressive framework for modelling dynamic networks with dependent edges. It encompasses models that accommodate, for example, transitivity, density-dependence and other stylized features often observed in real network data. By assuming the edges of networks at each time are independent conditionally on their lagged values, the models, which exhibit a close connection with temporal ERGMs, facilitate both simulation and the maximum likelihood estimation in a straightforward manner. Due to the possibly large number of parameters in the models, the natural MLEs may suffer from slow convergence rates. An improved estimator for each component parameter is proposed based on an iteration employing projection, which mitigates the impact of the other parameters (Chang et al., 2021, 2023). Leveraging a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the assumption of stationarity. The limiting distribution is not normal in general, although it reduces to normal when the underlying process satisfies some mixing conditions. Illustration with a transitivity model was carried out in both simulation and a real network data set.

Back to Table of Contents

Subhroshekhar Ghosh
National University of Singapore, Singapore

Learning with Latent Group Sparsity via Diffusions on Networks

Group or cluster structure on explanatory variables in machine learning problems is a very general phenomenon, which has attracted broad interest from practitioners and theoreticians alike. In this work we contribute an approach to learning under such group structure, that does not require prior information on the group identities. Our paradigm is motivated by the Laplacian geometry of an underlying network with a related community structure, and proceeds by directly incorporating this into a penalty that is effectively computed via a heat flow-based local network dynamics. In fact, we demonstrate a procedure to construct such a network based on the available data. Notably, we dispense with computationally intensive pre-processing involving clustering of variables, spectral or otherwise. Our technique is underpinned by rigorous theorems that guarantee its effective performance and provide bounds on its sample complexity. In particular, in a wide range of settings, it provably suffices to run the diffusion for time that is only logarithmic in the problem dimensions. We explore in detail the interfaces of our approach with key statistical physics models in network science, such as the Gaussian Free Field and the Stochastic Block Model. We validate our approach by successful applications to real-world data from a wide array of application domains, including computer science, genetics, climatology and economics. Our work raises the possibility of applying similar diffusion-based techniques to classical learning tasks, exploiting the interplay between geometric, dynamical and stochastic structures underlying the data.

Back to Table of Contents

Christophe Giraud
Université Paris-Saclay, France

Learning Latent Features from Network Data

Network structures are often shaped by some historical or latent features of the nodes. For example, communities drive the network properties in Stochastic Block Model, latent features determine the connection patterns in Graphons, and arrival time influences the local properties in random recursive trees. In this talk, we will discuss how to learn these features in various learning frameworks, where probabilities of connection fulfil some shape constraints, should it be tree constraints, or graphon's shape constraints.

Back to Table of Contents

Chenlei Leng
The University of Warwick, UK

Statistical Analysis of Reciprocity

Asymmetric relational data is increasingly prevalent across diverse fields, highlighting the need for directed network models to address the unique challenges posed by their complex structures. Unlike undirected models, directed models can explicitly capture reciprocity—the tendency of nodes to form mutual links. In this talk, we will progressively develop a series of models to examine reciprocity in directed networks. We begin with a simple extension of the Erdős–Rényi model to incorporate reciprocity, addressing the fundamental question of effective sample sizes for reciprocal and non-reciprocal effects. Next, we introduce covariates to refine the modelling of reciprocity. Finally, we present a model that accommodates both node-specific heterogeneity and reciprocity, achieving a fully general framework for modelling reciprocal relationships in directed networks.

Back to Table of Contents

Jinchi Lv
University of Southern California, USA

ATE-GL: Asymptotic Theory of Eigenvectors for Latent Embeddings with Generalized Laplacian Matrices

Laplacian matrices are commonly employed in many real applications, encoding the underlying latent structural information such as graphs and manifolds. The use of the normalization terms naturally gives rise to random matrices with dependency. It is well-known that dependency is a major bottleneck of new random matrix theory (RMT) developments. To this end, in this paper we formally introduce a class of generalized (regularized) Laplacian matrices, which contains the Laplacian matrix and the random adjacency matrix as a specific case, and suggest the new framework of asymptotic theory of eigenvectors for latent embeddings with generalized Laplacian matrices (ATE-GL). Our new theory is empowered by the tool of generalized quadratic vector equation for dealing with RMT under dependency, and delicate high-order asymptotic expansions of the empirical spiked eigenvectors and eigenvalues based on local laws. The asymptotic normalities established for both spiked eigenvectors and eigenvalues will enable us precise inference and uncertainty quantification for applications involving the generalized Laplacian matrices with flexibility. We discuss some applications of the suggested ATE-GL framework and showcase its validity through some numerical examples. This is a joint work with Jianqing Fan, Yingying Fan, Fan Yang and Xin Diwen Yu.

