Workshop Abstracts

Monday (June 2)

Session - Opening Remarks (08:50 - 09:00)

Session 1 - Network Model Assessment and Structure Inference (09:00 - 10:40)

Chair: Yang Feng

Qiwei Yao (London School of Economics and Political Science)

- Goodness-of-Fit and the Best Approximation: an Adversarial Approach

Abstract: Diagnostic checking for goodness-of-fit is one of the important and routine steps in building a statistical model. The most frequently used approach for checking the goodness-of-fit is the residual analysis in the context of regression analysis. However for many statistical models there exist no natural residuals, which includes the models for the underlying distributions behind data, the models for some complex dynamic structures such as the dynamic network models with dependent edges. Furthermore, there are scenarios in which there exist several competing models but none of them are the clear favourite. One then faces a task to choose the best approximation among the wrong models. We propose an adversarial approach in this paper. For checking the goodness-of-fit of a fitted model, we generate a synthetic sample from the fitted model and construct a classifier to classify the original sample and the synthetic sample into two different class. If the fitted model is adequate, the classifier will have difficulties in distinguish the two samples. For identifying the best model among several candidate models, the classifier will create a distance between the original sample and the synthetic sample generated from each of the candidate model, and the model with the shortest distance is chosen as the best approximation for the truth.

Ji Zhu (University of Michigan)

- Hyperbolic Network Latent Space Model with Learnable Curvature

Abstract: 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.

Masaaki Imaizumi (University of Tokyo)

- Statistical Analysis On In-Context Learning

Abstract: Deep learning and artificial intelligence technologies have made great progress, and the usage of foundation models has attracted strong attention by its general ability. Motivated by this fact, mathematical understanding is required to efficiently control and develop these technologies. In this talk, I will present a statistics-based analysis of a scheme called in-context learning, which is an useful framework of meta-learning to describe foundation models. I argue that in-context learning can efficiently learn the latent structure of the data, using the property of transformers

used in the learning scheme can efficiently handle the distribution of observations.

Wen Zhou (New York University)

- Multivariate Inference of Network Moments by Subsampling

Abstract: In this paper, we study the characterization of a network population by analyzing a single observed network, focusing on the counts of multiple network motifs or their corresponding multivariate network moments. We introduce an algorithm based on node subsampling to approximate the nontrivial joint distribution of the network moments, and prove its asymptotic accuracy. By examining the joint distribution of these moments, our approach captures complex dependencies among network motifs, making a significant advancement over earlier methods that rely on individual motifs marginally. This enables more accurate and robust network inference. Through real-world applications, such as comparing coexpression networks of distinct gene sets and analyzing collaboration patterns within the statistical community, we demonstrate that the multivariate inference of network moments provides deeper insights than marginal approaches, thereby enhancing our understanding of network mechanisms.

Session - Coffee Break (10:40 - 11:10)

Session 2 - Causal & Temporal Inference in Networks (11:10 - 12:25)

Chair: Ji Zhu

Jinchi Lv (University of Southern California)

- HNCI: High-Dimensional Network Causal Inference

Abstract: 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 neighborhood size for interference effect. We exploit the model setting in Belloni et al. (2022) and allow certain type of heterogeneity in node interference neighborhood 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 neighborhood 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. This is a joint work with Rundong Ding, Wenqin Du and Yingying Fan.

Jiaming Xu (Duke University)

- A Proof of The Changepoint Detection Threshold Conjecture in Preferential Attachment Models

Abstract: We investigate the problem of detecting and estimating a changepoint in the attachment function of a network evolving according to a preferential attachment model on n vertices, using only a single final snapshot of the network. Bet et al. show that a simple test based on thresholding the number of vertices with minimum degrees can detect the changepoint when the change occurs at time $n-\Omega(\sqrt{n})$. They further make the striking conjecture that detection becomes impossible for any test if the change occurs at time $n-o(\sqrt{n}).$ Kaddouri et al. make a step forward by proving the detection is impossible if the change occurs at time $n-o(n^{1/3}).$ In this paper, we resolve the conjecture affirmatively, proving that detection is indeed impossible if the change occurs at time $n-o(\sqrt{n}).$ Furthermore, we establish that estimating the changepoint

with an error smaller than $o(\sqrt{n})$ is also impossible, thereby confirming that the estimator proposed in Bhamidi et al. is order-optimal.

Yuting Wei (University of Pennsylvania)

- To intrinsic dimension and beyond: Efficient sampling in diffusion models

Abstract: The denoising diffusion probabilistic model (DDPM) has become a cornerstone of generative AI. While sharp convergence guarantees have been established for DDPM, the iteration complexity typically scales with the ambient data dimension of target distributions, leading to overly conservative theory that fails to explain its practical efficiency. This has sparked recent efforts to understand how DDPM can achieve sampling speed-ups through automatic exploitation of intrinsic low dimensionality of data. This talk explores two key scenarios: (1) For a broad class of data distributions with intrinsic dimension k, we prove that the iteration complexity of the DDPM scales nearly linearly with k, which is optimal under the KL divergence metric; (2) For mixtures of Gaussian distributions with k components, we show that DDPM learns the distribution with iteration complexity that grows only logarithmically in k. These results provide theoretical justification for the practical efficiency of diffusion models.

