VIASM
VIETNAM INSTITUTE FOR ADVANCED STUDY IN MATHEMATICS

STATISTICAL NETWORK ANALYSIS AND BEYOND

June 6–10, 2026

Vietnam Institute for Advanced Study in Mathematics (VIASM)
Hanoi, Vietnam

This is the booklet for the Statistical Network Analysis and Beyond 2026 workshop.
Further information can be found at:
`https://sites.google.com/view/snab-workshop-2026`

Contents

About

SNAB (Statistical Network Analysis and Beyond) 2026 is a three-day workshop that brings together experts in statistical network analysis and related areas. It provides a forum for presenting cutting-edge research and fostering new collaborations across disciplines.

Organizing Committee

Yuqi Gu (Columbia University, USA)
Lê Minh Hà (VIASM, Vietnam)
Can Le (University of California, Davis, USA)
Xiaodong Li (University of California, Davis, USA)
Jiashun Jin (Carnegie Mellon University, USA & Southeast University, China)
Ji Zhu (University of Michigan, USA)

Mini-course Lecturers

Can Le (University of California, Davis, USA)
Jiashun Jin (Carnegie Mellon University, USA & Southeast University, China)

Speakers (alphabetical)

Joshua Agterberg (University of Illinois Urbana–Champaign, USA)
Tony Cai (University of Pennsylvania, USA)
Hao Chen (University of California, Davis, USA)
Huimin Cheng (Boston University, USA)
Yuqi Gu (Columbia University, USA)
Yinqiu He (University of Wisconsin–Madison, USA)
Jiashun Jin (Carnegie Mellon University, USA, and Southeast University, China)
Yongdai Kim (Seoul National University, South Korea)
Can Le (University of California, Davis, USA)
Ngoc Le (Hanoi University of Science and Technology, Vietnam)
Chenlei Leng (Hong Kong Polytechnic University, Hong Kong)
Xiaodong Li (University of California, Davis, USA)
Zach Lubberts (University of Virginia, USA)
Vince Lyzinski (University of Maryland, USA)
Tan Minh Nguyen (National University of Singapore, Singapore)
Maggie Niu (Pennsylvania State University, USA)
Marianna Pensky (University of Central Florida, USA)
Yumou Qiu (Peking University, China)
Georgia Smits (University of Washington, USA)
Minh Tang (North Carolina State University, USA)
Jingming Wang (University of Virginia, USA)
Junhui Wang (Chinese University of Hong Kong, Hong Kong)
Wanjie Wang (National University of Singapore, Singapore)
Min Xu (Rutgers University, USA)
Yuhong Yang (Tsinghua University, China)
Anderson Ye Zhang (University of Pennsylvania, USA)
Emma Zhang (Emory University, USA)
Shurong Zheng (Northeast Normal University, China)
Wen Zhou (New York University, USA)
Ji Zhu (University of Michigan, USA)

Timetable

Monday, June 8

08:45 – 09:05	**Registration**	
09:05 – 09:15	**Opening Remarks**	
09:15 – 09:40	**Tony Cai** University of Pennsylvania	*Optimal Federated Learning for Functional Mean Estimation under Heterogeneous Privacy Constraints*
09:40 – 10:05	**Vince Lyzinski** University of Maryland, College Park	*Gaussian Mixture Models as a Proxy for Interacting Language Models*
10:05 – 10:30	**Jingming Wang** University of Virginia	*Estimating Spikes by Counting Cycles*
10:30–11:00	**Coffee Break**	
11:00–11:25	**Ji Zhu** University of Michigan	*Statistical Inference for Latent Space Models of Network Data with Edge Covariates*
11:25–11:50	**Junhui Wang** The Chinese University of Hong Kong	*Community Detection in Heterogeneous Signed Networks*
11:50–12:15	**Yuqi Gu** Columbia University	*Discrete Causal Representation Learning*
12:15–13:55	**Lunch**	
14:00–14:25	**Yongdai Kim** Seoul National University	*A Composite Activation Function for Learning Stable Binary Representations*
14:25–14:50	**Yinqiu He** University of Wisconsin–Madison	*Bridging Theory and Practice: Statistical Inference of Latent Space Models for Networks*
14:50–15:15	**Min Xu** Rutgers University	*Community Detection on a Randomly Growing Network*
15:15–15:45	**Coffee Break**	
15:45–16:10	**Emma Zhang** Emory University	*Estimating Treatment and Spillover Effects with the Ego-Cluster Experimental Design*
16:10–16:35	**Minh Tang** North Carolina State University	*Eigenvector Fluctuations and Limit Results for Random Graphs with Infinite Rank Kernels*
16:35–17:00	**Joshua Agterberg** University of Illinois Urbana–Champaign	*Statistically and Computationally Optimal Estimation and Inference of Common Subspaces*

