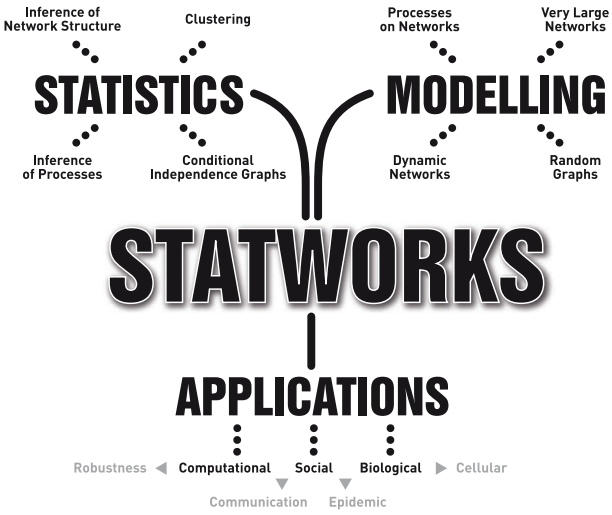
ABSTRACT BOOKLET



Statistical modelling and inference for networks
University of Bristol, UK
28th June – 1st July 2010

University of BRISTOL

SuSTaIn

Abstracts for invited speakers

 Stanley Wasserman *Indiana University*

Statistical models for networks: The past, present, and future

Monday 09:30–10:30

- Network science focuses on relationships between social entities. It is used widely in the social and behavioral sciences, as well as in political science, economics, organizational science, and industrial engineering. The social network perspective has been developed over the last seventy years by researchers in psychology, sociology, and anthropology, and more recently, in physics and computer science. The paradigm requires a new and different set of concepts and analytic tools, beyond those provided by standard statistical methods. Statistical/distributional approaches to networks have existed for many years. These concepts and tools are the topics of this talk – we focus on the past, review the present, and comment on the future.

 Michael Stumpf *Imperial College London*

To be announced

Monday 15:30–16:30

-

 Eric Kolaczyk *Boston University*

Drug Target Prediction: Finding Biological Needles in a Haystack of Networks

Monday 16:30–17:30

- Understanding the mechanism of action of putative drug compounds and locating potential genetic drug targets is a major focus in biomedicine. Compendia of mRNA microarray expression data, generated under various experimental conditions, and fairly readily available, can be a key source of information for this problem. However, there are a host of challenges that lie between the gathering of such data and the successful characterization of drug mechanism of action. In this talk, I will present work by our

group on one of these challenges: the problem of separating in expression data the response of gene targets to experimental perturbations from the background of typical cell activity. We pose the problem as one of extracting a sparse signal from a 'background' of an association network of gene correlations – that is, we pose a network filtering problem. Modeling the acquisition of network data, including the potential presence of targets, using a system of sparse simultaneous equation models (SSEMs), detection is approached as a two-step procedure, involving (i) statistical inference and removal of 'background' network structure, using tools of sparse inference, and (ii) outlier detection in the network-filtered residuals. I will discuss both the theoretical and practical capabilities of this approach, with the latter including applications to data in yeast and Ecoli.

Sanjeev Goyal *University of Cambridge*

Strategic Network Formation

Tuesday 09:00–10:00

•

Stephane Robin *AgroParisTech*

Uncovering structure in biological interaction networks

Tuesday 13:00–14:00

- Networks are a natural way to describe the interaction between biological entities (genes, species, etc). Flexible random graphs models are needed to describe the wide variety to topologies displayed by such networks. Stochastic block models have become a popular way to describe such topologies. These models are mixture models, the inference of which raises several statistical issues. We will present several advances such as some refinement of variational approximations, a variational Bayes inference or the introduction of covariates to explain the network structure. These results will be illustrated with application to molecular and ecological networks.

Geoffrey West *Santa Fe Institute*

Universal Scaling Laws, Network Structures, Sustainability and the Pace of Life from Cells and Ecosystems to Cities and Corporations

Wednesday 09:00–10:00

- Despite its extraordinary complexity and diversity, many of Life's most fundamental and complex phenomena scale with size in a surprisingly simple fashion. For example, metabolic rate scales approximately as the 3/4-power of mass over 27 orders of magnitude from complex molecules up to the largest multicellular organisms. Similarly, time-scales (such as lifespans and growth-rates) and sizes (such as genome lengths, RNA densities, and tree heights) scale as power laws with exponents which are typically simple multiples of 1/4. This universality and simplicity suggests that fundamental constraints underly much of the coarse-grained generic structure and organisation of living systems. It will be shown how these 1/4 power scaling laws follow from underlying principles embedded in the dynamics and geometry of space-filling, fractal-like, branching networks, presumed optimised by natural selection. These ideas lead to a general quantitative, predictive framework that potentially captures many essential features of diverse biological systems. Examples will include vascular systems, growth, cancer, aging and mortality, sleep, cell size, and evolutionary rates. These ideas will be extended to social organisations: to what extent are cities or corporations "just" very large organisms? Analogous scaling laws reflecting underlying social network structures point to general principles of organization common to all cities, but, counter to biology, the pace of social life systematically increases with size. This has dramatic implications for growth, development and sustainability: innovation and wealth creation that fuel social systems, if left unchecked, potentially sow the seeds for their inevitable collapse.

Sean Meyn *University of Illinois*

The Value of Volatile Resources in Electricity Markets

Wednesday 15:00–16:00

- While renewable resources most certainly provide environmental benefits, and also help to meet aggressive renewable energy targets, their deployment has pronounced impacts on system operations. There is an acute need to understand these impacts in order to fully harness the benefits of renewable resource integration. In this paper we focus on the integration of wind energy resources in a multi-settlement electricity market structure. We study the dynamic competitive equilibrium for a stochastic market model and obtain closed form expressions for the supplier and consumer surpluses. Numerical results based on these formulae show that the value of wind generation to consumers falls dramatically with volatility. In fact, we can establish thresholds for the coefficient

of variation beyond which the value of wind is questionable. These findings can help guide the integration of renewables in future electricity markets.

Sean Meyn, Matias Negrete-Pincetic, Gui Wang, Anupama Kowli, and Ehsan Shafieepoorfard.

