Workshop on Statistical Analysis of Networks

Timetable

Monday 18 Sept		
9:00-9:45	Steffen Lauritzen	Mixed convex exponential families
9:45-10:30	Wanjie Wang	Network-guided covariate selection and downstream applications
10:30-11:00	Coffee break	
11:00-11:45	Yi Yu	Multilayer random dot product graphs: estimation and online change point detection
11:45-12:30	Gesine Reinert	Stein's method for networks: characterisations, goodness-of-fit and synthetic data generation
12:30-14:00	Buffet lunch at Scarman restaurant	
14:00-14:45	Binyan Jiang	A two-way heterogeneity model for dynamic networks
14:45-15:30	Yuan Zhang	Distribution-free matrix prediction under arbitrary missing pattern
15:30-16:00	Coffee break	
16:00-16:45	Ji Zhu	A latent space model for hypergraphs with diversity and heterogeneous popularity
16:45-17:30	Emma Zhang	Network response regressions with applications in neuroimaging
17:30-18:30	Reception	
18:30	Invited dinner at Scarman restaurant	
Tuesday 19 Sept		
9:00-9:45	Catherine Matias	Model-based clustering in simple hypergraphs through a stochastic blockmodel
9:45-10:30	Kayvan Sadeghi	Axiomatization of interventional probability distributions
10:30-11:00	Coffee break	
11:00-11:45	Tracy Ke	Optimal network membership estimation under severe degree heterogeneity
11:45-12:30	Yang Feng	Semiparametric modeling and analysis for longitudinal network data
12:30-14:00	Buffet lunch at Scarman restaurant	
14:00-14:45	Swati Chandna	Nonparametric methods for network data
14:45-15:30	Francois Caron	Asymptotic analysis of statistical estimators related to multigraphex processes under misspecification
15:30-16:00	Coffee break and farewell	

18 Sept

Invited Speaker: Steffen Lauritzen
Title: Mixed convex exponential families
Abstract: The lecture presents the notion of a mixed convex exponential family and an associated mixed dual likelihood estimator. Standard examples involve Gaussian graphical models with restrictions that are partly linear in the covariance and partly linear in the concentration. We show that the mixed dual maximum likelihood estimator has the same asymptotic properties as the maximum likelihood estimator under standard sampling. Examples associated with random network models will be touched upon. The lecture is partly based on S. Lauritzen and P. Zwiernik. "Locally associated graphical models and mixed convex exponential families." Ann. Statist. 50 (5) 3009 - 3038, October 2022. https://doi.org/10.1214/22-AOS2219.

Invited Speaker: Wanjie Wang
Title: Network-guided covariate selection and downstream applications
Abstract: Nowadays, it is frequently seen that the data set often contains network information and covariates. Studies have shown that the covariates will be helpful in uncovering the network structure. The opposite direction should also work, that the network information helps to denoise the covariates and improve the statistical inference.
In this talk, I will present a covariate selection method based on the spectral information of the adjacency matrix. By the eigenvectors, we design a testing statistic of the covariates and select them by Higher-Criticism statistic. We prove the optimality of this method and the effect of it in the regression and clustering problems.

Invited Speaker: Yi Yu
Title: Multilayer random dot product graphs: Estimation and online change point detection
Abstract: We study the multilayer random dot product graph (MRDPG) model, an extension of the random dot product graph to multilayer networks. By modelling a multilayer network as an MRDPG, we deploy a tensor-based method and demonstrate its superiority over existing approaches. Moving to dynamic MRDPGs, we focus on online change point detection problems. At every time point, we observe a realisation from an MRDPG. Across layers, we assume shared common node sets and latent positions but allow for different connectivity matrices. We propose efficient algorithms for both fixed and random latent position cases, minimising detection delay while controlling false alarms. Notably, in the random latent position case, we devise a novel nonparametric change point detection algorithm with a kernel estimator in its core, allowing for the case when the density does not exist, accommodating stochastic block models as special cases. Our theoretical findings are supported by extensive numerical experiments, with the code available online. The paper is available on https://arxiv.org/abs/2306.15286.

