Titles and Abstracts

Graphon Cross-Validation with Application to Drug Repurposing

Huimin Cheng(成慧敏)
Boston University

Graphon, short for graph function, provides a generative model for networks. In recent decades, various methods for graphon estimation have been proposed. The success of most graphon estimation methods depends on the proper specification of hyperparameters. While some network cross-validation methods have been proposed, they suffer from restrictive model assumptions, expensive computational costs, and a lack of theoretical guarantees. To address these issues, we propose a graphon cross-validation (GraphonCV) method. The asymptotic properties of GraphonCV are established. The effectiveness of the proposed method in terms of both computation and accuracy is demonstrated through extensive simulation studies and real drug repurposing examples.

Imputation-based randomization tests for randomized experiments with interference via social network

Ke Deng(邓柯)
Tsinghua University

The presence of interference renders classic Fisher randomization tests infeasible due to nuisance unknowns. To address this issue, we propose imputing the nuisance unknowns and computing Fisher randomization p-values multiple times, then averaging them. We term this approach the imputation-based randomization test and provide theoretical results on its asymptotic validity. Our method leverages the merits of randomization and the flexibility of the Bayesian framework: for multiple imputations, we can either employ the empirical distribution of observed outcomes to achieve robustness against model mis-specification or utilize a parametric model to incorporate prior information. Simulation results demonstrate that our method effectively controls the type Ierror rate and significantly enhances the testing power compared to existing randomization tests for randomized experiments with interference. We apply our method to a two-round randomized experiment with multiple treatments and one-way interference, where existing randomization tests exhibit limited power.

Pseudo-Likelihood Ratio Screening based on Network Data with Applications

Danyang Huang (黄丹阳)
Renmin University of China

Social network data contain the network structure and personalized labels for each user. The dimension of personalized labels could be ultra-high, which would cause trouble for model analysis and response prediction. In this scenario, traditional categorical feature screening methods ignore the involvement of network structure, which may lead to an incorrect selected feature set and sub-optimal prediction accuracy. This study focuses on feature screening for network-involved label prediction problems. We first propose the definitions of individual-related features and network-related features, which are defined as those directly related to the response and those related to the network structure, respectively. Both contribute to prediction accuracy. In this way, we propose a pseudo-likelihood ratio feature screening procedure, which could obtain both types of features. Theoretical properties of the proposed procedure under different scenarios are carefully investigated, and its strong screening consistency property is established. Simulation and real data analysis on Sina Weibo corroborate our theoretical findings.

Modelling Homophily in Dynamic Networks

Binyan Jiang (蔣濱雁)
The Hong Kong Polytechnic University

Statistical modeling of network data is an important topic in various areas. Although many real networks are dynamic in nature, most existing statistical models and related inferences for network data are confined to static networks, and the development of the foundation for dynamic network models is still in its infancy. In particular, to the best of our knowledge, no attempts have been made to jointly address node heterogeneity and link homophily among dynamic networks. Being able to capture these network features simultaneously will only bring new insights on understanding how networks were formed, but also provide more sophisticated tools for the prediction of a future network with statistical guarantees. In this project, we take into account link homophily associated with both observed traits and latent traits of the nodes, and propose a novel convex loss based framework to generate stable estimations for the high dimensional parameters. We show that, with an appropriate initialization, the resulting estimator is consistent. The impressive performance of our proposed model is demonstrated through its application in community detection as well as various simulation studies.

Network modeling and Goodness-of-Fit

Jiashun Jin(金加顺)
Southeast University

The block-model family has four popular network models: SBM, MMSBM, DCBM, and DCMM. A fundamental problem is, how well each of these models fits with real networks. We propose GoF-MSCORE as a new Goodness-of-Fit (GoF) metric for DCMM (the broadest one among the four), with two main ideas. The first is to use cycle count statistics as a general recipe for GoF. The second is a novel network fitting scheme. GoF-MSCORE is a flexible GoF approach. We adapt it to all four models in the block-model family. We show that for each of the four models, if the assumed model is correct, then the corresponding GoF metric converges to N(0,1) as the network sizes diverge. We also analyze the powers and show that these metrics are optimal in many settings. For 12 real networks, we use the proposed GoF metrics to show that DCMM fits well with almost all of them. We also show that SBM, DCBM, and MMSBM do not fit well with many of these networks, especially when the networks are relatively large. Together with the mathematical tractability of the block-model family, these suggest that DCMM is a possible (or is close to the) sweet-spot for network modeling.

