Turing Workshop on Statistics of Network Analysis

Day 1: 29 May

9:30-10:00 Registration & Coffee

10:00-10:45 Eric Kolaczyk

Title: On the Propagation of Uncertainty in Network Summaries

Abstract: While it is common practice in applied network analysis to report various standard network summary statistics, these numbers are rarely accompanied by some quantification of uncertainty. Yet any error inherent in the measurements underlying the construction of the network, or in the network construction procedure itself, necessarily must propagate to any summary statistics reported. I summarize results from work on the problem of estimating the density of edges in a noisy network, as a canonical prototype of the more general problem of estimating density of arbitrary subgraphs. Under a simple model of network error, we show that consistent estimation of such densities is impossible when the rates of error are unknown and only a single network is observed. We then develop method-of-moment estimators of network edge density and error rates for the case where a minimal number of network replicates are available. These estimators are shown to be asymptotically normal as the number of vertices increases to infinity. We also provide the confidence intervals for quantifying the uncertainty in these estimates based on the asymptotic normality. We illustrate the use of our estimators in the context of gene coexpression networks. This is joint work with Qiwei Yao and Jinyuan Chang.

10:45-11:30 Sofia Olhede

Title: Nonparametric network summaries

Abstract: Nonparametrics for networks have attracted significant attention. I will discuss basic desiderata for constructing network summaries, and how we can use nonparametric summaries to characterise network structure. Key to understanding is being able to specify properties of scale, and also multiple summaries of structure.

11:30-12:15 Mingli Chen

Title: Sparse beta models

Abstract: We propose the sparse beta-model as a model to interpolate the classical Erdos-Renyi model well suited for studying sparse networks and the more recent beta model which is tractable for modelling relatively dense networks. We propose to use constrained minimization to estimate its parameters and study the properties of the resulting estimate.

12:15-13:45 Lunch

13:45-14:30 Bryan Graham

Title: Homophily and transitivity in dynamic network formation

Abstract: In social and economic networks linked agents often share additional links in common. There are two competing explanations for this phenomenon. First, agents may have a structural taste for transitive links – the returns to linking may be higher if two agents share links in common. Second, agents may assortatively match on unobserved attributes, a process called homophily. I study parameter identifiability in a simple model of dynamic network formation with both effects. Agents form, maintain, and sever links over time in order to maximize utility. The return to linking may be higher if agents share friends in common. A pair-specific utility component allows for arbitrary homophily on time-invariant agent attributes. I derive conditions under which it is possible to detect the presence of a taste for transitivity in the presence of assortative matching on unobservables. I leave the joint distribution of the initial network and the pair-specific utility component, a very high dimensional object, unrestricted. The analysis is of the 'fixed effects' type. The identification result is constructive, suggesting an analog estimator, whose single large network properties I characterize.

14:30-15:15 Michael Leung

Title: Normal Approximation in Strategic Network Formation

Abstract: We prove a CLT for network statistics and apply it to static and dynamic models of network formation with strategic interactions.

15:15-16:00 Koen Jochmans

Title: Fixed-effect regressions on network data

Abstract: This paper studies inference on fixed effects in a linear regression model estimated from network data. An important special case of our setup is the two-way regression model, which is a workhorse method in the analysis of matched data sets. Networks are typically quite sparse and it is difficult to see how the data carry information about certain parameters. We derive bounds on the variance of the fixed-effect estimator that uncover the importance of the structure of the network. These bounds depend on the smallest non-zero eigenvalue of the (normalized) Laplacian of the network and on the degree structure of the network. The Laplacian is a matrix that describes the network and its smallest non-zero eigenvalue is a measure of connectivity, with smaller values indicating less-connected networks. These bounds yield conditions for consistent estimation as well as convergence rates, and allow to evaluate the accuracy of first-order approximations to the variance of the fixed-effect estimator. The bounds can also be used to assess the bias and variance of estimators of moments of the fixed effects.