Invited Speaker: Gesine Reinert
Title: Stein's method for networks: characterisations, goodness-of-fit, and synthetic data generation
Abstract: To understand the distributions of random networks, characterisations can be derived using Stein's method. This talk details how these characterisations can be put to use for assessing goodness of fit. Moreover, synthetic data are increasingly used in computational statistics and machine learning. Some applications relate to privacy concerns, to data augmentation, and to method development. Synthetic data should reflect the underlying distribution of the real data, being faithful but also showing some variability. This talk addresses tests for goodness of fit

of synthetic data generators using Stein's method. Finally we shall see that ideas from Stein's method can even be used to generate synthetic network data.
This talk is based on papers with Nathan Ross and with Wenkai Xu.

Invited Speaker: Binyan Jiang
Title: A two-way heterogeneity model for dynamic networks
Abstract: Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. This paper proposes a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With observed networks each having nodes, we develop a new asymptotic theory for the maximum likelihood estimation of parameters. We overcome the global non-convexity of the negative log-likelihood function by the virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and a data analysis examining social interactions of ants.

Invited Speaker: Yuan Zhang
Title: Distribution-free matrix prediction under arbitrary missing pattern
Abstract: This paper studies the open problem of conformalized entry prediction in a row/column-exchangeable matrix. The matrix setting presents novel and unique challenges, but there exists little work on this interesting topic. We meticulously define the problem, differentiate it from closely related problems, and rigorously delineate the boundary between achievable and impossible goals. We then propose two practical algorithms. The first method provides a fast emulation of the full conformal prediction, while the second method leverages the technique of algorithmic stability for acceleration. Both methods are computationally efficient and can effectively safeguard coverage validity in presence of arbitrary missing pattern. Further, we quantify the impact of missingness on prediction accuracy and establish fundamental limit results. Empirical evidence from synthetic and real-world data sets corroborates the superior performance of our proposed methods.
This is a joint work with Meijia Shao.

Invited Speaker: Ji Zhu
Title: A latent space model for hypergraphs with diversity and heterogeneous popularity
Abstract: While relations among individuals make an important part of data with scientific and business interests, existing statistical modeling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. This work addresses the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, we propose a new latent space model for hypergraphs using determinantal point processes, which is driven by the diversity within hyperedges and each node's popularity. This model mechanism is in contrast to existing hypergraph models, which are predominantly driven by similarity rather than diversity. Additionally, the proposed model accommodates broad types of hypergraphs, with no restriction on the cardinality and multiplicity of hyperedges. Consistency and asymptotic normality of the maximum likelihood estimates of the model parameters have been established. Simulation studies and an application to the What's Cooking data show the

effectiveness of the proposed model.

Invited Speaker: Emma Zhang
Title: Network response regressions with applications in neuroimaging
Abstract: Recent advances in data collection technology have multiplied the availability of network data, leading to not only larger and more complex networks but also to instances where independent network samples are collected. In such data sets, a network serves as the basic data object and they are increasingly common in neuroscience and genetics. When analyzing these data, a fundamental scientific question of interest is to understand how the subject-level network connectivity changes as a function of clinical characteristics. In this talk, we propose a new network response model framework, in which the networks are treated as responses and the network-level covariates as predictors. Under the proposed framework, we discuss model identifiability, estimation and theoretical properties. Finally, we present our findings from the analyses of three resting-state and task-related neuroimaging studies.

19 Sept

Invited Speaker: Catherine Matias
Title: Model-based clustering in simple hypergraphs through a stochastic blockmodel
Abstract: We propose a new model to address the overlooked problem of node clustering in simple hypergraphs. Simple hypergraphs are suitable when a node may not appear multiple times in the same hyperedge, such as in co-authorship datasets. Our model assumes the existence of latent node groups and hyperedges are conditionally independent given these groups. We first establish the generic identifiability of the model parameters. We then develop a variational approximation Expectation-Maximization algorithm for parameter inference and node clustering, and derive a statistical criterion for model selection. To illustrate the performance of our R package HyperSBM, we compare it with other node clustering methods using synthetic data generated from the model, as well as from a line clustering experiment and a co-authorship dataset. As a by-product, our synthetic experiments demonstrate that the detectability thresholds for non-uniform sparse hypergraphs cannot be deduced from the uniform case.
This is a joint work with Luca Brusa.