David Barber *University College London*

Finding graph clusters using clique matrices

Thursday 11:00–12:00

- Finding clusters of well-connected nodes in a graph is a problem common to many domains, including social networks, the Internet and bioinformatics. From a computational viewpoint, finding these clusters or graph communities is a difficult problem. We use a clique matrix decomposition based on a statistical description that encourages clusters to be well connected and few in number. The formal intractability of inferring the clusters is addressed using a variational approximation inspired by mean-field theories in statistical mechanics. Clique matrices also play a natural role in parametrizing positive definite matrices under zero constraints on elements of the matrix.

Brendan Murphy *University College Dublin*

A mixture of experts latent position cluster model for social network data

Thursday 12:00–13:00

- Social network data represent the interactions between a group of social actors. Interactions between colleagues and friendship networks are typical examples of such data.

  The latent space model for social network data locates each actor in a network in a latent (social) space and models the probability of an interaction between two actors as a function of their locations. The latent position cluster model extends the latent space model to deal with network data in which clusters of actors exist - actor locations are drawn from a finite mixture model, each component of which represents a cluster of actors.

  A mixture of experts model builds on the structure of a mixture model by taking account of both observations and associated covariates when modeling a heterogeneous population. Herein, a mixture of experts extension of the latent position cluster model is developed. The mixture of experts framework allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations.

Estimates of the model parameters are derived in a Bayesian framework using a Markov Chain Monte Carlo algorithm. The algorithm is generally computationally expensive - surrogate proposal distributions which shadow the target distributions are derived, reducing the computational burden. Variational Bayesian inference for the model will also be discussed.

This is joint work with Claire Gormley and Michael Salter-Townshend.

Abstracts for contributed speakers

FORD, Ashley *Warwick University*

Statistically Equivalent Graphs and Product Space Representations

Monday 11:00–11:30

- A common approach to the study of stochastic processes on graphs is to select a family of graphs and determine features of the process on a randomly selected graph. Each of these families occupies a small part of the space of possible graphs and there is a risk of unjustified extrapolation of results from one family to all graphs. Another difficulty with this approach is the complexity of separating the effects of the random choice of graph from the stochastic process running on the graph.

  Building on standard statistical theory we develop a framework for defining statistical equivalence of graphs. An approach to identifying these classes is via a product space representation which makes the time evolution Markov. The approach is illustrated via epidemics on small graphs.

  Techniques currently being investigated for identifying membership of these classes include numerical and symbolic computation.

FYSON, Nick *University of Bristol*

Network Reconstruction by Set Covering

Monday 11:30–12:00

- We present a method for the reconstruction of networks, based on the order of nodes visited by a stochastic branching process. Distinct 'markers' are seeded into the underlying network at particular points, which then propagate along directed edges and result in pairs of locations and times for each 'report'. Our algorithm reconstructs a network of minimal size that ensures consistency with the observed data.

  We first formally define the notion of 'global consistency' between a network and observed data, in line with the inuition that the reconstructed network must be capable of producing the data actually observed. We then define the concept of 'local consistency', in which only the immediate neighbourhood of a node is considered. Crucially, we then demonstrate that local consistency across all nodes necessarily implies global consistency. Hence we can perform our reconstruction through purely local considerations, inferring the neighbourhood of each node in turn.

  We show that the optimisation problem for each individual node can be reduced to a Set Covering Problem, which is known to be NP-hard but in practice can be well approximated. We determine theoretical bounds on the performance of our reconstruction algorithm, considering both the amount of data required and the performance of the set covering heuristics. Experiments confirm these theoretical results, in which

we reconstruct networks from synthetic data generated by an SIR-like epidemiological model.

POLANSKI, Arnold *Queen's University Belfast*

Recovering Connection Structures from Individual Attributes

Monday 12:00–12:30

- One of the most important challenges of network analysis remains the scarcity of reliable information on existing connection structures. This work explores theoretical and empirical methods of inferring directed networks from nodes attributes and from functions of these attributes that are computed for connected nodes. We discuss the conditions under which an underlying connection structure can be (probabilistically) recovered and propose an iterative recovery algorithm. In an empirical application, we test the algorithm on the data from the European School Survey Project on Alcohol and Other Drugs (ESPAD).

WANG, Xueying *SAMSI*

Pairwise Closure Approximations in epidemic models on networks

Monday 13:30–14:00

- Mathematical modeling of contagion dynamics spreading on social networks is of great interest from both theoretical and numerical point of view. Pair approximations for epidemic models incorporate network structures, including the degree distribution and degree correlations, into the models, one that can capture the spatial correlation of the networks. To close the approximations for epidemic dynamics, various closure approximations have been considered.

  In this work, we compare pairwise closure approximations for the Susceptible-Infective-Susceptible (SIS) model [Eames and Keeling, 2002] and the Susceptible-Infective-Recovered-Susceptible (SIRS) model [Ganna Rozhnova and Ana Nunes, 2009] on a regular graph. We identify a connection between these work. It turns out that they used essentially the same idea for the closure approximations. Moreover, the former model (and Susceptible-Infective-Recovered (SIR) models respectively) can be obtained from the latter model in the limiting case where recovered individuals lose immunity instantaneously (and infective individuals will have life-long immunity as soon as they are recovered respectively).

  We are working on stochastic SIRS models and we are endeavored to understand how the distribution final epidemic size and the basic reproduction number, which is defined as the expected number of individuals that a single infectious individual infects

in an otherwise susceptible population, depend on the network topology and initial conditions when the large population. postertitle: Pairwise Closure Approximations in epidemic models on networks posterabstract: Mathematical modeling of contagion dynamics spreading on social networks is of great interest from both theoretical and numerical point of view. In this work, we compare pairwise closure approximations for the Susceptible-Infective-Susceptible (SIS) model [Eames and Keeling, 2002] and the Susceptible-Infective-Recovered-Susceptible (SIRS) model [Ganna Rozhnova and Ana Nunes, 2009] on a regular graph. We identify a connection between these work. It turns out that they used essentially the same idea for the closure approximations. Moreover, the former model (and Susceptible-Infective-Recovered (SIR) models respectively) can be obtained from the latter model in the limiting case where recovered individuals lose immunity instantaneously (and infective individuals will have life-long immunity as soon as they are recovered respectively).