Session - Lunch (12:30 - 14:00)

Session 3 - Generative, Diffusion & Attention Models (14:05 - 15:45)

Chair: Qingfeng Liu

Takeru Matsuda (University of Tokyo & RIKEN Center for Brain Science)

- Matrix estimation via singular value shrinkage

Abstract: In this talk, I will introduce recent studies on generalization of Stein's shrinkage estimation theory to matrices. Singular value shrinkage estimators and priors are shown to be minimax and work well when the unknown matrix is close to low-rank (e.g. reduced-rank regression). Further generalization to tensors will be also discussed.

Lexing Xie (Australian National University)

- Online Attention: Processes, Graphs, and Optimization

Abstract: What makes a video popular? What drives collective attention online? What are the similarities and differences between clicks and transactions in a market? This talk aims to address these three questions. First, I will discuss a physics-inspired stochastic time series model that explains and forecasts the seemingly unpredictable patterns of viewership over time. This model provides novel metrics for predicting expected popularity gains per share and assessing sensitivity to promotions. Next, I will describe new measurement studies and machine learning models that analyze how networks of online items influence each other's attention. Finally, I will introduce a macroscopic view of attention, offering mathematical descriptions of market equilibriums and distributed optimization. Our ongoing work seek computational descriptions of attention markets that can inform potential mechanisms for a healthier online ecosystem. Additionally, my group works on visualising intellectual influence and decision-making ly moral dilemmas, which opens up new questions on individual and collective attention.

Xiaotong Shen (University of Minnesota)

- Generative Score Inference for Multimodal Data

Abstract: Accurate uncertainty quantification is essential for reliable decision-making in various supervised learning scenarios, particularly when dealing with complex multimodal data such as images and text. Current approaches often face notable limitations, including rigid assumptions and limited generalizability, constraining their effectiveness across diverse supervised learning tasks. To

overcome these limitations, we introduce Generative Score Inference (GSI), a flexible inference framework capable of constructing statistically valid and informative prediction and confidence sets across a wide range of multimodal learning problems. GSI utilizes synthetic samples generated by deep generative models to approximate conditional score distributions, facilitating precise uncertainty quantification without imposing restrictive assumptions about the data or tasks. We empirically validate GSI's capabilities through two representative scenarios: hallucination detection in large language models and uncertainty estimation in image captioning. Our method achieves state-of-the-art performance in hallucination detection and robust predictive uncertainty in image captioning, and its performance is positively influenced by the quality of the underlying generative model. These findings underscore the potential of GSI as a versatile inference framework, significantly enhancing uncertainty quantification and trustworthiness in multimodal learning. Joint with Xinyu Tian of University of Minnesota.

Jiashun Jin (Carnegie Mellon University)

- TBD

Abstract: Abstract coming soon.

Session - Coffee Break (15:45 - 16:15)

Session 4 - Reception & Poster Session (16:15 - 17:45)

Chair: Yang Feng

Guillaume Braun (RIKEN Center for Advanced Intelligence Project)

- VEC-SBM: Optimal Community Detection with Vectorial Edge Covariates

Abstract: Social networks are often associated with rich side information, such as texts and images. While numerous methods have been developed to identify communities from pairwise interactions, they usually ignore such side information. In this work, we study an extension of the Stochastic Block Model (SBM), a widely used statistical framework for community detection, that integrates vectorial edges covariates: the Vectorial Edges Covariates Stochastic Block Model (VEC-SBM). We propose a novel algorithm based on iterative refinement techniques and show that it optimally recovers the latent communities under the VEC-SBM. Furthermore, we rigorously assess the added value of leveraging edge's side information in the community detection process. We complement our theoretical results with numerical experiments on synthetic and semi-synthetic data.

Takuya Koriyama (University of Chicago)

- Precise Asymptotics of Bagging Regularized M-estimators

Abstract: We characterize the squared prediction risk of ensemble estimators obtained through subagging (subsample bootstrap aggregating) regularized M-estimators and construct a consistent estimator for the risk. Specifically, we consider a heterogeneous collection of $M \ge 1$ regularized M-estimators, each trained with (possibly different) subsample sizes, convex differentiable losses, and convex regularizers. We operate under the proportional asymptotics regime, where the sample size n, feature size p, and subsample sizes k_m for $m \in [M]$ all diverge with fixed limiting ratios n/p and k_m/n. Key to our analysis is a new result on the joint asymptotic behavior of correlations between the estimator and residual errors on overlapping subsamples, governed through a (provably) contractible nonlinear system of equations. Of independent interest, we also establish convergence of trace functionals related to degrees of freedom in the non-ensemble setting (with $M = 1$) along the way, extending previously known cases for square loss and ridge, lasso regularizers. When specialized to homogeneous ensembles trained with a common loss, regularizer, and subsample size, the risk characterization sheds some light on the implicit regularization effect due to the ensemble and subsample sizes (M,k). For any ensemble size M,

optimally tuning subsample size yields sample-wise monotonic risk. For the full-ensemble estimator (when $M \to \infty$), the optimal subsample size $k^\star$ tends to be in the overparameterized regime $(k^\star \le \min\{n,p\})$, when explicit regularization is vanishing. Finally, joint optimization of subsample size, ensemble size, and regularization can significantly outperform regularizer optimization alone on the full data (without any subagging).