Back to Table of Contents

Michael Schweinberger
The Pennsylvania State University, USA

A Regression Framework for Studying Relationships among Attributes under Network Interference

To understand how the interconnected and interdependent world of the twenty-first century operates and make model-based predictions, joint probability models for networks and interdependent outcomes are needed. We propose a comprehensive regression framework for networks and interdependent outcomes with multiple advantages, including interpretability, scalability, and provable theoretical guarantees. The regression framework can be used for studying relationships among attributes of connected units and captures complex dependencies among connections and attributes, while retaining the virtues of linear regression, logistic regression, and other regression models by being interpretable and widely applicable. On the computational side, we show that the regression framework is amenable to scalable statistical computing based on convex optimization of pseudo-likelihoods using minorization-maximization methods. On the theoretical side, we establish convergence rates for pseudo-likelihood estimators based on a single observation of dependent connections and attributes. We demonstrate the regression framework using simulations and an application to hate speech on the social media platform X in the six months preceding the insurrection at the U.S. Capitol on January 6, 2021.

Back to Table of Contents

Dapeng Shi
Chinese University of Hong Kong, Hong Kong SAR

A Multilayer Probit Network Model for Community Detection with Dependent Layers

Community detection in multilayer networks, which aims to identify groups of nodes exhibiting similar connectivity patterns across multiple network layers, has attracted considerable attention in recent years. Most existing methods are developed by assuming that the network layers are either independent or exhibit some specific dependence structures. In this article, we propose a novel method for community detection in multilayer networks that accommodates a wide range of inter-layer dependence structures. The proposed method integrates the multilayer stochastic block model for community detection with a multivariate probit model to capture dependence structures between network layers. To facilitate the parameter estimation, we develop a constrained pairwise likelihood method coupled with an efficient alternating updating algorithm. The asymptotic properties of the proposed method are also established, with a focus on examining the influence of inter-layer dependence structures and strength on parameter estimation and community detection. The theoretical results are supported by extensive numerical experiments on both simulated networks and a real-world multilayer trade network.

Back to Table of Contents

Miaoyan Wang
University of Wisconsin-Madison, USA

Application and Methods for Structured Tensor Learning

High-order tensor datasets pose common challenges in applications such as recommendation systems, neuroimaging, and social networks. In this work, we introduce two approaches for learning with structured tensors: block models for higher-order clustering and sign-series models for tensor denoising. These approaches provide lens into the unique properties of tensor analysis. We establish statistical and computational efficiency results for each method. Additionally, we present polynomial-time algorithms with guaranteed efficiency. The effectiveness of our methods is demonstrated through applications to neuroimaging data analysis and social network analysis.

Back to Table of Contents

Dong Xia
Hong Kong University of Science and Technology, Hong Kong SAR

Online Decision Making:
Algorithm, Regret, Constraints and Uncertainty

We will discuss online decision-making problems in scenarios where covariates are high-dimensional or personalized covariates are unavailable. Our focus is on the ϵ-greedy algorithm for decision making and online gradient descent for estimating model parameters. By carefully balancing exploration and exploitation, we achieve a trade-off between regret performance and estimation accuracy. Additionally, we explore online decision-making under constraints (such as knapsack problems) within a primal-dual framework, demonstrating that sublinear regret is achievable. Finally, we propose an online debiasing approach based on inverse propensity weighting (IPW) for uncertainty quantification. Real data examples will also be discussed.

Back to Table of Contents

Qiwei Yao
London School of Economics and Political Science, UK

Cointegration Between Two Intrinsically Stationary Spatial Processes

The concept of the intrinsic processes proposed by Matheron (1973) provides an elegant mathematical framework for modeling nonstationary spatial phenomena. It can be viewed as a direct analogue of taking differences of nonstationary time series to achieving stationarity. But it is applicable to spatial data observed on irregular grids. The goal of this paper is to establish the inference methods and the relevant theory for identifying the cointegration between two simple intrinsic processes. We apply the least squares estimation, like Engle and Granger (1987). However, the asymptotic property of the estimation is much more complex, depending on the underlying processes as well as the way the observations were taken. We propose some bootstrap approximations for the asymptotic distribution of the estimators. It turns out that the wild bootstrap procedure is adaptive automatically to varying convergence rates under the different schemes of taking the observations. Therefore, it paves the way for constructing practically feasible confidence intervals for cointegration coefficients. A new and easy-to-use statistical tests is constructed for testing the cointegration. The proposed methods, as well as the associated asymptotic results under various settings, are illustrated in simulation.