16:00-16:30 Coffee

16:30-17:15 Yuguo Chen

Title: A blockmodel for node popularity in networks with community structure

Abstract: The community structure observed in empirical networks has been of particular interest in the statistics literature, with a strong emphasis on the study of blockmodels. We study an important network feature called node popularity, which is closely associated with community structure. Neither the classical stochastic blockmodel nor its degree-corrected extension can satisfactorily capture the dynamics of node popularity as observed in empirical networks. We propose a popularity-adjusted blockmodel for flexible and realistic modeling of node popularity. We establish consistency of likelihood modularity for community detection as well as estimation of node popularities and model parameters, and demonstrate the advantages of the new modularity over the degree-corrected blockmodel modularity in simulations. By analyzing the political blogs network, the British MP network, and the DBLP bibliographical network, we illustrate that improved empirical insights can be gained through this methodology.

17:15-18:00 David Choi

Title: Global Spectral Clustering in Dynamic Networks

Abstract: Networks are often dynamic, and it is of substantial interest to visualize and model their evolution over time. We present a novel method for spectral clustering in dynamic networks, with a smoothing penalty to promote similarity across time periods. We prove that the resulting optimization problem, while non-convex, can be solved globally when the smoothing penalty parameter is not too large. The optimization criteria can be shown to apply an “adaptive” level of smoothing; in particular, smoothing is automatically suppressed at change points in the data.

18:30 Workshop dinner (for speakers and organisers)

Day 2: 30 May

9:45-10:30 Zongming Ma

Title: Optimal hypothesis testing for stochastic block models with growing degrees

Abstract: In this talk, we discuss optimal hypothesis testing for distinguishing a stochastic block model from an Erdos-Renyi random graph. We derive central limit theorems for a collection of linear spectral statistics under both the null and local alternatives. In addition, we show that linear spectral statistics based on Chebyshev polynomials can be used to approximate signed cycles of growing lengths which in turn determine the likelihood ratio test asymptotically when the graph size and the average degree grow to infinity together. Therefore, one achieves sharp asymptotic optimal power of the testing problem within polynomial time complexity provided that the average degree grows sufficiently fast.

10:30-11:15 Chao Gao

Title: Testing for Global Network Structure Using Small Subgraph Statistics

Abstract: We study the problem of testing for community structure in networks using relations between the observed frequencies of small subgraphs. We propose a simple test for the existence of communities based only on the frequencies of three-node subgraphs. The test statistic is shown to be asymptotically normal under a null assumption of no community structure, and to have power approaching one under a composite alternative hypothesis of a degree-corrected stochastic block model. We also derive a version of the test that applies to multivariate Gaussian data. Our approach achieves near-optimal detection rates for the presence of community structure, in regimes where the signal-to-noise is too weak to explicitly estimate the communities themselves, using existing computationally efficient algorithms. We demonstrate how the method can be effective for detecting structure in social networks, citation networks for scientific articles, and correlations of stock returns between companies on the S&P 500.

11:15-11:45 Coffee break

11:45-12:30 Sonja Petrovic

Title: Goodness-of-fit tests for 3 SBM variants

Abstract: Stochastic block models (SBM) with unknown block structure are widely used in analysis of real-world network data. Testing goodness-of-fit of such models is an important practical question. We develop finite-sample goodness-of-fit tests for three different variants of SBMs with unknown block assignments. The main building block for the goodness-of-fit test is an exact test for SBM with observed block assignment, which is implemented using tools from algebraic statistics. The methodology extends to any mixture of log-linear models on discrete data.