Tuesday, June 9

08:45–09:10	**Jiashun Jin** Carnegie Mellon University	*Estimating the reciprocal effect by a cancellation trick*
09:10–09:35	**Maggie Niu** Pennsylvania State University	*A Multilayer Network Model for Aggregated Relational Data*
09:35–10:00	**Wen Zhou** New York University	*Adaptive Inference for Stratified Network Effects with an Application to Asymmetric Partisan Polarization*
10:00–10:30	**Coffee Break**	
10:30–10:55	**Marianna Pensky** University of Central Florida	*Scalable Community Detection in Massive Networks via Predictive Assignment*
10:55–11:20	**Zach Lubberts** University of Virginia	*Vertex Alignment and Changepoint Localization in Network Time Series*
11:20–11:45	**Yumou Qiu** Peking University	*A Human-guided AI Approach for Symbolic Cumulant Calculation*
11:45 – 13:25	**Lunch**	
13:30–13:55	**Yuhong Yang** Tsinghua University	*Penalized Network Cross-Validation*
13:55–14:20	**Wanjie Wang** National University of Singapore	*Optimal Differential Privacy on Networks and Bipartite Networks*
14:20–14:45	**Can Le** University of California, Davis	*Parametric Bootstrap for Fixed Edge-Probability Network Models*
14:45 – 15:15	**Coffee Break**	
15:15–15:40	**Georgia Smits** University of Washington	*Stable Motifs in a Rashomon Set of Networks*
15:40–16:05	**Huimin Cheng** Boston University	*Trustworthy Biomedical Knowledge Graph Construction and Reasoning with Human-in-the-Loop LLMs*
16:05–16:30	**Tan Minh Nguyen** National University of Singapore	*Tight Clusters Make Specialized Experts*
16:30 – 17:30	**Poster Session**	
18:00 – 20:30	**Banquet**	

Wednesday, June 10

08:45–09:10	**Shurong Zheng** Northeast Normal University	*Prediction Risk in High-Dimensional Ridge Regression*
09:10–09:35	**Anderson Ye Zhang** University of Pennsylvania	*Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery*
09:35–10:00	**Ngoc Le** Hanoi University of Science and Technology	*Network analysis for trustworthiness and applications*
10:00 – 10:30	**Coffee Break**	
10:30–10:55	**Chenlei Leng** The Hong Kong Polytechnic University	*A Propagation Framework for Network Regression*
10:55–11:20	**Hao Chen** University of California, Davis	*Community Detection Across Mixing Patterns for Two or More Communities*
11:20–11:45	**Xiaodong Li** University of California, Davis	*Sparse Multitask Regression with a Task Graph*
11:45 – 12:00	**Closing Remarks and Poster Awards**	
12:00 – 13:30	**Lunch**	

Mini-Courses

Saturday, June 6

09:00–10:00	**Can Le** University of California, Davis	*Network Data and Exploratory Analysis*
10:00–10:30	**Coffee Break**	
10:30–11:30	**Can Le** University of California, Davis	*Random Network Models*
11:30–13:30	**Lunch**	
13:30–14:30	**Can Le** University of California, Davis	*Spectral and Variational Methods for Network Analysis*
14:30–15:00	**Coffee Break**	
15:00–16:00	**Can Le** University of California, Davis	*Models for Network-Linked Data*

Sunday, June 7

09:00–10:00	**Jiashun Jin** Carnegie Mellon University and Southeast University	*Introduction: datasets and models*
10:00–10:30	**Coffee Break**	
10:30–11:30	**Jiashun Jin** Carnegie Mellon University and Southeast University	*Community detection by SCORE*
11:30–13:30	**Lunch**	
13:30–14:30	**Jiashun Jin** Carnegie Mellon University and Southeast University	*Network mixed-membership estimation and the statistical triangle*
14:30–15:00	**Coffee Break**	
15:00–16:00	**Jiashun Jin** Carnegie Mellon University and Southeast University	*Network testing by the cycle count statistics*

Talk Titles and Abstracts

Monday, June 8

Optimal Federated Learning for Functional Mean Estimation under Heterogeneous Privacy Constraints

Tony Cai **University of Pennsylvania**

Federated learning (FL) is a distributed machine learning technique designed to preserve data privacy and security, and it has gained significant importance due to its broad range of applications. In this talk, we discuss the problem of optimal functional mean estimation from discretely sampled data in a federated setting. We consider a heterogeneous framework where the number of individuals, measurements per individual, and privacy parameters vary across one or more servers, under both common and independent design settings. In the common design setting, the same design points are measured for each individual, whereas in the independent design, each individual has their own random collection of design points. Within this framework, we establish minimax upper and lower bounds for the estimation error of the underlying mean function, highlighting the nuanced differences between common and independent designs under distributed privacy constraints. We propose algorithms that achieve the optimal trade-off between privacy and accuracy and provide optimality results that quantify the fundamental limits of private functional mean estimation across diverse distributed settings. These results characterize the cost of privacy and offer practical insights into the potential for privacy-preserving statistical analysis in federated environments.