Issey Sukeda (University of Tokyo)

- Torus graph modeling for EEG analysis

Abstract: Identifying phase coupling recorded from multiple electrodes during electrophysiological diagnostics, such as electroencephalogram (EEG) and electrocorticography (ECoG), helps neuroscientists and clinicians understand the underlying brain structures or mechanisms. From a statistical perspective, these signals are multi-dimensional circular measurements that are correlated with one another and can be effectively modeled using a torus graph model designed for circular random variables. Using the torus graph model avoids the issue of detecting pseudo correlations. However, the naive estimation of this model tends to lead to a dense network structure, which is difficult to interpret. Therefore, to enhance the interpretability of the brain network structure, we propose to induce a sparse solution by implementing a regularized score matching estimation for the torus graph model based on the information criteria. In numerical simulations, our method successfully recovered the true dependence structure of the brain, from a synthetic dataset sampled from a pre-given torus graph model distribution. Furthermore, we present analyses of two real datasets, one involving human EEG and the other marmoset ECoG, demonstrating that our method can be widely applied to phase-coupling analysis across different types of neural data. Using our proposed method, the modularity of the estimated network structure revealed more resolved brain structures and demonstrated differences in trends among individuals.

Ergan Shang (Carnegie Mellon University)

- Inference for Balance Theory in Time-Varying Signed Network

Abstract: Dynamic signed networks are frequently used nowadays to describe the trend of two types of relationship, for example, alliances and disputes respectively, as time varies, using both positive edges and negative ones. To understand the connectivity property of signed network in different timestamps, the social balance theory considers the fully connected 3-node coliques and evaluate the portion of balanced triangles. To alleviate the hard assumption of relatively large sparsity level for making inference, we set up a statistical dynamic signed network model and propose a kernel smoothing estimator to make inference on the balanced portion parameter at the time even not observed in the data. In the process, we also give a bandwidth selection method to use in practice. Moreover, the theoretical guarantee of the error rate of our method is boosted by leveraging multiple graphs for inference under the alleviated requirement on the sparsity parameter. Finally, we apply our method on simulation studies and a real dataset of international alliances and disputes in the area of political science to exploit the inference power of our methodology.

Tao Shen (National University of Singapore)

- Optimal Network-Guided Covariate Selection for High-Dimensional Data Integration

Abstract: When integrating datasets from different studies, it is common that they have components of different formats. How to combine them organically for improved estimation is important and challenging. This work investigates this problem in a two-study scenario, where covariates are observed for all subjects, but network data is available in only one study, and response variables are available only in the other. To leverage the partially observed network information, we propose the Network-Guided Covariate Selection (NGCS) algorithm. It integrates the spectral information from network adjacency matrices with the Higher Criticism Thresholding

approach for informative covariates identification. Theoretically, we prove that NGCS achieves the optimal rate in covariate selection, which is the same rate in the supervised learning setting. Furthermore, this optimality is robust to network models and tuning parameters. This framework extends naturally to clustering and regression tasks, with two proposed algorithms: NG-clu and NG-reg. Empirical studies on synthetic and real-world datasets demonstrate the robustness and superior performance of our algorithms, underscoring their effectiveness in handling heterogeneous data formats.

Fan Wang (University of Warwick)

- Change Point Analysis in Dynamic Multilayer Networks

Abstract: In this talk, I will introduce the multilayer random dot product graph (MRDPG) model, a generalization of the random dot product graph model to multilayer networks. To estimate edge probabilities, I will present a tensor-based methodology and demonstrate its superiority over existing approaches. Moving to dynamic MRDPGs, I will formulate and analyze an online change point detection framework, where, at each time point, a realization from an MRDPG is observed. I will propose a novel nonparametric change point detection algorithm based on density kernel estimators and tensor-based methods. This approach is broadly applicable to various network settings, including stochastic block models as special cases. Theoretically, I will show that our methods effectively minimize the detection delay while controlling false alarms. Extensive numerical experiments, including an application to U.S. air transportation networks, supported our theoretical findings.

Bingcheng Sui (University of Science and Technology of China)

- Counting Cycles with AI

Abstract: Despite recent progress, AI still struggles on advanced mathematics. We consider a difficult open problem: how to derive an equivalent expression for the cycle count statistics that leads to a computational approach with minimal cost. The problem does not have known general solutions, and requires delicate combinatorics and tedious calculations. Such a task is hard to accomplish by human but is an ideal example where AI can be very helpful. We solve the problem by combining a novel approach we propose and the powerful coding skills of AI. Our results use delicate graph theory and contain new formula for general cases that have not been discovered before. We find that, while AI is unable to solve the problem all by itself, it is able to solve it if we provide it with a clear strategy, a step-by-step guidance and carefully written prompts. For simplicity, we focus our study on DeepSeek-R1 but we also investigate other AI approaches.