KYPRAIOS, Theodore *University of Nottingham*

Bayesian Inference for Stochastic Epidemic Models on Networks

Monday 14:00–14:30

- Bayesian Inference for Stochastic Epidemic Models on Networks talkabstract: Understanding the spread of an infectious disease is a highly important issue in order to prevent major outbreaks of an epidemic. Human infections such as influenza, malaria and HIV are still major causes of morbidity and mortality worldwide. In 2001, the UK experienced a range of severe economic and social effects of a Foot-and-Mouth (FMD) epidemic. Back in 2007, the Department for Environment Food and Rural Affairs (DEFRA) was expecting an Avian Influenza outbreak to hit the poultry industry due to many outbreaks occurring around Europe that time. Therefore, several modelling groups got involved in order to prepare for such an outbreak. In addition, very recently, the UK experienced an H1N1 outbreak in humans, mostly known as "Swine Flu". In many of these outbreaks, considerable transmission of the disease had already taken place even before the danger had been noticed. It is easy to realise that the available control strategies need to be imposed rapidly so as to effectively stop the spread of the infection. Epidemic models could be used to provide a better understanding of the transmission dynamics, the infection process, and the epidemiologically quantities of interest.

  In the early years of epidemic modelling, most of the epidemic models which have been introduced were fairly simple due to their mathematical tractability; for instance, the-so-called Susceptible-Infective-Removed (SIR) model and many variants of it, such as SIS, SI and SIRS. These models, basically, describe the different states at which an individual could be during the epidemic as well as the (stochastic) mechanism via which individuals move from one state to another. Nevertheless, in recent years there has been an increase in research activity regarding stochastic models for epidemics among populations with some kind of social structure. This was motivated by the need

for model realism, and in particular by the fact that real-life human populations are themselves structured.

Undertaking statistical inference based on stochastic epidemic models and data from disease outbreaks is generally a non-standard problem. This is due to both the nature of the data, which is highly dependent and typically partial observed, and also to the level of mathematical intractability of even the simplest stochastic epidemic models. Markov chain Monte Carlo (MCMC) methods offer, at least in principle, important advantages over standard methods such as maximum likelihood, most notable of which is the fact that they allow a much greater degree of modelling flexibility. However, the implementation of MCMC methods may be problematic, since algorithm convergence and mixing difficulties can arise due to the amount of missing data and correlation structures inherent within epidemic models. Consequently, algorithms often need to be designed with care.

In this talk, I will give a range of examples where networks have successfully been used in epidemic modelling (see, for example, Jewell et. al. 2009a, 200b) and also show how someone can draw Bayesian inference for the parameters of the epidemic model governing transmission as well as the network structure in some simple cases (Britton and O'Neill, 2002). Finally, we will also highlight the current challenges in the area of epidemic modelling and networks and discuss potential further research directions in this area.

ROBINSON, Katy *University of Bristol*

The dynamics of sexual contact networks: effects on disease spread and control

Monday 14:30–15:00

- Sexually transmitted infections continue to pose a challenge for public health despite the use of interventions such as screening programmes, vaccination, drug therapies and efforts to reduce high risk behaviors. In part this ongoing challenge is due to the heterogeneous and dynamic partnership networks over which such pathogens are spread. This has motivated the use of network-based models to study sexually transmitted infections, though most of the literature has focused on static networks. However, edges within the network will appear and disappear over time as partnerships are formed and dissolved, so that some individuals will be inaccessible to a particular pathogen even if located in the same network component.

  We develop a simple model for the formation of dynamic sexual contact networks based on partnership data from the National Survey of Sexual Attitudes and Lifestyles (Natsal 2000). Although this formation model is not based on preferential attachment, the networks created show long-tailed degree distributions as well as closely matching epidemiologically relevant characteristics such as gap length and concurrency of partnerships. We investigate the different effective networks available to pathogens with

different durations of infectiousness or of differing transmissabilities. We find that the complex interactions between pathogen characteristics and the behavioral network affect which types of individuals are most at risk of becoming infected. We use our modeling framework to identify high- and low-risk groups and to efficiently select and direct interventions within the population as a whole.

Authors: Katy Robinson (BCCS, University of Bristol), Ted Cohen (Global Health Equity, Brigham & Women's Hospital; Epidemiology, Harvard School of Public Health), Caroline Colijn (Engineering Mathematics, University of Bristol)

JONES, Nick *Oxford Physics*

A Taxonomy of Networks: Using a Mesoscopic Response Function to investigate structure in empirical networks

Tuesday 10:30–11:00

- The science of networks has grown into an enormous interdisciplinary endeavour across the natural, social, and information sciences. Yet many disciplines that employ network methods remain relatively unconnected. Here, we introduce a framework to establish a taxonomy of networks from various origins. The provision of this family tree not only helps understand the interrelatedness of networks, but also facilitates carrying over empirical analysis and modeling efforts across disciplinary boundaries. The framework is based on probing the mesoscopic properties of networks, an important source of heterogeneity for their structure and function, and in itself constitutes the first methodological application of community detection to classify networks. We provided a taxonomy for 699 networks, and used that to derive a stylized taxonomy. We also applied the method to three case studies in a political, sociological, and financial context, finding that the taxonomy produced by the framework can be externally validated in each case.

LAMBIOTTE, Renaud *Imperial College London*

Dynamics, Modularity and Robustness of Complex Networks

Tuesday 11:00–11:30

- The complex structure of many social, information and biological networks is underpinned by communities at different scales. These topological modules are often indicative of underlying features and functionalities, such as tightly-knit groups of metabolites or species in biological networks. The presence of well-defined communities also has an effect on the dynamics taking place on a network. A variety of methods and measures have been proposed to uncover these modules, most notably modularity and spectral partitioning. However, these approaches are based on structural, static properties of

the network. Here we introduce a definition for the quality of the partition of a network that is based on the statistical properties of a dynamical process taking place on the graph. This measure, denoted the stability of the partition, has an intrinsic dependence on the time-scale of the process, which can be used to uncover community structures at different resolutions. The stability extends and unifies standard community detection algorithms. In particular, both modularity and spectral partitioning are shown to have a dynamical interpretation in the case of undirected networks and can be seen as limiting cases of the stability. Similarly, several multi-resolution methods correspond to linearisations of our measure at short times. In the case of directed networks, however, stability differs from modularity by its non-local nature as it is based on the persistence of probabilistic flows in modules. We apply our method to find multi-scale partitions for different examples and show that stability can be computed efficiently through the use of extended versions of current algorithms that can deal with large networks. Finally, we discuss the possibility to detect the most significant scales of the system through statistical tests based on measures of robustness.