Gaussian Mixture Models as a Proxy for Interacting Language Models

Vince Lyzinski **University of Maryland, College Park**

Large language models (LLMs) are powerful tools that, in a number of settings, overlap with the results of human pattern recognition and reasoning. Retrieval-augmented generation (RAG) further allows LLMs to produce tailored output depending on the contents of their RAG databases. However, LLMs depend on complex, computationally expensive algorithms. In this paper, we introduce interacting Gaussian mixture models (GMMs) as a proxy for interacting LLMs. We construct a model of interacting GMMs, complete with an analogue to RAG updating, under which GMMs can generate, exchange, and update data and parameters. We show that this interacting system of Gaussian mixture models, which can be implemented at minimal computational cost, mimics certain aspects of experimental simulations of interacting LLMs whose iterative responses depend on feedback from other LLMs. We build a Markov chain from this system of interacting GMMs; formalize and interpret the notion of polarization for such a chain; and prove lower bounds on the probability of polarization. This provides theoretical insight into the use of interacting Gaussian mixture models as a computationally efficient proxy for interacting large language models.

Estimating Weak Spikes by Counting Cycles

Jingming Wang **University of Virginia**

How to estimate weak spikes is a challenging problem. A standard approach is to use empirical eigenvalues, but this method is significantly biased, and the bias is difficult to correct when the noise variances are heterogeneous and unknown. In this talk, I will introduce two new approaches for estimating weak spikes in the spiked Wigner model with general noise variance profiles. The main idea is to use cycle count statistics to estimate moments of the spikes, and then recover the individual spikes from these moments. On the theoretical side, we develop a refined analysis that introduces tools such as a Vandermonde Decomposition Trick, and we show that our methods achieve consistency with algebraic convergence rates. Our numerical studies demonstrate that in many weak signal regimes, the proposed methods significantly outperform existing approaches.

Statistical Inference for Latent Space Models of Network Data with Edge Covariates

Ji Zhu **University of Michigan**

Latent space models (LSMs) provide a powerful framework for analyzing network data by embedding nodes in a latent space. Incorporating covariate information via edge covariates offers an important generalization that strengthens both the interpretability and practical utility of the model. However, we show that coefficient estimates for edge covariate effects obtained through maximum likelihood estimation exhibit asymptotic bias due to high-order geometric effects and errors in latent variable estimation. To address this issue, we propose a plug-in bias-correction estimator that enables asymptotically valid and unbiased statistical inference for the effects of edge covariates. We establish theoretical guaranties, including consistency and asymptotic normality, under various network structures. Extensive simulations and real-world data examples demonstrate that our method effectively reduces estimation bias and improves the accuracy of inference. Our findings contribute to the statistical methodology of LSMs by providing a principled framework for unbiased parameter estimation in network models with edge covariates.

Community Detection in Heterogeneous Signed Networks

Junhui Wang **The Chinese University of Hong Kong**

Activation functions play a central role in neural networks by shaping internal representations. Recently, learning binary activation representations has attracted significant attention due to their advantages in computational and memory efficiency, as well as interpretability. However, training neural networks with Heaviside activations remains challenging, as their non-differentiability obstructs standard gradient-based optimization. In this talk, we propose *Heavy-Tailed Activation Function (HTAF)*, a smooth approximation to the Heaviside function that enables stable training with gradient-based optimization. We construct HTAF as a sigmoid–hyperbolic tangent composite function and theoretically show that it maintains a large gradient mass around zero inputs while exhibiting slower gradient decay in the tail regions. We show empirically that Spiking Neural Networks, Binary Neural Networks and Deep Heaviside neural Networks can be trained stably using HTAF with gradient-based optimization. Finally, we introduce Implicit Concept Bottleneck Models (ICBMs), an interpretable image model that leverages HTAF to induce discrete feature representations. Extensive experiments across various architectures and image datasets demonstrate that ICBM enables stable discretization while achieving prediction performance comparable to or better than standard models.

Discrete Causal Representation Learning

Yuqi Gu **Columbia University**

Causal representation learning seeks to uncover causal relationships among high-level latent variables from low-level, entangled, and noisy observations. Existing approaches often either rely on deep neural networks, which lack interpretability and formal guarantees, or impose restrictive assumptions like linearity, continuous-only observations, and strong structural priors. These limitations particularly challenge applications with a large number of discrete latent variables and mixed-type observations. To address these challenges, we propose discrete causal representation learning (DCRL), a generative framework that models a directed acyclic graph among discrete latent variables, along with a sparse bipartite graph linking latent and observed layers. This design accommodates continuous, count, and binary responses through flexible measurement models while maintaining interpretability. Under mild conditions, we prove that both the bipartite measurement graph and the latent causal graph are identifiable from the observed data distribution alone. We further propose a three-stage estimate-resample-discovery pipeline: penalized estimation of the generative model parameters, resampling of latent configurations from the fitted model, and score-based causal discovery on the resampled latents. We establish the consistency of this procedure, ensuring reliable recovery of the latent causal structure. Empirical studies on educational assessment and synthetic image data demonstrate that DCRL recovers sparse and interpretable latent causal structures.