Tuesday (June 3)

Session 5 - High-Dimensional Statistics and Tensor Methods (09:00 - 10:40)

Chair: Jiashun Jin

Cun-Hui Zhang (Rutgers University)

- Simultaneous Decorrelation of Matrix Time Series

Abstract: We propose a contemporaneous bilinear transformation for a p-by-q matrix time series to alleviate the difficulties in modeling and forecasting matrix time series when p and/or q are large. The resulting transformed matrix assumes a block structure consisting of several small matrices, and those small matrix series are uncorrelated across all times. Hence an overall parsimonious model is achieved by modelling each of those small matrix series separately without the loss of information on the linear dynamics. Such a parsimonious model often has better forecasting performance, even when the underlying true dynamics deviates from the assumed uncorrelated

block structure after transformation. The uniform convergence rates of the estimated transformation are derived. The proposed method is illustrated numerically via both simulated and real data examples. This is joint work with Yuefeng Han, Rong Chen and Qiwei Yao.

Lexin Li (University of California Berkeley)

- Tensor Data Analysis and Some Applications in Neuroscience

Abstract: 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 modeling 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.

Jeff Yao (Chinese University of Hong Kong Shenzhen)

- Alignment and matching tests for high-dimensional tensor signals via tensor contraction

Abstract: We consider two hypothesis testing problems for low-rank and high-dimensional tensor signals, namely the tensor signal alignment and tensor signal matching problems. These problems are challenging due to the high dimension of tensors and lack of meaningful test statistics. By exploiting a recent tensor contraction method, we propose and validate relevant test statistics using eigenvalues of a data matrix resulting from the tensor contraction. The matrix has a long range dependence among its entries, which makes the analysis of the matrix challenging, involved and distinct from standard random matrix theory. Our approach provides a novel framework for addressing hypothesis testing problems in the context of high-dimensional tensor signals. This is a joint work with Ruihan Liu (The University of Hong Kong) and Zhenggang Wang (Southeast University). Full paper available at https://arxiv.org/abs/2411.01732

Yuqi Gu (Columbia University)

- Minimax-Optimal Dimension-Reduced Clustering for High-Dimensional Nonspherical Mixtures

Abstract: In mixture models, nonspherical (anisotropic) noise within each cluster is widely present in real-world data. We study both the minimax rate and optimal statistical procedure for clustering under high-dimensional nonspherical mixture models. In high-dimensional settings, we first establish the information-theoretic limits for clustering under Gaussian mixtures. The minimax lower bound unveils an intriguing informational dimension-reduction phenomenon: there exists a substantial gap between the minimax rate and the oracle clustering risk, with the former determined solely by the projected centers and projected covariance matrices in a low-dimensional space. Motivated by the lower bound, we propose a novel computationally efficient clustering method: Covariance Projected Spectral Clustering (COPO). Its key step is to project the high-dimensional data onto the low-dimensional space spanned by the cluster centers and then use the projected covariance matrices in this space to enhance clustering. We establish tight algorithmic upper bounds for COPO, both for Gaussian noise with flexible covariance and general noise with local dependence. Our theory indicates the minimax-optimality of COPO in the Gaussian case and highlights its adaptivity to a broad spectrum of dependent noise. Extensive simulation studies under various noise structures and real data analysis demonstrate our method's superior performance.

Workshop Abstracts

Session - Coffee Break (10:40 - 11:10)

Session 6 - Privacy-Preserving & Dynamic Network Learning (11:10 - 12:25)

Chair: Xiaoyue Niu

Tony Cai (University of Pennsylvania)

- TBD

Abstract: Abstract coming soon.

Yi Yu (University of Warwick)

- Optimal federated learning under differential privacy constraints

Abstract: With the growing computational power and increasing awareness of privacy, federated learning has emerged as a pivotal framework for private, distributed data analysis. Depending on applications, diverse privacy constraints come into play, each imposing a unique cost on statistical accuracy. In this talk, I will start with an overview of the foundational concept of differential privacy (DP). I will then introduce three notions of DP tailored to the federated learning context, highlighting their relevance and implications in distributed settings. The core focus of this talk will be on a functional data estimation problem under a hierarchical and heterogeneous DP framework. I will discuss how privacy constraints impact estimation accuracy and quantify these tradeoffs through the lens of minimax theory. Key aspects of the proofs will also be outlined, as well as some numerical performances.