12:30-13:45 Lunch

13:45-14:30 Yang Feng

Title: Likelihood Ratio Test for Stochastic Block Models with Bounded Degrees

Abstract: A fundamental problem in network data analysis is to test whether a network contains statistical significant communities. We study this problem in the stochastic block model context by testing H0: Erdos-Renyi model vs. H1: stochastic block model. This problem serves as the foundation for many other problems including the testing-based methods for determining the number of communities and community detection. Existing work has been focusing on growing-degree regime while leaving the bounded-degree case untreated. Here, we propose a likelihood ratio type procedure based on regularization to test stochastic block models with bounded degrees. We derive the limiting distributions as power Poisson laws under both null and alternative hypotheses, based on which the limiting power of the test is carefully analyzed. The joint impact of signal-to-noise ratio and the number of communities on the asymptotic results is also unveiled. The proposed procedures are examined by both simulated and real-world network datasets. Our proofs depend on the contiguity theory for random regular graphs developed by Janson (1995).

14:30-15:15 Ginestra Bianconi

Title: Multilayer networks: Structure and Function

Abstract: Multilayer networks describe interacting complex systems formed by different interacting networks. Multilayer networks are ubiquitous and include social networks, financial markets, multimodal transportation systems, infrastructures, the brain and the cell. Multilayer networks cannot be reduced to a large single network. In this talk I will present recent results showing how we can extract from multilayer networks more relevant information than from its single layer taken in isolation. Secondly I will provide evidence that dynamical processes on multilayer networks can display very novel properties that reflect the rich interplay between structure and multiplexity.

15:15-16:00 Sharmodeep Bhattacharyya

Title: Dynamic Community Detection for Multiple Networks

Abstract: Multiple networks are currently becoming more common among network data sets. Usually, a number of network data sets, which share some form of connection between each other are known as multiple or multi-layer networks. We consider the problem of identifying the common and dynamic community structures for multiple networks. We also extend the existing nonparametric latent variable models in the context of multiple networks, and thereby propose a class of network models for multiple networks. We consider extensions of the spectral clustering methods for the multiple network models, and give theoretical guarantee that the spectral clustering methods produce consistent community detection in case of both multiple stochastic block model and multiple degree-corrected block models. The methods are shown to work under sufficiently mild conditions on the number of multiple networks to detect associative, dissociative and mixed

community structures, even if all the individual networks are very sparse and most of the individual networks are below community detectability threshold. We reinforce the validity of the theoretical results via simulations too.

16:00-16:30 Coffee break

16:30-17:15 Vishesh Karwa

Title: Sharing Social Network Data: Differentially Private Estimation of Exponential-Family Random Graph Models

Abstract: Motivated by a real-life problem of sharing social network data that contain sensitive personal information, we propose a novel approach to release and analyze synthetic graphs in order to protect privacy of individual relationships captured by the social network while maintaining the validity of statistical results. A case study using a version of the Enron e-mail corpus dataset demonstrates the application and usefulness of the proposed techniques in solving the challenging problem of maintaining privacy and supporting open access to network data to ensure reproducibility of existing studies and discovering new scientific insights that can be obtained by analyzing such data. We use a simple yet effective randomized response mechanism to generate synthetic networks under ε-edge differential privacy, and then use likelihood based inference for missing data and Markov chain Monte Carlo techniques to fit exponential-family random graph models to the generated synthetic networks.

17:15-18:00 Michael Schweinberger

Title: Concentration and consistency results for random graph models with transitivity

Abstract: Statistical inference for random graphs with transitivity and other features of random graphs inducing complex dependence is challenging. We stress the importance of additional structure and show that additional structure facilitates statistical inference. A simple and popular form of additional structure is a random graph with neighborhoods and local dependence within neighborhoods. Such structure is observed in a number of applications and is popular in network science, examples being school classes within schools, departments within companies, and units of armed forces. We develop the first concentration results for maximum likelihood and M-estimators of a wide range of canonical and curved exponential-family random graphs with local dependence. All results are non-asymptotic and cover random graphs of fixed and finite size, provided the neighborhoods are small relative to the size of the random graph. We discuss extensions to larger random graphs with more neighborhoods along with concentration results for subgraph-to-graph estimators. As applications, we consider canonical and curved exponential-family random graphs, with local dependence induced by sensible forms of transitivity and parameter vectors whose dimensions depend on the number of nodes.

18:00 Finish