A Composite Activation Function for Learning Stable Binary Representations

Yongdai Kim **Seoul National University**

Activation functions play a central role in neural networks by shaping internal representations. Recently, learning binary activation representations has attracted significant attention due to their advantages in computational and memory efficiency, as well as interpretability. However, training neural networks with Heaviside activations remains challenging, as their non-differentiability obstructs standard gradient-based optimization. In this talk, we propose *Heavy-Tailed Activation Function (HTAF)*, a smooth approximation to the Heaviside function that enables stable training with gradient-based optimization. We construct HTAF as a sigmoid–hyperbolic tangent composite function and theoretically show that it maintains a large gradient mass around zero inputs while exhibiting slower gradient decay in the tail regions. We show empirically that Spiking Neural Networks, Binary Neural Networks and Deep Heaviside neural Networks can be trained stably using HTAF with gradient-based optimization. Finally, we introduce Implicit Concept Bottleneck Models (ICBMs), an interpretable image model that leverages HTAF to induce discrete feature representations. Extensive experiments across various architectures and image datasets demonstrate that ICBM enables stable discretization while achieving prediction performance comparable to or better than standard models.

Bridging Theory and Practice: Statistical Inference of Latent Space Models for Networks

Yinqiu He **University of Wisconsin–Madison**

Latent space models have been widely adopted in modeling network data. Developing statistical inference for estimated model parameters enables quantifying associated uncertainty and is pivotal for downstream tasks. Despite recent progress on statistical inference of maximum likelihood estimation, crucial gaps remain between asymptotic theoretical guarantees and practical use. Specifically, how are the oracle maximum likelihood estimators related to the solutions produced by algorithms in practice? Can rigorous guarantees be established for existing algorithms without unnecessary restrictions? To address these fundamental questions, we develop a unified analytical framework that bridges theory and practice of statistical inference for latent space models. First, for the maximum likelihood estimation, we relax the spectral-multiplicity constraint in the existing asymptotic theory to broaden the applicability. Second, we overcome the dependence on unknown true parameters in prior algorithmic analyses by developing novel adaptive criteria and theoretical tools. For the widely used algorithm based on the projected gradient descent and the singular value thresholding, we explicitly connect their outputs to the maximum likelihood estimator without relying on unknown information. Our results provide a solid foundation for practically useful and statistically principled statistical inference in network analysis.

Community Detection on a Randomly Growing Network

Min Xu **Rutgers University**

We study community detection on Markovian random networks outside of the Stochastic Block Model framework. Specifically, we consider a random network growth process which generates K separate preferential attachment trees and connects them with Erdős–Rényi edges, so that each tree represents a community and each node inherits the label of the tree to which it belongs. This model is able to produce many features of real-world networks that are improbable under SBM, such as power law degree distribution and the existence of chains and hubs. Given only the final graph, without any knowledge of the growth process, we seek to recover the unobserved community membership of the nodes. We first prove that it is impossible for any algorithm to consistently recover the community label of all the nodes. However, we design algorithms which are provably able to recover the community labels of subsets of central nodes, for several different notions of node centrality such as arrival time or degree. Our procedure consists of two steps where, in the first step, we classify high degree nodes and then, in the second step, extend the community assignments to the remaining vertices. Numerical experiments and a real data application on a co-authorship network demonstrate the effectiveness of our proposed approach.

Estimating Treatment and Spillover Effects with the Ego-Cluster Experimental Design

Emma Zhang **Emory University**

Network interference occurs when a unit's outcome depends not only on its own treatment but also on the treatments received by connected units in the network. Experimental designs and analysis methods that ignore such interference can yield biased estimators of causal effects. In this talk, we develop a new experimental design for estimating the global treatment effect and spillover effect under a model-based framework and ego-cluster randomization. Under this design, the network is partitioned into a collection of ego-clusters, each consisting of a focal unit (the ego) and its network neighbors (the alters), with randomization conducted at the cluster level. We propose model-based estimators for the global treatment effect and spillover effect and establish their consistency and asymptotic normality, with asymptotic variances determined by the ego-cluster structure. Building on these theoretical results, we introduce an ego-clustering algorithm that sequentially selects egos and assigns alters to minimize asymptotic variances. Simulation studies and two empirical applications demonstrate that the proposed procedure yields accurate inference and efficiency improvements over existing network experimental designs.