AMBLARD, Pierre-olivier *CNRS/GIPSAlab*

Directed information theory to infer causality graphs

Tuesday 11:30–12:00

- The problem considered here is to infer directed connectivity in networks of interacting stochastic processes. We suppose to have access to a realization of a multivariate stochastic process, and we want to exhibit the directed links between the components of the process. When direction is not required, this problem is the problem of graphical modeling [1]. However, adding direction makes the problem harder. One possibility to assess direction is to use the notion of causality.

  Causality graphs are an extension of usual graphical models of multivariate processes to include directivity [2]. Directivity is assessed using Granger causality which is based on prediction. Causality graphs have been introduced and extensively studied by Eichler & Dahlaus. The inference problem has also received some attention by these authors, especially when considering weak forms of causality (weak causality relies on linear prediction: a process x weakly causes a process y if the linear prediction based on its past only is improved by including the past of x). Weak causality in graphs is then assessed using classical second order statistical tools such as correlation and coherence and their partial counterparts.

  Strict causality relies on probability distributions. x(t) does not cause y(t) if and only if y(t) and the past of x(t) are independent conditionally on the past of y(t). In such a case, considering the past of x will not improve the prediction of y. As studied by Granger, causality must be considered as a relative notion, relativity being based on the set of observations at hand. Indeed, the causality relationship between two processes can be changed when a third process is observed. This is very important in networks of processes and forbids the repeated use of bivariate analysis.

Strict causality graphs are well defined, but the inference problem for these graphs is a difficult problem. Indeed, it requires the development of tools to assess conditional independence between high dimensional random variables. One of the possible tools is the so-called directed information theory. It is a recent extension of information theory that allows to tackle with feedback and directivity. Usual information theory uses mutual information to measure the information shared by random variables. But mutual information is a symmetrical measure that can not assess the direction (if any) of the flow of information between random variables. In order to reintroduce directivity of information in the measures of information theory, the notion of causal conditioning can be introduced. We will discuss this notion and exhibit the links it has with Granger causality. This will allow us to show that the usefull information-theoretic measure for inferring causality graphs is the causal conditional directed information.

Once the theoretical link between directed information theory and causality graph is done, the important question of estimation and testing is raised. We will discuss the practical important issues for inferring causality graphs, especially the problem of estimation of information-theoretic measures from data.

McCORMICK, Tyler *Department of Statistics, Columbia University*

Latent Structure Models for Social Networks using Aggregated Relational Data

Tuesday 15:30–16:00

- Social networks have become an increasingly common framework for understanding and explaining social phenomena. But despite an abundance of sophisticated models, social network research has yet to realize its full potential, in part because of the difficulty of collecting social network data. In contrast, Aggregated Relational Data, commonly collected as questions of the form "How many X's do you know?", measure network relationships indirectly and are easily incorporated into standard surveys. We propose a latent space model where the propensity of an individual to know members of a given alter group (people named Michael, for example) is independent given the positions of the individual and the group in a latent "social space." This framework is similar in spirit to previous latent space models proposed for networks (Hoff , Raftery and Handcock (2002), for example) but doesn't require that the entire network be observed.

  Using this framework, we derive evidence of social structure in personal acquaintance networks, estimate homogeneity of groups, and estimate individual and population gregariousness. Our method makes information about more complicated network structure available to the multitude of researchers who cannot practically or financially collect data from the entire network.

Finding Rumor Sources in Networks

Tuesday 16:00–16:30

- We provide a systematic study of the problem of finding the rumor source in a network. We use a simple rumor spreading model based upon the SIR model and then cast finding the rumor source as a maximum likelihood (ML) estimation problem. For a general network this seems to be a daunting task, so we begin by addressing the rumor source estimation problem for trees. For regular trees, we are able to reduce the ML estimator to a novel combinatorial quantity we call rumor centrality. In principle, rumor centrality involves the sum of an exponential number of terms. However, for trees we find a structural property that allows for a linear time message-passing algorithm for evaluating rumor centrality. We extend this notion of rumor centrality to construct estimators for general trees. For a general graph, we note that there is an underlying tree which corresponds to the first time each node becomes infected. Using this intuition, we develop estimators for general graphs which utilize rumor centrality and breadth first search (BFS) trees (the idea being that the rumor would spread fastest along a tree that is close to the BFS tree).

  To understand the estimator performance in terms of it being able to correctly find the rumor source, we study its performance on general trees. Somewhat surprisingly, we find the following threshold phenomenon about the estimator's effectiveness. If a tree grows like a line, then the detection probability of the ML rumor source estimator will go to 0 as the network grows in size; but for trees growing faster than a line, the detection probability of our estimator will always be strictly greater than 0 (uniformly bounded away from 0) irrespective of the network size. In the latter case, we find that when the estimator makes an error, the wrong prediction is within a few hops of the actual source. Thus, our estimator is essentially the optimal for any tree network. The proofs of these results are non-trivial and require novel analytic techniques which may be of general interest in the context of graphical inference and percolation.

  We study the performance of the general graph rumor source estimator through extensive simulations. As representative results, we test the estimator's performance on the popular small-world and scale-free networks, and also on a real internet autonomous system (AS) network, the U.S. electrical power grid network, and online social networks. For online social networks, rumor spreading would correspond to a rumor or trend spreading, but rumor spreading on the AS network corresponds to the spread of a computer virus, while rumor spreading on the power grid network could instead represent a cascading failure or a blackout. We find that the estimator performs well on all of these different networks.

  We compare the new notion of rumor centrality with the more common distance centrality. We show that on trees, the rumor center is equivalent to the distance center. This indicates that distance centrality is the correct estimator for trees and tree-like networks. However, on general networks the rumor center and the distance center can be different. This is because distance centrality only considers the shortest paths in

the network, whereas rumor centrality utilizes a richer structure. Through simulations, we find that rumor centrality is a better estimator for the rumor source than distance centrality on networks which are not tree-like, such as the small-world and power grid networks.