An eigenvector theory for the normalized graph Laplacian and its application in network mixed membership estimation

Zheng Ke(柯峥)
Harvard University

TBA

Estimation of Grouped Time-Varying Network Vector Autoregression Models

Degui Li(李德櫃)
University of Macau

This paper introduces a flexible time-varying network vector autoregressive model framework for large-scale time series. A latent group structure is imposed on the heterogeneous and node-specific time-varying momentum and network spillover effects so that the number of unknown time-varying coefficients to be estimated can be reduced considerably. A classic agglomerative clustering algorithm with nonparametrically estimated distance matrix is combined with a ratio criterion to consistently estimate the latent group number and membership. A post-grouping local linear smoothing method is proposed to estimate the group-specific time-varying momentum and network effects, substantially improving the convergence rates of the preliminary estimates which ignore the latent structure. We further modify the methodology and theory to allow for structural breaks in either the group membership, group number or group-specific coefficient functions. Numerical studies including Monte-Carlo simulation and an empirical application are presented to examine the finite-sample performance of the developed model and methodology.

Quasi-Score Matching Estimation for Spatial Autoregressive Model with Random Weights Matrix and Regressors

Xuan Liang(梁萱)
Australian National University

With the rapid advancements in technology for data collection, the application of the spatial autoregressive (SAR) model has become increasingly prevalent in real-world analysis, particularly when dealing with large datasets. However, the commonly used quasi-maximum likelihood estimation (QMLE) for the SAR model is not computationally scalable to handle the data with a large size. In addition, when establishing the asymptotic properties of the parameter estimators of the SAR model, both weights matrix and regressors are assumed to be nonstochastic in classical spatial econometrics, which is perhaps not realistic in real applications. Motivated by the machine learning literature, this paper proposes quasi-score matching estimation for the SAR model. This new estimation approach is still likelihood-based, but significantly reduces the computational complexity of the QMLE. The asymptotic properties of parameter estimators under the random weights matrix and regressors are established, which provides a new theoretical framework for the asymptotic inference of the SAR-type models. The usefulness of the quasi-score matching estimation and its asymptotic inference is illustrated via extensive simulation studies and a case study of an anti-conflict social network experiment for middle school students.

Estimation and inference of average treatment effects under heterogeneous additive treatment effect model

Hanzhong Liu(刘汉中)
Tsinghua University

Randomized experiments are the gold standard for estimating treatment effects, yet network interference challenges the validity of traditional estimators by violating the stable unit treatment value assumption and introducing bias. While cluster randomized experiments mitigate this bias, they encounter limitations in handling network complexity and fail to distinguish between direct and indirect effects. To address these challenges, we develop a design-based asymptotic theory for the existing Horvitz-Thompson estimators of the direct, indirect, and global average treatment effects under Bernoulli trials. We assume the heterogeneous additive treatment effect model with a hidden network that drives interference. Observing that these estimators are inconsistent in dense networks, we introduce novel eigenvector-based regression adjustment estimators to ensure consistency. We establish the asymptotic normality of the proposed estimators and provide conservative variance estimators under the design-based inference framework, offering robust conclusions independent of the underlying stochastic processes of the network and model parameters. Our method's adaptability is demonstrated across various interference structures, including partial interference and local interference in a two-sided marketplace. Numerical studies further illustrate the efficacy of the proposed estimators, offering practical insights into handling network interference.