Eigenvector Fluctuations and Limit Results for Random Graphs with Infinite Rank Kernels

Minh Tang **North Carolina State University**

This paper systematically studies the behavior of the leading eigenvectors for independent edge undirected random graphs generated from a general latent position model whose link function is possibly infinite rank and also possibly indefinite. We first derive uniform error bounds in the two-to-infinity norm as well as row-wise normal approximations for the leading sample eigenvectors. We then build on these results to tackle two graph inference problems, namely (i) entrywise bounds for graphon estimation and (ii) testing for the equality of latent positions, the latter of which is achieved by proposing a rank-adaptive test statistic that converges in distribution to a weighted sum of independent chi-square random variables under the null hypothesis. Our fine-grained theoretical guarantees and applications differ from the existing literature which primarily considers first order upper bounds and more restrictive low rank or positive semidefinite model assumptions. Further, our results collectively quantify the statistical properties of eigenvector-based spectral embeddings with growing dimensionality for large graphs.

Statistically and Computationally Optimal Estimation and Inference of Common Subspaces

Joshua Agterberg **University of Illinois Urbana-Champaign**

In this talk we investigate the statistical and computational limits for the common subspace model, a model wherein one observes a collection of symmetric low-rank matrices perturbed by noise, where each low-rank matrix shares the same common subspace. First, we propose an estimator based on projected gradient descent initialized via a spectral sum of squared matrices and show that it achieves the optimal sin Theta error under a strong signal-to-noise ratio (SNR) condition, and we further give evidence that this SNR condition is necessary for a polynomial time estimator to exist. Next, we turn to estimation and inference for the sin Theta distance itself, and we show that our estimator achieves an asymptotically Gaussian distribution with a bias term that vanishes under a strong signal requirement. Based on this limiting result we propose confidence intervals and show that they are minimax optimal, though the resulting confidence intervals require knowledge of the SNR. We then turn to designing adaptive confidence intervals for the sin Theta error, and we show that adaptivity is information theoretically impossible unless the SNR is sufficiently strong. Consequently, our results unveil a novel phenomenon: despite the SNR being "above" the computational limit for estimation, adaptive statistical inference may still be information-theoretically impossible.

Tuesday, June 9

Estimating the reciprocal effect by a cancellation trick

Jiashun Jin **Carnegie Mellon University**

The p_1 model plays a fundamental role in modeling directed networks, where the reciprocal effect parameter ρ is of special interest in practice. However, due to nonlinear factors in this model, how to estimate ρ efficiently is a long-standing open problem. We tackle the problem by the cycle count approach. The challenge is, due to the nonlinear factors in the model, for any given type of generalized cycles, the expected count is a complicated function of many parameters in the model, so it is unclear how to use cycle counts to estimate ρ. However, somewhat surprisingly, we discover that, among many types of generalized cycles with the same length, we can carefully pick a pair of them such that in the ratio between the expected cycle counts of the two types, the non-linear factors *cancel out nicely with each other*, and as a result, the ratio equals to $\exp(\rho)$ exactly. The above discovery gives rise to a new estimator for the reciprocal effect. In a broad setting where we allow a wide range of reciprocal effects and a wide variety of network sparsity and degree heterogeneity, we show that our estimator achieves the optimal rates for convergence.

A Multilayer Network Model for Aggregated Relational Data

Maggie Niu **Pennsylvania State University**

The Network Scale-Up Method (NSUM) is a vital tool for estimating the sizes of hard-to-reach populations using Aggregated Relational Data (ARD). Recent advancements in survey design increasingly collect ARD across multiple definitions of tie strength, yielding a complex multilayer network structure. Existing statistical frameworks analyze these layers independently, sacrificing valuable shared information. We propose a latent space approach for the Multilayer ARD that enables principled joint estimation. We establish rigorous identifiability conditions for the shared latent space and mathematically prove that the joint estimator achieves asymptotic efficiency gains and finite sample bias reductions by "borrowing strength" across layers. We apply the framework to real-world multilayer NSUM survey data, demonstrating its practical efficacy in yielding more robust and precise subpopulation estimates.

Adaptive Inference for Stratified Network Effects with an Application to Asymmetric Partisan Polarization