HEARD, Nick *Imperial College London*

Bayesian Anomaly Detection Methods for Social Networks

Tuesday 16:30–17:00

- Anomaly detection on graphs of social or communication networks has important security applications. However, learning the network structure of a large graph is computationally demanding, and dynamically monitoring the network over time for any changes in structure threatens to be more challenging still.

  This talk presents a two-stage method for anomaly detection in dynamic graphs: the first stage uses simple, conjugate Bayesian models for discrete time counting processes to track the pairwise links of all nodes in the graph to assess normality of behaviour; the second stage applies standard network inference tools on a greatly reduced subset of potentially anomalous nodes.

  The utility of the propose method is demonstrated on simulated and real data sets. The real data come from the European Commission Joint Research Centres (JRC) European Media Monitor (EMM) (http://emm.jrc.it). EMM is a web intelligence service, providing real-time press and media summaries to Commission cabinets and services, including a breaking news and alerting service. The simulated data come from the VAST Challenge 2008 (http://www.cs.umd.edu/hcil/VASTchallenge08); we consider the simulated cell phone data from the Mini Challenge focused in the area of social network analysis.

PENFOLD, Christopher *University of Warwick*

Systems Biology Networks

Wednesday 10:30–11:00

- A key objective of Systems Biology lies with inferring a gene-regulatory network from time-series observations of gene expression. Whilst there exist many methods for inferring networks, the sheer quantity of genes measured by high-throughput microarrays compared to the relatively sparse number of measured observations means that identifying genome-wide networks is a problematic, if we are to avoid overfitting.

  Rather than identifying global networks, one might instead wish to identify a local network (M) operating about a particular (set of) gene(s) of interest G, where G c M.

Identifying the complementary genes, M' = M, is nontrivial, however, since there is no way to known a priori whether a particular gene acts on G or vice versa.

Here we present a method identifying potential genes, M;, using a Metropolis-style wrapper for a variational Bayesian State Space model (VBSSM) as follows:

1. A particular (set of) gene(s), G , is chosen for biological reasons and will be conditionally admitted to all suggested networks. 2. Additionally a random set of N genes is chosen from a pool J to be modelled alongside G to give an intial guess for the network G . Using the state-space model of Beal et al. (cite) the marginal likelihood for the combination of genes may be evaluated. 3. At each step, a new network M is chosen by randomly picking n genes from M; = M and replacing with n genes from the pool J . Note that the number of genes to be swapped may be assumed to be Binomially distributed with tunable parameter p. The marginal likelihood calculated at each step may then be compared to the previous step and the new set of genes accepted with the Metropolis criterion.

The above approach may be run for a suffcient number of steps such that the marginal likelihood converges to an upper limit, to yield an estimate for M'.

We validate the method using simulated data from in silico networks and measured expression data from the Arabidopsis thaliana circadian clock. These studies demonstrate that the above approach is adept at picking out groups of expression profiles belonging to genes from the same network, compared to groups of profiles belonging to a perturbation of the network.

We illustrate an application of the method to developing local networks for 2 genes implicated in plant senesence BrotherFT, HAT.

IQBAL, Mudassar *University of Warwick*

An Integrative Bayesian Analysis of Transcription Regulation in S. coelicolor

Wednesday 11:00–11:30

- Given the abundance of different types of data, e.g., gene expression as well as sequence data, one of the main goals of systems biology is utilize this data to identify regulatory relationships between transcription factors and their target genes. To that end, efficient methods need to be developed which are capable of integrating diverse datasets and make biologically plausible predictions. We present a generic method for poorly studied organisms, requiring microarray data and a sequenced genome. As part of European consortium (SysMO-STREAM Project), we are studying an important soil bacterium, S. coelicolor, a model organism within the actinomyces, a genus responsible for the production of most of the antibiotics currently in use. S. coelicolor produces four antibiotics; however despite decades of research the regulatory network that induces their production across a variety of conditions is still only partially known. The

shift from primary to secondary metabolism (antibiotic production) was studied using high-resolution time series (35-50 time points) on the sequenced strain M145 across a variety of conditions and for a number of knock-out mutants. We integrate, using a Bayesian approach, gene expression data with de novo binding site data; only a hand full of motifs being known to date. Motif searches include those for dyad type motifs (two conserved words with variable spacer between them, which is quite common in bacterial genomes) and conserved motifs across orthologous species. We modified the Factor model of Sabatti & James (Bioinformatics 2006) to predict the motif activity across our time series data, activity profiles that we subsequently match to transcriptional regulator expression. This is a linear regression model based on the assumption that gene expression values are determined by the activity of the unknown factors (motif activities) while connectivity is restricted by the presence of the motif. Our model allows for varying motif confidence and knock-outs. We implemented our model using Markov chain Monte Carlo methodology, utilising two techniques to improve performance, specifically a local M-H move for the connectivity, and parallel tempering to improve mixing and convergence. The latter was found to be essential for large data sets (¿30 motifs, ¿2000 genes). We tested our combined motif search and transcription factor activity modelling on E.coli, comparing to the known transcription factor binding motifs, demonstrating our models explanatory potential. When applied to the S. coelicolor data we identified a number of key motifs which are being experimentally verified for protein binding (gel retardation).

JUAREZ, Miguel *University of Warwick*

Inferring the topology of a non-linear gene regulatory network using fully Bayesian spline regression

Wednesday 11:30–12:00

- There is significant interest in retrieving (reverse engineering) the topology of an interaction network, mainly driven by biological applications. Using time series data, our main objective is the study of genetic regulatory networks (GRN). The problem is to infer the relations between a group of units (genes). When stated in a graphical manner such units are represented as nodes and their relations as edges; the edges can be directed or not, and loops, feedbacks and cliques may or may not be allowed. One very well known example of such graphs is a directed acyclic graph, characterised by directed edges and a tree structure. From a Bayesian perspective, these are usually estimated using Bayesian Networks.