Invited Speaker: Kayvan Sadeghi
Title: Axiomatization of interventional probability distributions
Abstract: Causal intervention is an essential tool in causal inference. It is axiomatized under the rules of do-calculus in the case of structure causal models. We provide simple axiomatizations for families of probability distributions to be different types of interventional distributions. Our axiomatizations neatly lead to a simple and clear theory of causality that has several advantages: it does not need to make use of any modeling assumptions such as those imposed by structural causal models; it only relies on interventions on single variables; it includes most cases with latent variables and causal cycles; and more importantly, it does not assume the existence of an underlying true causal graph–in fact, a causal graph is a by-product of our theory. We show that, under our axiomatizations, the intervened distributions are Markovian to the defined intervened causal graphs, and an observed joint probability distribution is Markovian to the obtained causal graph; these results are consistent with the case of structural causal models, and as a result, the existing theory of causal inference applies. We also show that a large class of natural structural causal models satisfy the theory presented here. We also show how these results can be specialised for interventions on networks. This is joint work with Terry Soo.

Invited Speaker: Tracy Ke
Title: Optimal network membership estimation under severe degree heterogeneity
Abstract: Real networks often have severe degree heterogeneity, with maximum, average, and minimum node degrees differing significantly. This paper examines the impact of degree heterogeneity on statistical limits of network data analysis. Introducing the empirical heterogeneity distribution (EHD) under a degree-corrected mixed membership model, we demonstrate that the optimal rate of mixed membership estimation is directly linked to EHD. Surprisingly, severe degree heterogeneity can decelerate the error rate, even when the overall sparsity remains unchanged.
To develop a spectral algorithm that is rate-optimal for arbitrary EHD, we propose the "two normalizations" approach to enhance the performance of an existing spectral algorithm, Mixed-SCORE. Our approach involves a pre-PCA normalization parameterized by $b \in R$. The crucial question is how to choose b to simultaneously optimize the signal-to-noise ratios for all entries of leading empirical eigenvectors. Remarkably, we find that universally choosing $b = 1/2$ yields

favorable results. The resulting algorithm is rate-optimal for networks with arbitrary degree heterogeneity.
Our findings have two significant implications: (1) Degree heterogeneity indeed influences the fundamental statistical limits; and (2) Achieving optimal adaptivity to degree heterogeneity is possible, provided that the algorithm is thoughtfully designed.

Invited Speaker: Yang Feng
Title: Semiparametric modeling and analysis for longitudinal network data
Abstract: We introduce a semiparametric latent space model for analyzing longitudinal network data. The model consists of a static latent space component and a time-varying node-specific baseline component. We develop a semiparametric efficient score equation for the latent space parameter by adjusting for the baseline nuisance component. Estimation is accomplished through a one-step Newton-Raphson update and an appropriately penalized log-likelihood function. We derive oracle error bounds for the estimators and address identifiability concerns from a quotient manifold perspective. Our approach is demonstrated using the New York Citi Bike Dataset.

Invited Speaker: Swati Chandra
Title: Nonparametric methods for network data
Abstract: Network data are commonly observed in a wide variety of applications. Such data may arise in the form of a single network observed at a given point in time, or as multiple networks on the same set of nodes, for example, social networks on the same set of individuals over time, or from different sources at a given point in time (e.g., structural brain networks from different subjects). A nonparametric approach to studying structure in an unlabeled network is offered by the graphon function. There has been a growing interest on the problem of graphon estimation as well as its application to important problems such as bootstrapping networks, estimation of missing links etc. In this talk, I will present results on graphon estimation from a single network observed with node covariates and a natural extension of the graphon model to the bivariate setting where a pair of possibly correlated networks on the same set of nodes are observed.

Invited Speaker: Francois Caron
Title: Asymptotic analysis of statistical estimators related to multigraphex processes under misspecification
Abstract: We consider here the asymptotic properties of Bayesian or frequentist estimators of a vector of parameters related to structural properties of sequences of graphs. The estimators studied originate from a particular class of graphex model introduced by Caron and Fox (2017). The analysis is however performed here under very weak assumptions on the underlying data generating process, which may be different from the model of Caron and Fox or from a graphex model. In particular, we consider generic sparse graph models, with unbounded degree, whose degree distribution satisfies some assumptions. We show that one can relate the limit of the estimator of one of the parameters to the sparsity constant of the true graph generating process. When taking a Bayesian approach, we also show that the posterior distribution is asymptotically normal. We discuss situations where classical random graphs models satisfy our assumptions.
This is joint work with Zacharie Naulet and Judith Rousseau, paper available on https://arxiv.org/abs/2107.01120.