Wen Zhou **New York University**

U.S. legislatures exhibit persistent and evolving partisan polarization, which is often studied using voting networks. Bill cosponsorship networks, which link legislators through jointly supported bills, provide a richer relational alternative that includes a larger set of bills and allows for the study of asymmetric polarization through directed links. However, such networks pose substantial challenges for inference, since directionality renders standard parametric models overly restrictive and unreliable. Statistically, inference on directed network effects can be based on network U-statistics, yet their degeneracy may be indeterminate. These difficulties are further amplified by the bi-strata structure induced by party affiliation: dependence patterns differ across strata, invalidating methods such as cluster-based bootstrap and forcing the coexistence of distinct degeneracy regimes across unequally sized strata. To address these, we show that a broad class of network effect estimators in bi-strata directed networks can be reformulated as two-sample network U-statistics, yielding a unified nonparametric framework. We identify three regimes of network effect estimators: degenerate, partially degenerate, and nondegenerate, derive their asymptotics within a single framework, and propose a data-adaptive variance estimator that yields a unified test with Berry-Esseen bounds in all regimes, providing finite-sample guarantees. Applied to the U.S. Congressional cosponsorship networks from 2003-2023, our method reveals that Republican minorities were more likely to be dissatisfied with Democratic majority proposals than the reverse, and that Democratic majorities drove a novel decline in polarization during the 116th House and 117th Senate by backing minority party proposals.

Scalable Community Detection in Massive Networks via Predictive Assignment

Marianna Pensky **University of Central Florida**

Massive network datasets are becoming increasingly common in scientific applications. Existing community detection methods encounter significant computational challenges for such massive networks due to two reasons. First, the full network needs to be stored and analyzed on a single server, leading to high memory costs. Second, existing methods typically use matrix factorization or iterative optimization using the full network, resulting in high runtimes. We propose a strategy called predictive assignment to enable computationally efficient community detection while ensuring statistical accuracy. The core idea is to avoid large-scale matrix computations by breaking up the task into a smaller matrix computation plus a large number of vector computations that can be carried out in parallel. Under the proposed method, community detection is carried out on a small subgraph to estimate the relevant model parameters. Next, each remaining node is assigned to a community based on these estimates. We prove that predictive assignment achieves strong consistency under the stochastic block model and its degree-corrected version. We also demonstrate the empirical performance of predictive assignment on simulated networks and two large real-world datasets: DBLP (Digital Bibliography Library Project), a computer science bibliographical database, and the Twitch Gamers Social Network.

Vertex Alignment and Changepoint Localization in Network Time Series

Zach Lubberts **University of Virginia**

Existing methodology for changepoint localization in an evolving time series of networks generally relies on accurately prescribed vertex correspondence between network realizations at different times. However, such vertex alignments are often misspecified or even unknown. To understand the impact of vertex misalignment on inference for dynamic networks, we construct two illustrative models for network evolution, each with a similar changepoint, and compare methods for changepoint localization. In one model, vertex misalignment causes comparatively little error, while in the other, it seriously impairs localization in a way that cannot be recovered. We show how misalignment between network realizations at different times weakens their underlying correlation, impeding inference procedures that rely on accurate inference of this quantity. We also discuss the merits of graph matching and optimal transport as potential mechanisms for mitigating errors from misalignment.

A Human-guided AI approach for Symbolic Cumulant Calculation

Yumou Qiu **Peking University**

Cumulants play an important role in probability and statistics, but an explicit formula for high-order cumulants are hard to derive, especially in multi-variate case. To tackle the problem, we develop a human-guided AI (hugAI) approach. First, we outline a clear strategy by developing a graph-based theoretical framework, which provides the most simplified mathematical expression of cumulants in terms of moments. Second, to explicitly compute the coefficients, we divide the task into many small steps, and for each step, we communicate with AI (sometimes for several rounds) by carefully written prompts. Finally, the AI execute the formula with a python code. This way, our hugAI approach is able to execute formulas for cumulants of multivariate random variables at any given order. We have carefully validated the formulas with several approaches to make sure they are correct. As applications, we apply our results to Edgeworth expansion for multivariate random variables. Our study primarily uses GPT but we also investigate a handful other LLMs.

Penalized Network Cross-Validation

Yuhong Yang **Tsinghua University**

In both unipartite and bipartite network modeling, one needs to select the most appropriate model among a number of candidates. For choosing the number of communities in stochastic block models (SBMs), cross-validation (CV) methods have been proposed. Interestingly, they are typically shown to be half-consistent, i.e., the underfitting probability is guaranted to converge to zero. To establish full-consistency in both unipartite and bipartite SBMs, we advocate the use of penalized CV. Importantly, this new CV approach allows the comparison across different network model classes (e.g., DCBM versus graphon).

Optimal Differential Privacy on Networks and Bipartite Networks

Wanjie Wang **National University of Singapore**

Differential privacy for social networks is challenging because standard edge-perturbation mechanisms, such as edge flipping, often destroy the low-rank structure needed for statistical inference. We develop a differentially private framework for community detection in network data that is tailored to spectral methods. The key idea is an efficient propose-test-release procedure based on a nonconvex "good set" defined through spectral regularity conditions, including degree control, eigengap separation, and eigenvector incoherence. When the observed graph lies in this regular region, the algorithm releases a noisy spectral embedding with calibrated sensitivity; otherwise, it safely withholds the informative output. Under the degree-corrected stochastic block model, the proposed edge-DP spectral clustering method achieves strong consistency when the privacy budget is as small as $\epsilon \gg 1/n$, implying essentially no asymptotic statistical cost of privacy in the relevant regime. The paper also proves a matching lower bound showing that no (ϵ, δ)-edge-DP algorithm can be even weakly consistent when $\epsilon \ll 1/n$, establishing optimality of the privacy rate. The framework is further extended to bipartite networks under node-level privacy, with analogous consistency and lower-bound results. Simulations and real-data studies on Flickr and Senate roll-call networks show that the proposed GapPTR procedure substantially improves the privacy–utility tradeoff over edge flipping while preserving interpretable community structure.