Yang Feng (New York University)

- Variational Nonparametric Inference in Functional Stochastic Block Model

Abstract: We propose a functional stochastic block model whose vertices involve functional data information. This new model extends the classic stochastic block model with vector-valued nodal information, and finds applications in real-world networks whose nodal information could be functional curves. Examples include international trade data in which a network vertex (country) is associated with the annual or quarterly GDP over certain time period, and MyFitnessPal data in which a network vertex (MyFitnessPal user) is associated with daily calorie information measured over certain time period. Two statistical tasks will be jointly executed. First, we will detect community structures of the network vertices assisted by the functional nodal information. Second, we propose computationally efficient variational test to examine the significance of the functional nodal information. We show that the community detection algorithms achieve weak and strong consistency, and the variational test is asymptotically chi-square with diverging degrees of freedom. As a byproduct, we propose pointwise confidence intervals for the slop function of the functional nodal information. Our methods are examined through both simulated and real datasets.

Session - Group Photo (12:25 - 12:30)

Session - Lunch (12:30 - 14:00)

Session 7 - Data Integration & Applications (14:05 - 15:45)

Chair: Emma Jingfei Zhang

Annie Qu (University of California Irvine)

- High-order Joint Embedding for Multi-Level Link Prediction

Abstract: Link prediction infers potential links from observed networks, and is one of the essential

problems in network analyses. In contrast to traditional graph representation modeling which only predicts two-way pairwise relations, we propose a novel tensor-based joint network embedding approach on simultaneously encoding pairwise links and hyperlinks onto a latent space, which captures the dependency between pairwise and multi-way links in inferring potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to capture richer network information. In addition, the proposed method introduces a hierarchical dependency among links to infer potential hyperlinks, and leads to better link prediction. In theory we establish the estimation consistency for the proposed embedding approach, and provide a faster convergence rate compared to link prediction utilizing pairwise links or hyperlinks only. Numerical studies on both simulation settings and Facebook ego-networks indicate that the proposed method improves both hyperlink and pairwise link prediction accuracy compared to existing link prediction algorithms. This is a joint work with Prof. Yubai Yuan at Penn State.

Peter Song (University of Michigan)

- Network Structural Equation Models for Causal Mediation and Spillover Effects

Abstract: Social network interference induces spillover effects from neighbors' exposures, and the complexity of statistical analysis increases when mediators are involved with network interference. We develop a theoretical framework employing a structural graphical modeling approach to investigate both mediation and interference effects within network data. Our framework enables us to capture the multifaceted mechanistic pathways through which neighboring units' exposures and mediators exert direct and indirect influences on an individual's outcome. We extend the exposure mapping paradigm in the context of a random-effects network structural equation models (REN-SEM), establishing its capacity to delineate spillover effects of interest. Identifiability conditions for both causal mediation and interference estimands are postulated rigorously. Our proposed methodology contributions include maximum likelihood estimation for REN-SEM and inference procedures with theoretical guarantees. Such guarantees encompass consistent asymptotic variance estimators, derived under a non-i.i.d. asymptotic theory. The robustness and practical utility of our methodology are demonstrated through simulation experiments, underscoring its effectiveness in capturing the intricate dynamics of network-mediated exposure effects.

Tracy Ke (Harvard University)

- Poisson-Process Topic Model for Integrating Knowledge from Pre-trained Language Models

Abstract: Topic modeling is traditionally applied to word counts without accounting for the context in which words appear. Recent advancements in large language models (LLMs) offer contextualized word embeddings, which capture deeper meaning and relationships between words. We aim to leverage such embeddings to improve topic modeling. We use a pre-trained LLM to convert each document into a sequence of word embeddings. This sequence is then modeled as a Poisson point process, with its intensity measure expressed as a convex combination of K base measures, each corresponding to a topic. To estimate these topics, we propose a flexible algorithm that integrates traditional topic modeling methods, enhanced by net-rounding applied before and kernel smoothing applied after. One advantage of this framework is that it treats the LLM as a black box, requiring no fine-tuning of its parameters. Another advantage is its ability to seamlessly integrate any traditional topic modeling approach as a plug-in module, without the need for modifications. Assuming each topic is a Î²-HÃ¶lder smooth intensity measure on the embedded space, we establish the rate of convergence of our method. We apply our method to several datasets, providing evidence that it offers an advantage over traditional topic modeling approaches.

Hui Shen (McGill University)

- Consistent Identification of Top-K Nodes in Noisy Networks

Abstract: Identifying the most important nodes in a network, often defined via centrality measures, is a fundamental challenge in applied network analysis. However, real-world networks are frequently constructed from noisy or incomplete data, leading to distortions in centrality rankings and misidentification of key nodes. In this paper, we systematically examine how network noise affects the accurate recovery of the true top-k node set, as measured by degree centrality. Specifically, we analyze the performance of the empirical top-k set obtained from a noisy observation of the network, where edges are randomly added or removed under a probabilistic noise model. We derive conditions for consistent recovery and characterize regimes where reliable identification is provably impossible. To assess the stability of the empirical ranking, we establish sharp lower and upper bounds on the expected set difference between the empirical and true top-k sets, both in general and under specific network models. Simulation studies corroborate our theoretical findings and demonstrate the practical value of these bounds across a variety of network settings.

