Titles and Abstracts Cyber Workshop

Haeran Cho

Model selection in change-point problems

Lately, there has been a surge of interest in research for computationally fast and statistically efficient methods for change-point detection, as nonstationarities frequently observed in real-life datasets are often attributed to structural breaks in the underlying stochastic properties. In multiple change-point detection, model selection via estimating the total number of change-points poses as a challenge, particularly when the dimensionality of the data is large. In this talk, I will address the model selection in change-point problems in two different settings: when $p = 1$ where one can benefit from localised application of an information criterion, and when p is large where the change-point detection problem can be translated to that of detecting pervasive and latent 'factors'.

Nick Heard

Bayesian changepoint detection in cyber-security

Changepoint models have numerous applications within statistical cyber-security: Examples include determining the genealogy of a malware executable; providing flexible models for densities and intensities of normal computer network traffic; and, of course, network anomaly detection to potentially reveal the presence of a network intrusion. This talk will review Bayesian changepoint inferential tools, and demonstrate how these tools can be applied and adapted in the cyber context.

Kathryn Leeming

The Generalised Network AutoRegression Model for Network Time Series

(Joint work with M. I. Knight, G. P. Nason and M. A. Nunes)

In this talk, the Generalised Network AutoRegression (GNAR) model will be presented, a model designed for use on time series recorded at nodes of a network. This model encapsulates autoregressive behaviour both within and between series at nodes of the network, using the network structure via neighbour sets. Useful features of our GNAR model will be discussed, such as sparsity, flexibility to different networks, and its potential for use on data with missing observations. We will illustrate the model on an application involving prediction of a multivariate time series, including finding a useful network for the data to aid prediction.

Carey Priebe

On a two-truths phenomenon in spectral graph clustering

Clustering is a many-splendored thing. As the ill-defined cousin of classification, in which the observation to be classified X comes with a true but unobserved class label Y, clustering is concerned with coherently grouping observations without any explicit concept of true groupings. Spectral graph clustering – clustering the vertices of a graph based on their spectral embedding – is all the rage, and recent theoretical results provide new understanding of the problem and solutions. In particular, we reset the field of spectral graph clustering, demonstrating that spectral graph clustering should not be thought of as kmeans clustering composed with Laplacian spectral embedding, but rather

Gaussian mixture model (GMM) clustering composed with either Laplacian or Adjacency spectral embedding (LSE or ASE); in the context of the stochastic blockmodel (SBM), we use eigenvector CLTs & Chernoff analysis to show that (1) GMM dominates kmeans and (2) neither LSE nor ASE dominates, and we present an LSE vs ASE characterization in terms of affinity vs core-periphery SBMs. Along the way, we describe our recent asymptotic efficiency results, as well as an interesting twist on the eigenvector CLT when the block connectivity probability matrix is not positive semidefinite. (And, time permitting, we will touch on essential results using the matrix two-to-infinity norm.) We conclude with a 'Two Truths' LSE vs ASE spectral graph clustering result – necessarily including model selection for both embedding dimension & number of clusters – convincingly illustrated via an exciting new diffusion MRI connectome data set: different embedding methods yield different clustering results, with one (ASE) capturing gray matter/white matter separation and the other (LSE) capturing left hemisphere/right hemisphere characterization.

Gesine Reinert

Detecting financial fraud is a global challenge. This talk will mainly focus on financial transaction networks. In such networks, examples of anomalies are long paths of large transaction amounts, rings of large payments, and cliques of accounts. There are many methods available to detect specific anomalies. Our aim is to detect unknown anomalies. To that purpose we use a strategy with derives features from network comparison methods and spectral analysis, and then apply a random forest method to classify nodes as normal or anomalous. We test the method on synthetic data which we generated, and then on synthetic data without us having had access to the ground truth.