Parametric Bootstrap for Fixed Edge-Probability Network Models

Can Le **University of California, Davis**

This paper studies parametric bootstrap methods for network data, with the goal of quantifying the uncertainty of network statistics of interest. While existing network resampling methods primarily focus on count statistics under node-exchangeable graphon models, we consider more general network statistics, including local statistics, under the Chung-Lu model without assuming node exchangeability. We show that the natural network parametric bootstrap, which first estimates the network-generating model and then draws bootstrap samples from the estimated model, generally suffers from bootstrap bias. As a general remedy, we show that a two-level bootstrap procedure provably reduces this bias. This extends the classical idea of the iterative bootstrap to the network setting, where the number of parameters grows with the network size. Moreover, for many network statistics, the second-level bootstrap provides a way to construct confidence intervals with higher accuracy. As a by-product of this analysis, we also obtain a central limit theorem for subgraph counts under the inhomogeneous Erdős–Rényi model, which may be of independent interest.

Stable Motifs in a Rashomon Set of Networks

Georgia Smits **University of Washington**

Network motifs, which summarize how a graph's connections organize into patterns of absent links and shared strengths, are widely used to interpret and compare network data. In finite samples, however, substantively different motif structures often fit the data comparably well, so the single best-fitting motif can change under modest sampling variation, a concern whenever decisions are based on the recovered structure. To address this, we build a Bayesian network motif model, then leverage the Rashomon effect, the phenomenon in which many distinct models explain the data almost equally well, enumerate the set of high-posterior, motif-based representations and report the features that remain stable across it. Simulations and a cyberphysical attack-detection pilot demonstrate that this multi-model perspective improves robustness, calibration of edge detection, and changepoint interpretation for network-valued data.

Trustworthy Biomedical Knowledge Graph Construction and Reasoning with Human-in-the-Loop LLMs

Huimin Cheng **Boston University**

Large language models (LLM) provide a scalable tool for extracting biomedical entities and relations from rapidly expanding scientific literature, but LLM-only systems can produce unsupported entities, noisy relations, and hallucinated explanations. In this talk, I will present a human-in-the-loop LLM framework for constructing evidence-preserving biomedical knowledge graphs from literature. The framework uses LLM-assisted pre-annotation, verifier-prioritized human correction, supervised entity and relation extraction, and provenance-aware KG construction. We further integrate the constructed KG with a GraphRAG framework for evidence-grounded question answering. GraphRAG first detects community structure within the KG to identify semantically coherent subgraphs and supporting evidence, and then uses an LLM to generate responses grounded in the retrieved graph context. By combining structured KG retrieval with LLM reasoning, the framework improves traceability and reduces unsupported or hallucinated outputs. Diabetes–nutrition is used as a motivating case study, but the framework is general and can be extended to broader biomedical domains requiring trustworthy knowledge integration and reasoning.

Tight Clusters Make Specialized Experts

Tan Minh Nguyen **National University of Singapore**

Sparse Mixture-of-Experts (MoE) architectures have emerged as a promising approach to decoupling model capacity from computational cost. At the core of the MoE model is the router, which learns the underlying clustering structure of the input distribution in order to send input tokens to appropriate experts. However, latent clusters may be unidentifiable in high dimension, which causes slow convergence, susceptibility to data contamination, and overall degraded representations as the router is unable to perform appropriate token-expert matching. We examine the router through the lens of clustering optimization and derive optimal feature weights that maximally identify the latent clusters. We use these weights to compute the token-expert routing assignments in an adaptively transformed space that promotes well-separated clusters, which helps identify the best-matched expert for each token. In particular, for each expert cluster, we compute a set of weights that scales features according to whether that expert clusters tightly along that feature. We term this novel router the Adaptive Clustering (AC) router. Our AC router enables the MoE model to obtain three connected benefits: 1) faster convergence, 2) better robustness to data corruption, and 3) overall performance improvement, as experts are specialized in semantically distinct regions of the input space. We empirically demonstrate the advantages of our AC router over baseline routing methods when applied on a variety of MoE backbones for language modeling and image recognition tasks in both clean and corrupted settings.