Session - Coffee Break (15:45 - 16:15)

Session 8 - Spectral / Hypergraph & Signal-Detection Methods (16:15 - 17:30)

Chair: Rachel Wang

Emma Jingfei Zhang (Emory University)

- Modeling Non-Uniform Hypergraphs Using Determinantal Point Processes

Abstract: Most statistical models for networks focus on pairwise interactions between nodes. However, many real-world networks involve higher-order interactions among 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 (i.e., edges of the same size) or rely on diversity among nodes. In this work, we propose a new hypergraph model based on non-symmetric determinantal point processes. The proposed model naturally accommodates non-uniform hyperedges, has tractable probability mass functions, and accounts for both node similarity and diversity in hyperedges. For model estimation, we maximize the likelihood function under constraints using a computationally efficient projected adaptive gradient descent algorithm. We 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 edge predictions on several real-world datasets.

Zhigang Bao (Hong Kong University of Science and Technology)

- Signal Detection from Spiked Noise via Asymmetrization

Abstract: The signal-plus-noise model is a fundamental framework in signal detection, where a low-rank signal is corrupted by noise. In the high-dimensional setting, one often uses the leading singular values and corresponding singular vectors of the data matrix to conduct statistical inference on the signal component. Specifically, when the noise consists of i.i.d. random entries, the singular values of the signal component can be estimated from those of the data matrix, provided the signal is sufficiently strong. However, when the noise entries are heteroscedastic or correlated, this standard approach may fail. In particular, this talk considers a challenging scenario that arises with heteroscedastic noise: when the noise itself can create spiked singular values. This raises the recurring question of how to distinguish the signal from the spikes in the noise. To address this, we study the eigenvalues of an asymmetrized model when two samples are available. We demonstrate that by examining the leading eigenvalues (in magnitude) of the asymmetrized model, one can reliably detect the signal. This approach is effective even in the heavy-tailed regime, where the singular value method fails.

Aaron Bramson (GA technologies)

- Measuring Housing Demand Using Multimodal Geospatial Networks

Abstract: Residential demand estimation is important for determining rents, sales prices, vacancy risk, areas underserviced and overserviced by transportation resources, proper zoning, appropriate/optimal land usage, as well as other economic and policy related decisions. Our approach calculates demand as the time-weighted potential flow of people from places of work, across a detailed multimodal transportation network, to each location in the greater Tokyo area. Unfortunately, diffusion on the road/walking network is too computationally expensive. As a solution, we create a 250m inner diameter hexagonal grid covering the greater Tokyo area (126,440 hexes) and convert it into a network using interhex traversal times determined from the road network. The resulting hex network has origin-destination traversal times within a 4% error of direct road network traversals. We find that our demand flow scores alone have a 0.829 Pearson correlation with rental prices. We further find that this geospatial network feature can fully replace implicitly spatial variables (such as coordinates and neighborhood names) in an ï¿½explicable Alï¿½ LGBM pricing model with an R2 of 0.935. Refinements to the hex network, the inclusion of buses, and additional sources of demand are expected to increase the quality of these assessments.

Session - Banquet (18:30 - 21:00)

Wednesday (June 4)

Session 9 - Complex Networks and Multilayer Networks (09:00 - 10:40)

Chair: Kaizheng Wang

Xinyu Zhang (Chinese Academy of Sciences)

- A Transfer Learning Framework for Multilayer Networks via Model Averaging

Abstract: Link prediction in multilayer networks is a key challenge in applications such as recommendation systems and protein-protein interaction prediction. While many techniques have been developed, most rely on assumptions about shared structures and require access to raw auxiliary data, limiting their practicality. To address these issues, we propose a transfer learning framework for multilayer networks using a bi-level model averaging method. Our approach introduces a cross-validation criterion based on edges to automatically weight inter-layer and intra-layer candidate models. Theoretically, we prove the optimality and weight convergence of our method under mild conditions. Computationally, our framework is efficient and privacy-preserving, as it avoids raw data sharing and supports parallel processing across multiple servers. Simulations show our method outperforms others in predictive accuracy and robustness. We further demonstrate its practical value through two real-world recommendation system applications.

Rachel Wang (University of Sydney)

- Network autoregression for the propagation of binary responses in social networks

Abstract: Studying the propagation of binary responses on nodes in a large-scale social network is critical for understanding how individual behaviors and decisions are shaped by social structures and for predicting collective outcomes. We propose a network autoregressive model for binary-valued responses, in which the probability of response at each node is influenced by its neighbors' past decisions, its own past decision, and node-specific covariates, through a logistic link function. The model accounts for network noise and community structure by assuming the underlying network is generated from a block model, with autoregressive parameters that are community-specific. We establish conditions under which the long term behavior of the

high-dimensional binary vector converges to a community-specific distribution and the associated convergence rate, illustrating when individuals in the same community or across the whole network reach a consensus regardless of their initial positions. Given an observed network and response vectors, we show asymptotic consistency and normality of the maximum likelihood estimators. We demonstrate the efficiency and validity of the inference procedure through simulated and real data. In particular, we show the model can be used to study the dynamics of strike occurrences in China and highlight the impact of online social network in facilitating collective actions.