  Bayesian networks (BN) have been used previously in gene network determination. However, it is well known that biological processes have feedback oops and thus the validity of BN is questionable when modelling such systems. Dynamic Bayesian networks (DBN) have been proposed for modelling time course (longitudinal) gene expression data. These can be thought of as "unfolding" a BN for every time point and when folding back the network self-regulation and cliques may be obtained. Formally, a DBN

is characterised by a set of conditional relations, $p(y^{t+1} \mid \mathbf{y}^t)$. In the case of a regression based DBN these relations can be written as $y_i^{t+1} = f_i(\mathbf{y}^t) + \varepsilon_i^{t+1}$, where y_i^t is the measurement of unit $i = 1, \ldots, G$, at time $t = 1, \ldots, T$, $\mathbf{y}^t = \{y_1^t, y_2^t, \ldots, y_G^t\}$ and ε_i^t is an idiosyncratic error term. The functional forms of the interactions, $f_i(\cdot)$, are usually unknown. Whether $\partial y_i^{t+1}/\partial y_j^t \equiv 0$ or not defines the structure (topology) of the network. The interaction topology is key in GRN, as it determines the causal relations in the gene regulatory dynamics for a given biological process. In turn, the estimated network can be used to propose new experiments –*e.g.* gene knockouts– to further understand said process. This programme of experimenting-modelling-hypothesising-experimenting is at the heart of Systems Biology.

Irrespective of the actual form of $f_i(\cdot)$, DBN are typically heavily parameterised. If we consider a network of G units, we need to estimate G^2 interactions (edges), in addition to any other parameters involved in the model. To avoid further complications, the common approach is to assume a linear form for the interactions. Within a GRN framework, regulatory relations are frequently modelled as systems of ordinary differential equations (ODEs), with monotonic functional interactions, and thus this simplification is justified as a first order Taylor expansion. Frequently, however, the linear assumption does not hold in practice; either because some of the interactions are indeed highly non-linear or due to a large spacing between measurements, thus rendering the linear approximation invalid. Specifying a different form for the interactions (power, exponential, Michaelis-Menten, etc.) entails a larger number of parameters than in the linear case, and may be feasible only if a sufficient number of data points are available. Moreover, misspecification of the actual shape may yield a spurious estimate of the network topology.

A flexible way of including unknown non-linearities is to use a semi-parametric specification by letting the interactions be described by spline functions. There is a vast literature on spline curve smoothing and spline regression. One fundamental problem when using spline regression is knot selection, which greatly influences the curve fitting. One efficient solution would be to select only a few well placed knots, for a given spline degree. This implies determining both the optimal number and position of the knots, which is typically addressed by means of a trans-dimensional MCMC scheme or by cross-validation. The efficiency gained in the modelling may be offset by mixing problems in the sampler, due mainly to the vast space that must be explored and the associated computational problems, or by the unwieldy amount of comparisons required for cross-validation.

Our approach avoids such issues by relying on P-splines, which are characterised by specifying a rather large number of evenly spaced knots. Then, in order to avoid overfitting and also to control for the effective number of parameters to be estimated, a penalty that shrinks the spline coefficients towards the origin is specified. Such a penalty depends crucially on a so-called smoothness parameter. Semi-parametric P-spline regression models for GRN retrieval have been proposed, which optimise the smoothness parameter using a modified BIC and then by performing a greedy search on the network topology space. In this paper, we propose a fully Bayesian set up for dealing with this smoothness parameter and discuss in some detail the implications of alternative prior specifications for this key parameter.

Given that it is common to specify an improper prior for this kind of model, we provide sufficient conditions for posterior propriety. The resulting posterior distribution is intractable analytically and we provide an MCMC scheme for sampling from it, with a novel Metropolis-Hastings step which improves mixing and convergence of the chain. The application of our model is illustrated with synthetic and real data sets, where we reconstruct the corresponding networks and assess their accuracy. We also provide some guidelines for calibrating the prior.

BOWSHER, Clive *University of Cambridge*

Biomolecular Networks: Dynamic Independence, Modularisation and Information Processing

Wednesday 13:00–13:30

- A Stochastic Kinetic Model (SKM) is a highly multivariate jump process used to model the biochemical reaction networks inside cells. We introduce a directed, cyclic graph (the Kinetic Independence Graph or KIG) which encodes the local independence structure of a given SKM. Given a partition [A,D,B] of the vertices, the graphical separation of A and B by D in the undirected KIG has an intuitive biochemical interpretation and, under conditions we derive, implies the dynamic conditional independence of the internal histories (up to any time t) of A and B given the internal history of D. Such global conditional independences are used to address modularisation of and information processing by biochemical networks. We discuss the automation of our methods using the MIDIA algorithm which is based on graphical decompositions and junction tree manipulations, and their application to signal transduction networks in systems biology.

SMITH, Andrew *University of Bristol*

Nonparametric regression on a graph

Wednesday 13:30–14:00

- We will look at the problem of nonparametric regression in the context of removing noise from observations taken at the vertices of a graph. So rather than making inferences about distributions on the edges, we make estimates and inferences at all the vertices. There are many existing regression situations that contain a graphical structure, and we will consider examples of scatterplot smoothing, image analysis and denoising UK house price data.

  Regression on a graph involves fitting an estimate that somehow explains the observations, some of which may be missing, taken at the vertices. Nonparametric regression

is more appropriate since the underlying trend in the observations, and the graph itself, may be completely arbitrary. The edges of the graph provide information about the distance between observations, and in some applications this is the only such information available.

Borrowing ideas from penalised regression and total variation denoising, we penalise distance from data on the vertices, and roughness on the edges, measured in the L2 and L1 norms respectively. There are computational challenges associated with implementing these penalty terms, so we will see the results of a new, fast algorithm for denoising on a graph.

The presence of graphical structures in regression problems is often surprising, and may suggest some new insights into the connections between networks and other statistical models. The algorithm can also be adapted to identify clusters of neighbouring vertices, tackle classification problems, and identify important edges in penalised regression.