Back to Table of Contents

Emma Jingfei Zhang
Emory University, USA

Modelling Non-Uniform Hypergraphs Using Determinantal Point Processes

Most statistical models for networks focus on pairwise interactions between nodes. However, many real-world networks feature higher-order interactions involving multiple nodes, such as co-authors collaborating on a paper. Hypergraphs provide a natural representation for these networks, with each hyperedge representing a set of nodes. The majority of existing hypergraph models assume uniform hyperedges, that is, edges are of the same size, or are driven by diversity amongst nodes. In this work, we propose a new hypergraph model formulated based on non-symmetric determinantal point processes. The proposed model naturally accommodates non-uniform hyperedges, has tractable probability mass functions, and allows for node similarity or diversity in hyperedges. For model estimation, we maximize the likelihood function under constraints via a computationally efficient projected adaptive gradient descent algorithm and establish the consistency and asymptotic normality of the estimator. Simulation studies confirm the efficacy of the proposed model, and its utility is further demonstrated through predictions on several real-world datasets.

Back to Table of Contents

Jingnan Zhang
University of Science and Technology of China, China

A Dynamic Network Autoregressive Model for Time-varying Network-link Data

Network-linked data, where different units are linked through a network has been extensively studied in literature. However, its extension, specifically time-varying network-link data, has received less investigation. Existing methods for time-varying network-link data only assume that units' attributes change over time, neglecting network evolution. To address this gap, we propose a dynamic network autoregressive model for time-varying network-link data, where both units' attributes and networks are allowed to vary over time. A tensor decomposition method is employed to provide low-dimensional embedding vectors, which are further used to reformulate the traditional network autoregressive model. Interestingly, node-embedding vectors are concentrated around some group centers but are not exactly the same within some groups. Meanwhile ,both within-group and global homogeneities are considered for the effect of covariate vectors. To tackle the resultant optimization task, we develop the power update algorithm and an efficient alternative updating algorithm. Furthermore, the asymptotic consistencies of the proposed method are established, irrespective of the presence of the global effect of covariate vector. These consistencies are demonstrated by extensive simulated examples and a real example of time-varying network-linked fund data.

Back to Table of Contents

Yuan Zhang
Ohio State University, USA

Higher-order Accurate Two-sample Network Inference and Network Hashing

Two-sample hypothesis testing for network comparison presents many significant challenges, including: leveraging repeated network observations and known node registration, but without requiring them to operate; relaxing strong structural assumptions; achieving finite-sample higher-order accuracy; handling different network sizes and sparsity levels; fast computation and memory parsimony; controlling false discovery rate (FDR) in multiple testing; and theoretical understandings, particularly regarding finite-sample accuracy and minimax optimality. In this paper, we develop a comprehensive toolbox, featuring a novel main method and its variants, all accompanied by strong theoretical guarantees, to address these challenges. Our method outperforms existing tools in speed and accuracy, and it is proved power-optimal. Our algorithms are user-friendly and versatile in handling various data structures (single or repeated network observations; known or unknown node registration). We also develop an innovative framework for offline hashing and fast querying as a very useful tool for large network databases. We showcase the effectiveness of our method through comprehensive simulations and applications to two real-world datasets, which revealed intriguing new structures.

Back to Table of Contents

Ji Zhu
University of Michigan, USA

Hyperbolic Network Latent Space Model with Learnable Curvature

Network data is ubiquitous in various scientific disciplines, including sociology, economics, and neuroscience. Latent space models are often employed in network data analysis, but the geometric effect of latent space curvature remains a significant, unresolved issue. In this work, we propose a hyperbolic network latent space model with a learnable curvature parameter. We theoretically justify that learning the optimal curvature is essential to minimizing the embedding error across all hyperbolic embedding methods beyond network latent space models. A maximum-likelihood estimation strategy, employing manifold gradient optimization, is developed, and we establish the consistency and convergence rates for the maximum-likelihood estimators, both of which are technically challenging due to the non-linearity and non-convexity of the hyperbolic distance metric. We further demonstrate the geometric effect of latent space curvature and the superior performance of the proposed model through extensive simulation studies and an application using a Facebook friendship network.

Back to Table of Contents