Ji Oon Lee (Korea Advanced Institute of Science & Technology)

- Detection problems in spiked Wigner matrices

Abstract: The spiked Wigner matrix model is one of the most basic yet fundamental models for signal-plus-noise type data, where the signal is a vector and the noise is a symmetric random matrix. One of the main questions in the study of the spiked Wigner matrices is the detection problem, where the main goal is to detect the presence or the absence of the signal in a given data matrix. In this talk, I will explain various mathematical results on the detection problem, such as the fundamental limit and detecting algorithms, which are based on the study of random matrices and spin glass models.

Wanjie Wang (National University of Singapore)

- Data Integration: Network-Guided Covariate Selection

Abstract: In the era of data, the integration of data becomes more and more important. The data may come from different studies, different clients, or the same client but on different aspects. In this work, we are interested in the social platform, where we can observe both the user-user connection data (network) and the user profile/tags/posts (covariates). Our question arises: can we find the influential covariates, i.e. covariates that are related to the hidden information of users? Without network information, this is an unsupervised learning problem. Based on the standard procedure, we propose a network-guided covariate selection algorithm. Leveraging the network information, we significantly improve the selection power. The algorithm is efficient and robust to various network models. Finally, we discuss the downstream applications with selected covariates, including clustering and regression.

Session - Coffee Break (10:40 - 11:10)

Session 10 - Advances in Statistical Learning and Uncertainty Quantification (11:10 - 12:25)

Chair: Qiyang Han

Xinghua Zheng (Hong Kong University of Science and Technology)

- Stock Co-jump Network Models based on Site Percolation

Abstract: Stock prices often exhibit co-jumps, even in the absence of market jumps. To capture such phenomena, we introduce a class of network models based on site-percolation, which contrasts with usual network models that rely on edge-based connections to represent dependencies. We discuss the fundamental differences between these two modeling approaches, develop community detection methods tailored to our proposed framework, and demonstrate their economic significance through empirical applications.

Robert Lunde (Washington University at St Louis)

- Conformal Prediction for Dyadic Regression Under Structured Missingness

Abstract: Dyadic regression, which involves modeling a relational matrix given covariate

information, is an important task in statistical network analysis. We consider uncertainty quantification for dyadic regression models using conformal prediction. We establish finite-sample validity of our procedures for various sampling mechanisms under a joint exchangeability assumption. Our proof uses new results related to the validity of conformal prediction beyond exchangeability, which may be of independent interest. We also show that, under certain conditions, it is possible to construct asymptotically valid prediction sets for a missing entry under a structured missingness assumption.

Kaizheng Wang (Columbia University)

- Uncertainty Quantification for LLM-Based Survey Simulations

Abstract: We investigate the reliable use of simulated survey responses from large language models (LLMs) through the lens of uncertainty quantification. Our approach converts synthetic data into confidence sets for population parameters of human responses, addressing the distribution shift between the simulated and real populations. A key innovation lies in determining the optimal number of simulated responses: too many produce overly narrow confidence sets with poor coverage, while too few yield excessively loose estimates. To resolve this, our method adaptively selects the simulation sample size, ensuring valid average-case coverage guarantees. It is broadly applicable to any LLM, irrespective of its fidelity, and any procedure for constructing confidence sets. Additionally, the selected sample size quantifies the degree of misalignment between the LLM and the target human population. We illustrate our method on real datasets and LLMs.

Session - Lunch (12:30 - 14:00)

Session 11 - Graphical Models, Hypergraphs, and Generative Approaches (14:05 - 15:20)

Chair: Xinyu Zhang

Nick Whiteley (University of Edinburgh)

- Statistical exploration of the Manifold Hypothesis

Abstract: The Manifold Hypothesis is a widely held tenet of Machine Learning which asserts that nominally high-dimensional data are in fact concentrated near a low-dimensional manifold, embedded in high-dimensional space. This phenomenon is observed empirically in many real world situations, has led to development of a wide range of statistical methods in the last few decades, and has been suggested as a key factor in the success of modern AI technologies. We show that rich manifold structure in data can emerge from a generic and remarkably simple statistical model -- the Latent Metric Model -- via elementary concepts such as latent variables, correlation and stationarity. This establishes a general statistical explanation for why the Manifold Hypothesis seems to hold in so many situations. Informed by the Latent Metric Model we derive procedures to discover and interpret the geometry of high-dimensional data, and explore hypotheses about the true data generating mechanism. This is joint work with Annie Gray and Patrick Rubin-Delanchy.

Jingming Wang (University of Virginia)

- Network Goodness-of-Fit for the block-model family

Abstract: The block model family is widely used in network modeling and includes four popular models: SBM, DCBM, MMSBM and DCMM. However, the question of which block model best fits real networks has received limited attention in the literature. In this talk, I will introduce a novel approach using cycle count statistics to address the Goodness-of-Fit for these block models. By leveraging the cycle count statistics and a network fitting scheme, we construct four GoF metrics with parameter-free limiting distributions of $N(0,1)$ under the assumed models. We apply these

GoF-metrics to some frequently-used real networks for comparison. The numerical results suggest that DCMM is particularly promising for modeling undirected networks. This talk is based on joint work with Jiashun Jin, Tracy Ke, and Jiajun Tang.