PERRY, Patrick O. *Harvard University*

A graph log-linear model for characterizing repeated interactions

Wednesday 14:00–14:30

- We are surrounded by interaction data. More and more, this data is being harnessed for inferential purposes. Phone and email records are being used to study communications networks. Records of legislation cosponsorship and journal coauthorship are being used to study collaboration networks. Animal co-occurrence data is being used to study herding and association behaviors. Network scientists have risen to the challenge of analyzing these new types of data, delivering a host of promising and powerful new methods and models. Most network models are designed for modeling binary relations (e.g. whether or not two people are friends). These methods are not directly applicable to data with varying frequencies of interactions between actors. We propose a simple modeling framework to handle data with repeated interactions.

  Our main idea is that for many types of interaction data, the network of interactions is just a convenient summary of the data. The full data set is a sequence of time-stamped interactions. We know more than the number that actors i and j interacted–we also know the times of their interactions. Suppose actors labeled 1, 2, ..., n are under observation for times in the interval (0, T]. Suppose that each interaction takes place instantaneously. Some interactions (e.g., phone calls) do not take place instantaneously; in cases like these, identify the interactions with their starting times. Now, consider the interactions between actors i and j as a point process indexed by time.

  In studying the interactions of all actors under observation, we introduce a graph with point processes on its edges. Each node in the graph is an actor, and the point process on edge (i, j) represents the times of the interactions between actors i and j. We start by supposing that the set of interaction times for a pair of actors is a stationary Poisson process, which we define below. Next, we assume that the interactions between

different pairs of actors are independent. That is, if (i, j) and (i', j') are two different pairs of actors, then the interactions of these pairs are independent. Under the two simplifying conditions above, the total interaction counts are sufficient statistics for the edge intensities.

We introduce a "graph log-linear model", or graph-LLM, to predict the frequency of interaction between pairs of actors. This is a graph with a log-linear model on each edge. The number of interactions on edge (i,j) is modeled in terms of the edge's endpoints and other edge-specific covariates. We show how to estimate the parameters of the model and also give asymptotic results showing that the estimates are consistent when the observation interval or the number of nodes goes to infinity.

We validate our modeling framework using a subset of the emails sent within the Enron corporation. Over small windows, we demonstrate the interactions between people to be approximately pairwise independent and time homogeneous. Using past patterns of interaction to estimate the connectivity graph and model parameters, the graph-LLM model is able to outperform a log-linear model that ignores the network structure.

This is joint work with Patrick Wolfe, Harvard Statistics and Information Sciences Lab.

BEJAN, Andrei *University of Cambridge*

Statistical Modelling and Analysis of Sparse Bus Probe Data in Urban Areas

Wednesday 16:00–16:30

- Congestion in urban areas causes financial loss to business and increased use of energy compared with free-flowing traffic. Providing citizens with accurate information on traffic conditions can encourage journeys at times of low congestion, and uptake of public transport. Metricating a city to provide this information is expensive and potentially invades privacy. Increasingly, public transport vehicles are equipped with sensors to provide real-time arrival time estimates, but these data are sparse. Our work shows that these data can be used to estimate journey times experienced by road users generally.

  In this study we investigated how access to a large repository of bus trajectories combined with other publicly available data, such as OpenStreetMap (OSM) data and National Public Transport Access Node (NaPTAN) database, can be used in order to provide a detailed account of journey times and some of the important factors that affect them. Specifically, we (i) describe what a typical data set from a fleet of over 100 buses looks like; (ii) describe an algorithm to extract bus journeys and estimate their duration along a single route; (iii) show how to best visualise journey times and the influence of contextual factors; (iv) validate our approach for recovering speed information from the sparse movement data.

  Several types of graphical displays, a nonparametric quantile regression and monotonic splines technique are used to visualise these effects on journey time and recover speed

information from sparse data. We have also been able to study in some detail the sources of delay within individual journeys and have presented our findings using a novel and insightful technique based on a notion of local time profile which we define. Our techniques provide a detailed basis for understanding journey time behaviour in the urban environment and lead to further important research areas such as how to incorporate real-time data to improve prediction of journey times.

This is a joint work with Richard Gibbens, David Evans, Alastair Beresford, Jean Bacon, and Adrian Friday.

KEYWORDS: sensor networks, bus probe data, journey times, large scale data analysis, knowledge discovery, quantile regression, spline interpolation

---

GIBBENS, Richard *University of Cambridge*

An investigation of proportionally fair ramp metering

Wednesday 16:30–17:00

- This talk concerns ramp metering which is one important approach to dealing with congestion on motorways. Congestion occurs when demand exceeds available resources and can significantly reduce the capacity of the motorway network at peak times. Reduced capacity results in additional delays, increased environmental pollution and hinders passenger safety. Congestion is observed to cause low but highly volatile speeds resulting in more uncertain journey times (referred to as flow breakdown or stop-and-go behaviour) (Gibbens & Saatci (2008)).

  Ramp metering is intended to control the entry of new flow in such a way as to maintain steady flow on the motorway and to avoid the flow breakdown associated with congestion. The rate of entry of flow is set according to the particular ramp metering strategy. Such strategies have been the subject of much attention in the transport literature. One of the key issues is the trade-off between efficiency and fair use of resources. This is a trade-off that has been considered extensively in the modelling and control of communication networks.

  We describe a study that adds to recent work (Kelly & Williams (2010)) on a ramp metering strategy, proportionally fair metering, inspired by rate control mechanisms developed for the Internet. Specifically, we use simulation results to compare proportionally fair metering with a greedy strategy for a linear network with a series of entry points leading towards a single common destination for all the traffic, such as a radial route towards a city centre. Under our modelling assumptions, the greedy strategy is provably optimal for exogenously determined arrival streams of traffic, but it is unfair, in a certain precise sense, between different entry points and may well have perverse and suboptimal consequences if it influences traffic demand. We further consider a network with parallel roads where flows of traffic may have route choice according to the levels of queueing at the individual entry points.

Joint work with F.P. Kelly. Based on work to be presented at 5th IMA Conference on Mathematics in Transport, UCL, London, April 2010.

GASTNER, Michael *Imperial College London*

The complex network of global cargo ship movements

Wednesday 17:00–17:30

- The ability to travel, trade commodities and share information around the world with unprecedented efficiency is a defining feature of the modern globalised economy. Among the different means of transport, ocean shipping stands out as the most energy efficient mode of long-distance transport for large quantities of goods. According to estimates, as much as 90 per cent of world trade is hauled by ships. In 2006, 7.4 billion tons of goods were loaded at the worlds ports and the trade volume exceeded 30 trillion ton-miles.