Wednesday, June 10

Prediction Risk in High-Dimensional Ridge Regression

Shurong Zheng **Northeast Normal University**

This paper studies the limiting behavior of the generation error in high-dimensional ridge regression. We consider a linear independent component model in the proportional asymptotic regime where both the sample size n and the dimension p diverge with a fixed ratio $\gamma = p/n \in (0, \infty)$. Under this framework, we establish a central limit theorem for the empirical generation error and derive an explicit formula for its asymptotic mean and variance. As an application, we derive the sample size and the dimension under a given generation error level. Numerical experiments confirm the theoretical results and illustrate the effectiveness of the proposed scaling rule.

Misspecified Maximum Likelihood Estimation for Non-uniform Group Orbit Recovery

Anderson Ye Zhang **University of Pennsylvania**

We study maximum likelihood estimation (MLE) in the generalized group orbit recovery model, where each observation is generated by applying a random group action and a known, fixed linear operator to an unknown signal, followed by additive noise. This model is motivated by single-particle cryo-electron microscopy (cryo-EM) and can be viewed primarily as a structured continuous Gaussian mixture model. In practice, signal estimation is often performed by marginalizing over the group using a uniform distribution—an assumption that typically does not hold and renders the MLE misspecified. This raises a fundamental question: how does the misspecified MLE perform? We address this question from several angles. First, we show that in the absence of projection, the misspecified population log-likelihood has desired optimization landscape that leads to correct signal recovery. In contrast, when projections are present, the global optimizers of the misspecified likelihood deviate from the true signal, with the magnitude of the bias depending on the noise level. To address this issue, we propose a joint estimation approach tailored to the cryo-EM setting, which parameterizes the unknown distribution of the group elements and estimates both the signal and distribution parameters simultaneously.

Network analysis for trustworthiness and applications

Ngoc Le **Hanoi University of Science and Technology**

In this talk, we present our research on social network analysis for trustworthiness ranking. The two mentioned problems are fake account detection and spam phone detection. Traditional content based method and network based method are also presented.

A Propagation Framework for Network Regression

Chenlei Leng **The Hong Kong Polytechnic University**

We introduce a unified and computationally efficient framework for regression on network data, addressing limitations of existing models that require specialized estimation procedures or impose restrictive decay assumptions. Our Network Propagation Regression (NPR) models outcomes as functions of covariates propagated through network connections, capturing both direct and indirect effects. NPR is estimable via ordinary least squares for continuous outcomes and standard routines for binary, categorical, and time-to-event data, all within a single interpretable framework. We establish consistency and asymptotic normality under weak conditions and develop valid hypothesis tests for the order of network influence. Simulation studies demonstrate that NPR consistently outperforms established approaches, such as the linear-in-means model and regression with network. This is joint work with Yingying Ma.

Community Detection Across Mixing Patterns for Two or More Communities

Hao Chen **University of California, Davis**

Community structure in networks can arise through a variety of mixing patterns, including assortative mixing, disassortative mixing, core-periphery structure, and combinations of these patterns across multiple communities. This talk presents a unified framework for community detection across such settings, built on standardized edge-count statistics and a recursive bi-partitioning strategy. The two-community case serves as the basic building block, providing criteria for distinguishing different mixing structures and selecting an appropriate splitting rule. Building on this foundation, the framework extends to multi-community networks by recursively partitioning subnetworks, allowing different parts of the network to exhibit different forms of organization. Simulation studies and real-data examples illustrate the performance of the proposed approach and its ability to recover interpretable community structures in networks with heterogeneous mixing patterns.

Sparse Multitask Regression with a Task Graph

Xiaodong Li **University of California, Davis**

We study high-dimensional multitask regression when the tasks are connected by a graph. The graph is meant to describe which tasks are similar, but it may be noisy and imperfect. Our estimator combines shared sparsity across predictors with graph fusion across tasks. We give global error bounds when the regression coefficients vary smoothly on the graph or have sparse changes across graph edges. We also study a simple stochastic block model setting for the task graph, where tasks form communities and edges are denser within communities than between communities. In this setting, graph fusion can recover a local pooling effect: each task borrows information mainly from tasks in the same community, while the between-community edges create a leakage term. The theory explains when the method behaves close to an oracle that knows the task communities in advance.

Useful Information

Lectures and **Talks** will be held at the **Laurent Schwartz Lecture Hall** of the **Vietnam Institute for Advanced Study in Mathematics (VIASM)**.

Coffee Break will be at the hallway outside of the Lecture Hall. **Wi-Fi** will be available; the detailed information of Wi-Fi access will be provided during the workshop.

How to get to VIASM?

The address of the Vietnam Institute for Advanced Study in Mathematics (VIASM) is:

161 Huynh Thuc Khang Street, Lang Ward, Hanoi, Vietnam.

Booklet Template

This template originates from `LaTeXTemplates.com` and is based on the original version at: `https://github.com/maximelucas/AMCOS_booklet`