Supervised Centrality via Sparse Network Influence Regression: An Application to the 2021 Henan Floods' Social Network

Yingying Ma(马莹莹)
Beihang University

The social characteristics of players in a social network are closely associated with their network positions and relational importance. Identifying those influential players in a network is of great importance as it helps to understand how ties are formed, how information is propagated, and, in turn, can guide the dissemination of new information. Motivated by a Sina Weibo social network analysis of the 2021 Henan Floods, where response variables for each node are available, we propose a new notion of supervised centrality that emphasizes the task-specific nature of a player's centrality. To estimate the supervised centrality and identify important players, we develop a novel sparse network influence regression by introducing individual heterogeneity for each user. To overcome the computational difficulties in fitting the model for large social networks, we further develop a forward-addition algorithm and show that it can consistently identify a superset of the influential nodes. We apply our method to analyze three responses in the Henan Floods data: the number of comments, reposts, and likes, and obtain meaningful

results. A further simulation study corroborates the developed method.

Sequential data integration under dataset shift

Ying Sheng(盛赢)
University of Chinese Academy of Sciences, UCAS

With the rapidly increasing availability of large-scale and high-velocity streaming data, efficient algorithms that can process data in batches without requiring expensive storage and computation resources have drawn considerable attention. An emerging challenge in developing efficient batch processing techniques is dataset shift, where the joint distribution of the collected data varies across batches. If not recognized and addressed properly, dataset shift often leads to erroneous statistical inferences when integrating data from different batches. In this paper, two shift-adjusted estimation procedures are developed for updated estimation of the parameter in the presence of dataset shift. Under prior probability shift, we can obtain parameter estimation and assess the degree of dataset shift simultaneously. We study the asymptotic properties of the proposed estimators and evaluate their performance in numerical studies. The proposed methodologies are illustrated with an analysis of the Ford GoBike docked bike-sharing data. This is a joint work with Professor Jing Qin (NIH) and Professor Chiung-Yu Huang (UCSF).

Inference and Learning for Signed Networks Guided by Social Theory

Weijing Tang(唐沩婧)
Carnegie Mellon University, CMU

In many real-world networks, relationships often go beyond simple presence or absence; they can be positive (e.g., friendship, alliance, and mutualism) or negative (e.g., enmity, disputes, and competition). These negative relationships display substantially different properties from positive ones, and more importantly, their presence interacts in unique ways. The balance theory originating from social psychology, illustrated by proverbs like “a friend of my friend is my friend” and “an enemy of my enemy is my friend”, provides insight into the formation mechanism of positive and negative connections. In this talk, we characterize the balance theory with a novel and natural notion of population-level balance. We propose a nonparametric inference method to evaluate the real-world evidence of population-level balance in signed networks. Inspired by the empirical findings, we further develop a general latent space framework that incorporates the balance theory for modeling signed networks.

Uniform accurate approximation on large graphical models

Xin Tong(童心)
National University of Singapore

Distributions on large graphical model often exhibit sparse or incoherent features. Intuitively, such incoherency implies low intrinsic dimensionality, which can be exploited for efficient approximation and computation of complex distributions. Existing approximation theory mainly considers the joint distributions, which does not guarantee that the marginal error is small. In this work, we establish a dimension independent error bound for the marginals of approximate distributions. Such l-infty approximation error is achieved by Stein's method, and we propose a condition that quantifies how a distribution is incoherent. We also show how to obtain uniform error bound given different sparsity conditions that characterize incoherence. The l_inf approximation bound motivates to sparsify existing approximation methods to respect the locality. As examples, we show how to use score matching to avoid the dimension dependence in the approximation error. We also show how to use parallel Gibbs sampling to efficiently accelerate the sampling.