  The worldwide maritime network also plays a crucial role in todays spread of invasive species. Two major pathways for marine bio-invasion are discharged water from ships ballast tanks and hull fouling. Even terrestrial species such as insects are sometimes inadvertently transported in shipping containers. In several parts of the world, invasive species have caused dramatic levels of species extinction and landscape alteration, thus damaging ecosystems and creating hazards for human livelihoods, health and local economies.

  In the spirit of current network research, I take here a large-scale perspective on the global cargo ship network as a complex system defined as the network of ports that are connected by links if ship traffic passes between them. I will present how the world-wide network of ports can be inferred from itineraries of 16,363 cargo ships during the year 2007. The networks properties will be compared to other transportation networks and previous models of network flows.

  I will show that the global cargo ship network has a small-world topology where the combined cargo capacity of ships calling at a given port (measured in gross tonnage) follows a heavy-tailed distribution. This capacity scales super-linearly with the number of directly connected ports. Complex network theory can identify the most central ports in the network and finds several groups of highly interconnected ports showing the importance of regional geopolitical and trading blocks.

  I will compare the empirical data with theoretical traffic flows calculated by the gravity model. Simulation results, based on the empirical network or the gravity model, differ significantly in a population-dynamic model for the spread of invasive species between the worlds ports. Predictions based on the real network are thus more informative for international policy decisions concerning the stability of worldwide trade and for reducing the risks of bio-invasion. At the end of the talk I will discuss implications and open questions.


SENDOVA-FRANKS, Ana B. *University of the West of England, Bristol*

Emergency networking in ant colonies

Thursday 9:30–10:00

- Network analysis is a growing area in studies of animal social structure (Croft et al. 2008, Whitehead 2008) particularly when looking at patterns of association among members of vertebrate groups. However, networks based on functional behavioural contacts are rare. Social interactions reach a zenith in insect societies. Here we present the first results from the analysis of resource distribution in colonies of the ant Temnothorax albipennis (Sendova-Franks et al. 2009). These ants live in approximately 2-D rock crevices and have been established by the Franks Ant Lab in Bristol as a model system for experimental manipulation and the development of new theory.

  Resource distribution is fundamental to social organization, but it poses a dilemma. How to facilitate the spread of useful resources but restrict harmful substances? This dilemma could precipitate a crisis during famine relief. Survival depends on distributing food fast but that could increase vulnerability to poisons. We tested how T. albipennis ants solve this dilemma in the distribution of honey solution after 48 h of starvation in four colonies with individually marked workers. We constructed the complete network of liquid food transmission (trophallaxis) between individuals. Within the first 30 min of famine relief, 95and the distribution rate was an order of magnitude faster compared to the controls. We tested the assumptions of a simple analytical model that best fitted our data. Good mixing during famine relief was facilitated by the movement of internal workers away from the brood pile and the movement of foragers with food away from the nest entrance. This is intriguing because T. albipennis workers have spatial fidelity zones and in the controls internal and external workers were segregated. We discovered that colony vulnerability to poisons during famine relief might be mitigated by: (1) the dilution of food from the same source through social mixing, (2) the concentration of food in workers positioned midway between the colony centre and its periphery and (3) the existence of living silos. The latter are expendable foragers, who stay inside the nest and store food during famine relief, thus acting as potential disposable testers for food toxicity.

  In addition to the importance of the temporal structure in the sequence of interactions, such studies also point to the significance of combining social interaction networks with spatial graphs of the interacting components.

  References:

  Croft, D. P., James, R. & Krause, J. 2008. Exploring Animal Social Networks. Princeton, New Jersey: Princeton University Press.

  Sendova-Franks, A. B., Hayward, R. K., Wulf, B., Klimek, T., James, R., Planqu, R., Britton, N. F. & Franks, N. R. 2009. Emergency networking: famine relief in ant colonies. Animal Behaviour, 79, 473485.

Whitehead, H. 2008. Analyzing Animal Societies: Quantitative Methods for Vertebrate Social Analysis. Chicago: University of Chicago Press.

Ana B. Sendova-Franks (a), Rebecca K. Hayward (b), Richard James (b), Nigel R. Franks (c)

a - Department of Mathematics and Statistics, BIT, UWE, Bristol b - Department of Physics, University of Bath c - School of Biological Sciences, University of Bristol

---

JAMES, Dick *University of Bath*

Animal Social Networks

Thursday 10:00–10:30

- The networks approach is rising in popularity among those trying to unravel the details of social structure and dynamics in animals. For the most part, the approach taken has been to use measures of network structure to relate patterns of association to fairly simple explanatory variables, such as space-use, body size and sex, behavioural traits and genetic relatedness. Some are beginning to study processes on social networks such as the transmission of information (good or bad).

  I work with biologists trying to understand the social structure of populations of animals that spend all or part of their time living in groups. The range of species is wide, from large mammals through small freshwater fish to social insects. In all cases we are trying to use networks of contacts, interactions or associations to try to relate social structure and dynamics to the biology of an individual. My role in these collaborations is to try to find useful measures of structure that are statistically robust.

  Many of the populations studied are in the wild, and form a "fission-fusion" society, in which animals are apparently free to leave and join groups, and do so frequently. It is then relatively easy to use point sampling to construct affiliation networks of association, to accumulate them over many surveys of group membership, and to analyse the resulting social networks in terms of simple explanatory variables. Since such networks are strongly affected by variation in recapture frequency and group-size distribution, we construct a biological null model of association as the basis of randomisation tests for statistically significant network structure.

  Other examples use wild-caught animals in a more controlled setting where, at the expense of losing natural conditions, we are able to replicate and manipulate observations to help tease out key aspects of social structure. Here the problems are different. There is less of an issue with sampling protocol, or emigration and immigration of individuals, but now we must take more care to ascribe biologically meaningful weights to our network edges.

  I will present a brief review of some of the projects I am involved in, highlighting the questions in social evolution that motivate our use of network analysis, one or two of the things we think we have learnt so far, some of the statistical approaches we have used and unanswered questions we would like some help with.

Invited speaker index

Contributed speaker index