Yao Xie (Georgia Institute of Technology)

- Scalable flow-based generative models for network data

Abstract: Generative models have become a transformative technology in machine learning, with significant advancements made for vector-valued data. However, generative modeling for network data, i.e., data with underlying graph topology remains an emerging area. In this work, we develop a deep generative model, the invertible graph neural network(iGNN), for network data by formulating it as a conditional generative task based on flow-based generative models. The proposed model is capable of generating synthetic samples and performing probabilistic prediction by capturing correlations among nodal observations while quantifying uncertainty. It consists of an invertible sub-network that establishes a one-to-one mapping from data to encoded features, enabling forward prediction via a linear classification sub-network and efficient generation from output labels through a parametric mixture model. The invertibility of the encoding sub-network is enforced via Wasserstein-2 regularization, which allows flexible, free-form layers in the residual blocks. A key feature of the model is its scalability to large graphs, achieved through a factorized parametric mixture model of the encoded features, along with computational efficiency provided by GNN layers. The existence of an invertible flow mapping is supported by theories from optimal transport and diffusion processes, and we prove the expressiveness of graph convolution layers in approximating the theoretical flows of graph data. The proposed model is evaluated on synthetic datasets, including large-scale graph examples, and its empirical advantages are demonstrated on real-world applications such as solar ramping event prediction and traffic flow anomaly detection.

Session - Coffee Break (15:20 - 15:50)

Session 12 - Random Matrix Theory and Differential Privacy (15:50 - 17:05)

Chair: Wanjie Wang

Qiyang Han (Rutgers University)

- Algorithmic inference via gradient descent: from linear regression to neural networks

Abstract: Classical statistical methods typically rely on constructing a single estimator with desirable properties, assuming a clear separation between statistical optimality and algorithmic achievability. In contrast, modern machine learning applies simple first-order methods to complex non-convex models, often without any provable guarantees of convergence to traditional empirical risk minimizers. Can we still perform precise statistical inference in such settings? This talk introduces a new framework for algorithmic inference using the simplest gradient descent algorithm. Under suitable random data models, we show that gradient descent can be augmented, via a few auxiliary computations, to perform various valid inference tasks. We illustrate this approach in both classical linear and logistic regression models, as well as in significantly more complex multi-layer neural network models. In both cases, the augmented gradient descent algorithm produces, at each iteration, consistent estimates of the generalization error and valid confidence intervals for the unknown signal. In principle, these iteration-wise estimates can inform practical decisions such as early stopping and hyperparameter tuning during gradient descent training. This new algorithmic inference approach relies on a recent state evolution theory for a class of General First-Order Methods (GFOMs) developed by the author.

Yumou Qiu (Peking University)

- Versatile differentially private learning for general loss functions

Abstract: This paper aims to provide a versatile privacy-preserving release mechanism along with a unified approach for subsequent parameter estimation and statistical inference. We propose the ZIL privacy mechanism based on zero-inflated symmetric multivariate Laplace noise, which requires no prior specification of subsequent analysis tasks, allows for general loss functions under minimal conditions, imposes no limit on the number of analyses, and is adaptable to the increasing data volume in online scenarios. We derive the trade-off function for the proposed ZIL mechanism that characterizes its privacy protection level. Within the M-estimation framework, we propose a novel doubly random corrected loss (DRCL) for the ZIL mechanism, which provides consistent and asymptotic normal M-estimates for the parameters of the target population under differential privacy constraints. The proposed approach is easy to compute without numerical integration and differentiation for noisy data. It is applicable for a general class of loss functions, including non-smooth loss functions like check loss and hinge loss. Simulation studies, including logistic regression and quantile regression, are conducted to evaluate the performance of the proposed method.

Guangming Pan (Nanyang Technological University)

- Eigenvector overlaps in large sample covariance matrices and nonlinear shrinkage estimators

Abstract: Consider a data matrix $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$ of size $M \times N$, where the columns are independent observations from a random vector $\mathbf{y}$ with zero mean and population covariance Σ. Let u_i and v_j denote the left and right singular vectors of Y, respectively. This study investigates the eigenvector/singular vector overlaps $\langle{u_i, D_1 u_j}\rangle$, $\langle{v_i, D_2 v_j}\rangle$ and $\langle{u_i, D_3 v_j}\rangle$, where D_k are general deterministic matrices with bounded operator norms. In the high-dimensional regime, where the dimension M scales proportionally with the sample size N, we establish the convergence in probability of these eigenvector overlaps towards their deterministic counterparts with explicit convergence rates. Building on these findings, we offer a more precise characterization of the loss associated with Ledoit and Wolf's nonlinear shrinkage estimators of the population covariance Σ.

Session - Closing Remarks and Poster Awards (17:05 - 17:30)