This talk is based on joint work with Andrew Elliott and Mihai Cucuringu, (Alan Turing Institute) as well as Milton Martinez Luaces and Paul Reidy (Accenture).

https://www.pnas.org/content/pnas/early/2019/03/07/1814462116.full.pdf

Patrick Rubin-Delanchy

A statistical interpretation of spectral embedding: the generalised random dot product graph

A generalisation of a latent position network model known as the random dot product graph model is considered. The resulting model may be of independent interest because it has the unique property of representing a mixture of connectivity behaviours as the corresponding convex combination in latent space. We show that, whether the normalised Laplacian or adjacency matrix is used, the vector representations of nodes obtained by spectral embedding provide strongly consistent latent position estimates with asymptotically Gaussian error. Direct methodological consequences follow from the observation that the well-known mixed membership and standard stochastic block models are special cases where the latent positions live respectively inside or on the vertices of a simplex. Estimation via spectral embedding can therefore be achieved by respectively estimating this simplicial support, or fitting a Gaussian mixture model. In the latter case, the use of k-means, as has been previously recommended, is suboptimal and for identifiability reasons unsound. Empirical improvements in link prediction, as well as the potential to uncover much richer latent structure (than available under the mixed membership or standard stochastic block models) are demonstrated in a cyber-security example.

Francesco Sanna Passino

Some ideas on Bayesian modelling of networks for cyber-security applications

This talk discusses two Bayesian models for network data, aimed at two common tasks: link prediction and community detection. In the first part of the talk, extensions to the well-known Poisson matrix factorisation (PMF) model are presented. Poisson matrix factorisation is a popular model for link prediction in large networks, particularly useful for its scalability. In this talk, PMF is extended to include scenarios that are commonly encountered in computer networks. In particular, an extension is proposed to explicitly include known covariates associated with the nodes. Furthermore, a doubly sparse PMF with Indian Buffet Process priors, which further refines the edge probabilities, is introduced. Finally, a seasonal version of PMF is presented to handle dynamic networks. Fast inference schemes are discussed. The extensions of the PMF model are tested and applied on the user-authentication dataset publicly released by the Los Alamos National Laboratory. Results show improved performance over the standard PMF model and other common link prediction techniques.

The second part of the talk discusses a model for Bayesian community detection based on spectral embedding. Spectral embedding of adjacency or Laplacian matrices of undirected graphs is a common technique for representing the network in a lower dimensional latent space, with optimal theoretical guarantees. The embedding can be used to estimate the community structure of the network, with strong consistency results in the stochastic blockmodel framework. One of the main limitations of standard algorithms for community detection from spectral embeddings is that the number of communities and the latent dimension of the embedding must be specified in advance. A Bayesian model for simultaneous and automatic selection of the appropriate dimension of the latent space and the number of blocks is discussed in this talk. The model is tested on simulated and real world datasets, showing promising performance for recovering the latent community structure of networks.

Vinesh Solanki

TDA and the geometry of spectral embedding

Spectral embedding of a network results in a point cloud of latent positions, the distribution of which can often reveal important underlying structure. For example, under the stochastic block model, the point cloud separates into clusters which represent network communities and under the mixed membership stochastic block model, the point cloud concentrates around a simplex. The distributions of latent positions in both cases exhibit distinct geometric structure. In this talk, I will explain how the tools of topological data analysis can be used to describe this structure. A substantial part of this talk will consist of an introduction to the methods of TDA. This is joint work with Patrick Rubin-Delanchy and Ian Gallagher.

Yi Yu

Optimal change point detection and localisation

This talk will be a combination of 4 papers. I will start from the best-studied univariate mean change point detection problem [1], which serves as a blueprint for more complex settings, and then talk about high-dimensional covariance [2], sparse dynamic networks [3] and distribution functions (ongoing project) change point detection problems. I will show the phase transitions, minimax

optimality and binary segmentation type algorithms in all settings. To conclude the talk, I will talk about the penalisation techniques used in the change point detection problem and the limitations thereof.

[1] https://arxiv.org/abs/1810.09498

[2] https://arxiv.org/abs/1712.09912

[3] https://arxiv.org/abs/1809.09602