Tracking structural changes in dynamic heterogeneous networks

Junhui Wang(王軍輝)
The Chinese University of Hong Kong, CUHK

Dynamic networks consist of a sequence of time-varying heterogeneous networks, and it is of great importance to detect the structural changes. Most existing methods focus on detecting abrupt network changes, necessitating the assumption that the underlying network probability matrix remains constant between adjacent change points. This assumption can be overly strict in many real-life scenarios due to their versatile network dynamics. In this talk, we introduce a new subspace tracking method to detect network structural changes in dynamic networks, whose network connection probabilities may still undergo continuous changes. Particularly, two new detection statistics are proposed to jointly detect the network structural changes, followed by a carefully refined detection procedure. Theoretically, we show that the proposed method is asymptotically consistent in terms of detecting the network structural changes, and also establish the impossibility region in a minimax fashion. The advantage of the proposed method is supported by extensive numerical experiments on both synthetic networks and a series of UK politician social networks.

Empirical Likelihood Inference over Decentralized Networks

Qihua Wang(王启华)
University of Chinese Academy of Sciences, UCAS

As a nonparametric statistical inference approach, empirical likelihood has been found very useful in numerous occasions.

However, it encounters serious computational challenges when applied directly to the modern massive dataset. This article studies empirical likelihood inference over decentralized distributed networks, where the data are locally collected and stored by different nodes. To fully utilize the data, this article fuses Lagrange multipliers calculated in different nodes by employing a penalization technique. The proposed distributed empirical log-likelihood ratio statistic with Lagrange multipliers solved by the penalized function is asymptotically standard chi-squared under regular conditions even for a divergent machine number. Nevertheless, the optimization problem with the fused penalty is still hard to solve in the decentralized distributed network. To address the problem, two alternating direction method of multipliers (ADMM) based algorithms are proposed, which both have simple node-based implementation schemes. Theoretically, this article establishes convergence properties for proposed algorithms, and further proves the linear convergence of the second algorithm in some specific network structures. The proposed methods are evaluated by numerical simulations and illustrated with analyses of census income and Ford gobike datasets.

Spectral-Based Community Detection for Multilayer Networks with Mixed Sparsity and Layer-Specific Memberships

Wanjie Wang(王婉洁)
National University of Singapore, NUS

Complex networks are increasingly common. In this work, we focus on multilayer networks with extreme sparsity in some communities. Unlike most multilayer works, we allow the community memberships to vary across layers, with only a proportion p of the structure shared between layers. To leverage both the information specific to the layer of interest and the common information across layers, we extend community detection methods for networks with covariates. For each layer l, we treat its adjacency matrix A^(l) as the primary network and construct a "covariate" matrix X^(l) using spectral information from other layers. By applying the Network-Adjusted Covariate (NAC) clustering algorithm to A^(l) and X^(l), we achieve community detection in a layer-specific context. A critical challenge lies in constructing an effective “covariate" matrix. For Degree-Corrected SBM, leading eigenvectors of the adjacency matrices preserve essential information, and stacking these eigenvectors from all layers naturally forms a covariate matrix. However, this naive approach struggles with extreme sparsity, leading to significant detection errors. To address this, we propose a degree-corrected version, and incorporate eigenvalues as weights. This correction significantly enhances accuracy and guarantees strong consistency in community detection across all layers. The effectiveness is also shown in numerical analysis.

Network Regression with low rank and sparse structure

Weining Wang(王玮宁)
University of Groningen

We propose to study the interaction effects of social and spatial networks in the presence of a noisy adjacency matrix. First, we provide evidence that existing network datasets exhibit low-rank, sparse, and noisy structures, and we utilize this information to create a de-noised version of the network. We employ the Least Absolute Shrinkage and Selection Operator (LASSO) in conjunction with nuclear norm penalization to simultaneously regularize the sparse and low-rank components. We introduce two procedures: a two-step estimator, where we first de-noise the adjacency matrix before using it in regression analysis, and a one-step supervised Generalized Method of Moments (GMM) estimator using proximal gradient methods for efficient computation. Our results show that our estimation method performs favorably compared to GMM, especially when dense errors are present and networks are endogenous to measurement errors. Simulation exercises indicate that our method outperforms GMM by up to 30\% in root mean squared error (RMSE) terms when noise is present in the network, and maintains a significant advantage of approximately 40\% on average with endogenous networks. Additionally, we apply our method to two international trade datasets and find our methods deliver estimated spillover and multiplier effects that differ significantly from those obtained using standard GMM. Furthermore, we show how our decomposition can be used to provide reliable and more detailed guidance for policy targeting under constraints.

Network Gradient Descent and Expectation-Maximization Algorithms in Decentralized Federated Learning

Shuyuan Wu(伍书缘)
Shanghai University of Finance and Economics

We study two fully decentralized federated learning algorithms: Network Gradient Descent (NGD) and Network Expectation-Maximization (NEM) algorithms. In these methods, clients communicate only parameter estimators, minimizing privacy risks and enhancing reliability through a specially designed network structure. Our theoretical analysis demonstrates that the learning rate and network structure significantly affect the statistical efficiency of the resulting estimators. With a sufficiently small learning rate and a well-balanced network, the estimators achieve statistical efficiency comparable to that of the global estimator, even with heterogeneous data distributions. Extensive simulations and real data analyses are conducted to validate our theoretical findings.

Mixture Multi-layer Stochastic Block Model: Community Detection, Network Clustering, and Minimax Optimality

Dong Xia(夏冬)
The Hong Kong University of Science and Technology, HKUST

We introduce a mixture multi-layer stochastic block model that accommodates heterogeneous community assignments across layers. We propose computationally efficient algorithm based on tensor decomposition and spectral clustering for community detection and layer clustering. A two-stage clustering algorithm is also designed showing that minimax optimal rate of layer clustering can be achieved.

Likelihood ratio tests in random graph models

Ting Yan (晏挺)
Central China Normal University

In this talk, we present likelihood ratio tests in some random graph models including the beta model for undirected graphs, the Bradley-Terry model for paired companions and the p0 model for directed graphs. For growing dimensional specified and homogeneous null hypotheses, we reveal high dimensional Wilks' phenomena that the normalized log-likelihood ratio statistic converges in distribution to a standard normal distribution.For fixed dimensional homogeneous null, we establish the Wilks-type theorem that the log-likelihood ratio test statistic converges in distribution to a chi-square distribution, not depending on the nuisance parameters.

Graph alignment problems with partially correlated nodes

Pengkun Yang (杨朋昆)
Tsinghua University

The graph alignment problem aims to find the underlying node correspondence present in two correlated graphs. Applications span various fields such as networks de-anonymization, computer vision, natural language processing, and computational biology. In this talk, I will discuss the model under which not all nodes from both graphs are correlated. By extending the functional digraph under partial correlation, we provide a tight characterization of the information-theoretic thresholds of the problem. I will also discuss the efficient algorithms and the challenges stemming from potential statistical-computational gaps.

Consistent Community Detection via Cross Validation

Yuhong Yang(杨宇红)
Tsinghua University

The problem of identifying the number of communities in a stochastic block model (SBM) is an interesting question that has been much studied in the literature, though with rather limited theoretical advancements. In this talk, we will consider cross validation methods based on nodes or edge splittings to choose the number of communities. It turns out that how the data are split plays a crucial role on if the likelihood of the true number of communities is high. Indeed, under proper conditions on data splitting, we show the true number of communities is identified with probability approaching 1. Simulations and real data examples demonstrate competitive performances of our approaches.

Autoregressive networks and stylized features

Qiwei Yao(姚琦伟)
The London School of Economics and Political Science, LSE

We give a brief introduction on the autoregressive (AR) model for dynamic network processes. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as MLEs and a permutation test for model diagnostic checking. We illustrate how this AR model can serve as a building block to accommodate more complex structures such as stochastic latent blocks, change-points. We also elucidate how some stylized features often observed in real network data, including node heterogeneity, edge sparsity, persistence, transitivity and density dependence, can be embedded in the AR framework. Then the framework needs to be extended for dynamic networks with dependent edges, which poses new technical challenges. Illustration with real network data for the practical relevance of the proposed AR framework is also presented.

TBA

Zhigang Yao(姚志刚)
National University of Singapore, NUS

TBA

Preferential Latent Space Models for Networks with Textual Edges

Emma Zhang(张菁菲)
Emory University

Many real-world networks contain rich textual information in the edges, such as email networks where an edge between two nodes is an email exchange. The useful textual information carried in the edges is often discarded in most network analyses, resulting in an incomplete view of the relationships between nodes. In this work, we represent each text document as a generalized multi-layer network and introduce a new and flexible preferential latent space network model that can capture how node-layer preferences directly modulate edge probabilities. We establish identifiability conditions for the proposed model and tackle model estimation with a computationally efficient projected gradient descent algorithm. We further derive the non-asymptotic error bound of the estimator from each step of the algorithm. The efficacy of our proposed method is demonstrated through simulations and an analysis of the Enron email network.

Efficient Estimation for Longitudinal Networks via Adaptive Merging

Haoran Zhang(张浩然)
Southern University of Science and Technology

Longitudinal network consists of a sequence of temporal edges among multiple nodes, where the temporal edges are observed in real time. It has become ubiquitous with the rise of online social platform and e-commerce, but largely under-investigated in literature. In this paper, we propose an efficient estimation framework for longitudinal network, leveraging strengths of adaptive network merging, tensor decomposition and point process. It merges neighboring sparse networks so as to enlarge the number of observed edges and reduce estimation variance, whereas the estimation bias introduced by network merging is controlled by exploiting local temporal structures for adaptive network neighborhood. A projected gradient descent algorithm is proposed to facilitate estimation, where the upper bound of the estimation error in each iteration is established. A thorough analysis is conducted to quantify the asymptotic behavior of the proposed method, which shows that it can significantly reduce the estimation error and also provides guideline for network merging under various scenarios. We further demonstrate the advantage of the proposed method through extensive numerical experiments on synthetic datasets and a militarized interstate dispute dataset.

Higher-order accurate two-sample network inference and network hashing

Yuan Zhang(张源)
The Ohio State University, OSU

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.

Inward and Outward Network Influence Analysis and A Mutual Influence Model for Two-Mode Network Data

Tao Zou(邹韬)
Australian National University

Measuring heterogeneous influence across nodes in a network is critical in network analysis. This article proposes an inward and outward network influence (IONI) model to assess nodal heterogeneity. Specifically, we allow for two types of influence parameters; one measures the magnitude of influence that each node exerts on others (outward influence), while we introduce a new parameter to quantify the receptivity of each node to being influenced by others (inward influence). Accordingly, these two types of influence measures naturally classify all nodes into four quadrants (high inward and high outward, low inward and high outward, low inward and low outward, and high inward and low outward). To demonstrate our four- quadrant clustering method in practice, we apply the quasi-maximum likelihood approach to estimate the influence parameters, and we show the asymptotic properties of the resulting estimators. In addition, score tests are proposed to examine the homogeneity of the two types of influence parameters. To improve the accuracy of inferences about nodal influences, we introduce a Bayesian information criterion that selects the optimal influence model. The usefulness of the IONI model and the four-quadrant clustering method is illustrated via

simulation studies and an empirical example involving customer segmentation. This is a joint work with Wei Lan, Yujia Wu and Chih-Ling Tsai.

A two-mode network is a distinctive network structure where nodes are divided into two specific types, and edges exclusively connect nodes of differing types. These network connections often lead to interdependencies between nodes of one type and nodes of the other type, and the nature of these relationships can vary across nodes. To examine and model this heterogeneity, we introduce a mutual influence model tailored for two-mode network data. In this model, we account for nodal heterogeneity by incorporating two sets of influence parameters. The first set of parameters gauges the extent of influence each node exerts on others, while the second set of parameters quantifies how receptive each node is to being influenced by others. To estimate this model, we introduce the quasi-maximum likelihood estimator and establish its asymptotic properties. To evaluate the performance of our proposed model and estimation in finite samples, we conduct simulation studies and provide an empirical example. This is a joint work with Rui Shen and Xuan Liang.